POLITECNICO DI TORINO

Department of Electronics and Telecommunications

Master's Degree in ICT for Smart Societies

Politecnico di Torino

NEC

Semantic Transfer of Images through Terrestrial and Non-terrestrial Networks Leveraging Generative AI

Supervisors

Prof. Carla Fabiana Chiasserini

Dr. Marco Palena

Candidate

Federico Villata

March 2026

Table of Contents

List of Figures

List of Tables

Abstract

Selective-fidelity semantic image transfer aims to preserve task-relevant content at high quality while conveying only sparse evidence from the remaining regions, delegating completion to a generative decoder. Under strict rate constraints, two coupled transmitter decisions become critical: selecting the Region of Interest (ROI) among multiple plausible segments and sampling a limited set of patches over the Region of Non-Interest (RONI) that anchor generative inpainting.

In this work, a state-of-the-art selective-fidelity semantic image transfer framework called SPIFF is considered. An encoder-side instantiation is developed within a SPIFF-like pipeline: the receiver-side diffusion inpainting module is kept fixed, and the study intervenes exclusively on transmitter-side decisions and policies. ROI selection is formulated as a semantic ranking over candidate semantic categories produced by a GroundingDINO+SAM backend on a fixed label set. The main ranking pipeline is implemented both in a task-agnostic form, via CLIP image-to-image similarity between the input and segment-isolated views, and in a task-oriented form. Task-oriented relevance is instantiated as a face-recognition case study to provide a concrete example of task-driven ranking, and is implemented via a face-relevance heatmap derived from occlusion sensitivity of face-embedding similarity. In addition, an exploratory LLM-based ranking study is conducted separately on a reduced subset as a preliminary multimodal extension.

RONI evidence sampling is driven by informativeness measures computed from entropy, edge density, saliency, and an adaptive mixture, with patch selection strategies balancing relevance and spatial dispersion under a common coverage budget. Experiments on COCO-Stuff images with paired captions evaluate semantic ranking by agreement with a manually annotated ground truth, and assess RONI patch sampling through end-to-end reconstruction quality under a fixed decoder using perceptual and semantic fidelity metrics (LPIPS- and CLIP-based), patch-geometry indicators, and runtime measurements.

Under the adopted face-driven reference criterion, the task-oriented ranking better matches the manual annotations, reaching rank-weighted F_1 scores of 0.741 and 0.803 under rank-averaged and global weighted aggregation, respectively, while the task-agnostic variant reaches 0.283 and 0.290 under the same protocol. The exploratory LLM-based study follows the same pattern: face-oriented prompting reaches 0.559 and 0.753, whereas generic prompting reaches 0.280 and 0.346, although both remain below

the dedicated task-oriented pipeline. For RONI patch sampling, deterministic farthest point sampling yields the strongest overall fidelity profiles, with Perceptual Fidelity (PF) up to 0.631 (about 0.630 on average across cues, above the overall mean of about 0.613) and the best Semantic Fidelity (SF) of 0.966, though with much narrower margins in a near-saturated SF regime. These configurations are also associated with near-zero patch overlap and higher spatial dispersion, while decoder inference time remains essentially unchanged across configurations. Overall, this work provides a modular encoder-side study that links task-dependent semantic ranking and dispersion-aware evidence sampling to reconstruction outcomes under strict budgets.

Executive Summary

Semantic Transfer of Images through Terrestrial and Non-terrestrial Networks Leveraging Generative AI

Candidate: **Federico Villata**

Supervisors: **Prof. Carla Fabiana Chiasserini**, **Dr. Marco Palena**

Semantic image transfer targets communication settings where bandwidth, latency, and link quality are constrained and time-varying, including heterogeneous terrestrial and non-terrestrial paths. In these conditions, transmitting every pixel is often inefficient. Many scenes remain usable when only the most relevant content is explicitly preserved, and the rest is inferred by the receiver. The research context is semantic communications for heterogeneous terrestrial and non-terrestrial links, where resource variability motivates encoder policies that prioritize task-relevant meaning under strict budgets.

This thesis builds on the state-of-the-art SPIFF selective-fidelity framework by developing an encoder-side instantiation under a strict evidence budget. High-relevance regions are preserved at high quality, while lower-relevance regions are conveyed through a sparse set of localized observations (patches). Missing content is reconstructed at the receiver via conditional generative inpainting, with the decoder treated as fixed and the emphasis placed on encoder-side allocation policies.

Two coupled encoder decisions are addressed.

- **Region of Interest (ROI) selection.** Multiple plausible segments are semantically

ranked to select the ROI to preserve.

- **Region of Non-Interest (RONI) evidence sampling.** Given the preserved ROI, a limited set of RONI patches must be chosen to constrain inpainting. Patch sampling must balance informativeness and spatial diversity.

ROI selection is formulated as a semantic-ranking problem over candidate semantic categories produced by a GroundingDINO+SAM backend with a fixed label set. Two main semantic ranking families are implemented on the same candidate semantic categories.

- **Task-agnostic semantic ranking.** Segment relevance is estimated through CLIP-based similarity between the original image and segment-isolated views.

- **Task-oriented semantic ranking.** Segment relevance is instantiated as a face-recognition case study, chosen based on the available ground-truth task signal, and is implemented via a face-relevance heatmap computed through occlusion sensitivity of face-embedding similarity. Segment scores are obtained by aggregating heatmap evidence within masks.

In addition, an exploratory LLM-based ranking study is conducted separately on a reduced subset as a preliminary multimodal extension.

RONI evidence sampling is guided by pixel-wise informativeness measures computed on the RONI domain. Several cues are implemented, including local entropy, edge density, and visual saliency, together with an adaptive mixture that fuses these cues into a single informativeness measure. Under a fixed patch budget, each patch selection strategy is evaluated under each informativeness measure.

- **Score-Guided Covering (SGC).** A probabilistic patch sampler in which patch locations are drawn from a score-induced probability distribution, under feasibility constraints.

- **Farthest Point Sampling (FPS).** A dispersion-driven strategy that incrementally constructs the patch set by maximizing the minimum distance between selected patch centers. Both a deterministic variant and a probabilistic variant are considered.

- **Determinantal Point Processes (DPP).** A quality–diversity selector that favors informative yet diverse patch subsets. Patch quality is derived from informativeness

scores, while a kernel encoding spatial similarity promotes mutual repulsion. Deterministic and probabilistic variants are implemented, encouraging low-overlap, spatially dispersed evidence.

Evaluation is organized into two complementary tracks under a fixed decoder and a shared evidence budget. ROI semantic ranking is assessed by agreement with manually curated annotations under a face-driven relevance criterion. The exploratory LLM-based ranking study is evaluated separately on a reduced subset under the same agreement protocol. RONI evidence sampling is assessed through end-to-end reconstruction using an experimental protocol that jointly reports perceptual and semantic fidelity metrics (LPIPS- and CLIP-based), patch-set geometry (overlap and spatial dispersion), seed robustness for probabilistic variants, and runtime measurements.

The results highlight three main outcomes. First, under the adopted face-driven reference criterion, the task-oriented branch reaches rank-weighted F_1 values of 0.741 when performance is summarized by rank and then averaged, and 0.803 under the corresponding global weighted aggregation, whereas the task-agnostic branch reaches 0.283 and 0.290 under the same evaluation protocol. This pattern is consistent with the closer alignment of the task-oriented criterion with the adopted face-recognition notion of relevance. The exploratory LLM-based study follows the same pattern: under the same face-driven reference criterion, the face-oriented prompting regime reaches rank-weighted F_1 values of 0.559 and 0.753, whereas the generic one reaches 0.280 and 0.346, although both remain below the dedicated task-oriented ranking pipeline. Second, for RONI evidence sampling, deterministic farthest point sampling yields the strongest overall fidelity profiles. Across cues, it reaches an average Perceptual Fidelity (PF) of about 0.630, above the overall mean of about 0.613 across the 20 evaluated configurations, with the best PF = 0.631 attained under edge and mix. Under Semantic Fidelity (SF), deterministic FPS remains the best-performing family at mean SF = 0.966, compared with an overall average of about 0.963; this best mean is shared by all deterministic FPS cue variants, while the margin over the alternatives is much smaller, consistent with the near-saturated SF regime. These configurations are also associated with near-zero patch overlap and higher spatial dispersion. Third, higher dispersion is associated with better reconstruction outcomes, while decoder inference time remains essentially unchanged across evidence-sampling configurations. In contrast, probabilistic variants in-

troduce seed-dependent variability and therefore require robustness to be reported alongside mean performance.

Overall, the study provides an encoder-side characterization of selective-fidelity design choices within a SPIFF-like architecture under a fixed generative decoder. It delivers a modular implementation of semantic ranking and patch selection strategies, together with an evaluation framework that links ranking criteria, patch-set geometry, and reconstruction outcomes under strict evidence budgets.

Acronyms

- **ROI** – Region of Interest
- **RONI** – Region of Non-Interest
- **FPS** – Farthest Point Sampling
- **DPPs** – Determinantal Point Processes
- **MAP** – Maximum A Posteriori
- **RBF** – Radial Basis Function
- **CLIP** – Contrastive Language–Image Pre-training
- **LLM** – Large Language Model
- **SAM** – Segment Anything Model
- **LPIPS** – Learned Perceptual Image Patch Similarity
- **PF** – Perceptual Fidelity
- **SF** – Semantic Fidelity
- **SPIFF** – Selective Preservation of Image Fidelity Framework
- **SGC** – Score-Guided Stochastic Covering
- **IoU** – Intersection over Union

- **MCD** – Mean Center Distance
- **CV** – Coefficient of Variation
- **NTN** – Non-Terrestrial Network(s)
- **RTT** – Round-Trip Time
- **LEO** – Low Earth Orbit
- **SNR** – Signal-to-Noise Ratio
- **NMS** – Non Maximum Suppression

1 Introduction

1.1 Background and Motivation

Conventional image transmission is largely pixel-centric: the encoder aims to preserve all pixels under rate constraints, typically optimizing for reconstruction distortion. While effective for general multimedia delivery, this paradigm becomes inefficient when bandwidth or latency are tightly constrained, since the cost of transmitting every pixel is often disproportionate to the information that actually matters for understanding the scene. In many cases, semantic consistency can be maintained even when only a compact description of the most relevant content is preserved.

This observation reflects the inherently non-uniform nature of visual information: the image portions that are most critical for scene interpretation are referred to as the Region of Interest (ROI), whereas the remaining content constitutes the Region of Non-Interest (RONI). Large portions of the RONI primarily provide contextual support. Semantic-aware pipelines, therefore, aim to allocate resources unevenly, preserving the ROI at higher fidelity while spending fewer bits on secondary content.

A key challenge in such pipelines is deciding *which* region should be treated as the ROI and preserved with high fidelity. When semantic segmentation yields multiple plausible regions, identifying the true semantic focus can be formulated as a *semantic ranking* problem and strongly impacts the perceived coherence of the final reconstruction. In parallel, recent diffusion-based generative models enable realistic reconstruction from partial observations, shifting part of the reconstruction burden to the receiver.

These considerations motivate the focus on two coupled encoder-side challenges under a constrained budget:

- selecting the ROI via semantic ranking over a set of candidate semantic categories;
- transmitting a limited set of RONI patches that provide localized evidence to anchor generative reconstruction.

The goal is not uniform pixel-level accuracy, but an efficient allocation of bits that preserves what matters most for semantic fidelity and downstream understanding.

1.2 Reference Framework (SPIFF)

This thesis builds upon SPIFF (Selective Preservation of Image Fidelity Framework). This semantic image transmission pipeline achieves selective fidelity, preserving high-relevance content (ROI) at high quality and transmitting limited evidence from the remaining regions (RONI), thereby delegating completion to a generative decoder at the receiver.

In SPIFF, the transmitter performs:

1. semantic segmentation of the input image;
2. assignment of semantic categories to discrete relevance ranks;
3. preservation of the highest-rank content;
4. sampling a limited set of patches from lower-rank regions, assembling them into a compact *patch grid* with spatial metadata, and compressing the result into a single bitstream.

At the receiver, the transmitted crop/masks/patch grid is recomposed into a partial observation, and a diffusion-based inpainting model reconstructs the missing content.

In this thesis, the SPIFF decoder is treated as a fixed black-box module, while the analysis focuses on encoder-side choices: semantic ranking for ROI selection and patch selection over the RONI under a constrained budget.

1.3 Research Direction

This thesis investigates encoder-side design choices for semantic image transmission under strict rate constraints, within a pipeline that combines selective fidelity at the transmitter with generative reconstruction at the receiver.

Following the SPIFF pipeline outlined above, the transmitter-side design problem is formulated around two coupled challenges. Given a semantic segmentation into candidate regions, the following aspects are considered:

- **ROI selection:** when multiple segments may plausibly represent the subject, the region(s) to preserve at high fidelity must be selected at the encoder. Semantic ranking methodologies are primarily studied along two main complementary axes: *task-agnostic* semantic relevance and *task-oriented* relevance derived from task-specific evidence aggregated within segment masks. In addition, an exploratory multimodal LLM-based ranking study is treated as a preliminary extension and evaluated on a reduced sample set.

- **RONI evidence selection:** conditioned on the chosen ROI, the remaining content is conveyed through a limited set of RONI patches with spatial metadata, acting as anchors for generative reconstruction. Candidate locations are scored using complementary *informativeness measures* (e.g., entropy, edge density, saliency, and their combinations). A small patch set is then selected to balance *informativeness* and *spatial diversity* under a fixed budget, enabling a comparison between relevance-driven baselines and diversity-aware *patch selection strategies*.

ROI selection is primarily evaluated against curated ground-truth annotations for the two main ranking methodologies, while the exploratory LLM-based ranking is assessed separately under the same agreement protocol on a reduced subset. Reconstruction quality is assessed through perceptual and semantic fidelity metrics (e.g., LPIPS and CLIP-based similarity). By relating patch-set properties (informativeness and diversity) to reconstruction outcomes, the analysis aims to identify the encoder-side mechanisms that most reliably support high-quality reconstruction under bandwidth constraints.

1.4 Contributions

The main contributions of this thesis are summarized as follows:

- **Semantic ranking for ROI selection:** the implementation of a segment-level ranking module that orders candidate semantic regions using two main complementary methodologies: a task-agnostic semantic alignment score based on CLIP similarity and a task-oriented score derived by aggregating a task-specific relevance signal within each segment mask. The reference pipeline then preserves the top-ranked segment(s) as ROI.

- **Exploratory LLM-based semantic ranking:** an additional prompt-based ranking study in which a vision-language LLM assigns discrete category-level relevance ranks conditioned on the input image, its caption, and the detected semantic labels. This study is evaluated separately from the main ranking pipeline on a reduced subset of samples and is treated as a preliminary extension.

- **Implementation of RONI informativeness measures:** the design and implementation of multiple RONI scoring functions to guide patch sampling, including local entropy, edge density, and visual saliency, as well as a combined strategy that integrates heterogeneous cues into a unified informativeness score.

- **Implementation of patch selection strategies:** a modular and extensible framework for selecting patches over the RONI, including a relevance-driven probabilistic baseline and diversity-aware strategies based on Farthest Point Sampling (FPS) and Determinantal Point Processes (DPPs), with both deterministic and probabilistic variants.

- **Experimental analysis framework for encoder-side selection:** a structured evaluation pipeline that compares semantic ranking outputs against curated ground-truth annotations, including a separate exploratory evaluation of the LLM-based ranker under the same category-level agreement protocol on a reduced subset, and analyzes the geometric properties of sampled patch sets (e.g., overlap and spatial dispersion), relating encoder-side selection properties to reconstruction outcomes assessed through perceptual and semantic fidelity metrics.

1.5 Objectives

The main objective of this thesis is to characterize encoder-side decision strategies for semantic image transmission within a fixed generative reconstruction pipeline. The study is organized around two evaluation tracks: ROI-selection accuracy (semantic ranking) and RONI evidence sampling. Accordingly, the thesis aims to:

1. **Quantify ROI-selection accuracy:** determine how task-agnostic and task-oriented relevance definitions affect semantic ranking agreement with manually curated ground-truth annotations.

2. **Characterize the effect of RONI informativeness measures:** assess how different informativeness measures (entropy, edge density, saliency, and their adaptive combination), under a constrained patch budget, bias the spatial distribution and content of sampled (and transmitted) patches and influence generative reconstruction at the receiver.

3. **Link patch-set geometry to reconstruction outcomes:** establish relationships between sampled patch-set properties (e.g., overlap, spatial dispersion, coverage) and perceptual/semantic reconstruction quality (e.g., LPIPS and CLIP-based similarity).

4. **Evaluate robustness of probabilistic patch selection:** analyze how deterministic versus probabilistic patch selection strategies impact reconstruction variability, stability across random seeds, and sensitivity to the patch budget.

5. **Identify favorable operating regimes:** map combinations of scoring metrics and selection strategies to transmission–quality trade-offs, highlighting configurations that best balance bandwidth efficiency, spatial diversity, and reconstruction fidelity.

1.6 Outline

The remainder of this thesis is structured as follows:

- **Chapter 2 – State-of-the-Art.** Prior work is reviewed to contextualize selective-fidelity pipelines, generative inpainting at the receiver, and encoder-side policies for semantic ranking and RONI evidence sampling.

- **Chapter 3 – Methodology.** The adopted pipeline is described. ROI selection is formulated as a semantic ranking problem, using both task-agnostic and task-oriented methodologies, and is complemented by an exploratory LLM-based ranking implementation. RONI evidence sampling is defined through informativeness measures and patch selection strategies under a fixed budget.

- **Chapter 4 – Results and Discussion.** Quantitative results are reported for the two evaluation tracks: encoder-side semantic ranking agreement against ground truth and end-to-end reconstruction quality under fixed decoding, relating PF/SF outcomes to

patch-set geometry and computational cost. Additionally, a separate exploratory evaluation of the LLM-based ranker is provided.

- **Chapter 5 – Conclusion and Future Work.** The main findings are summarized, limitations are discussed, and directions for future extensions are outlined.

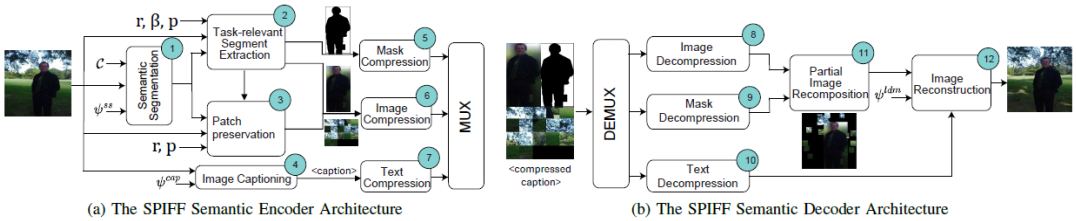
2 State-of-the-Art

2.1 Reference Framework and Scope: SPIFF as Architectural Baseline

SPIFF (Selective Preservation of Image Fidelity Framework) [1] is adopted as the architectural baseline for selective-fidelity semantic image transfer. Transmission resources are concentrated on task-relevant content (ROI), while the remaining regions (RONI) are conveyed through sparse evidence and completed at the receiver via generative inpainting. As Sec. 1.2 details the SPIFF pipeline, attention is restricted here to the architectural decision points that structure the state-of-the-art.

Throughout this thesis, SPIFF is interpreted as an encoder–decoder pipeline (Fig. 1) in which end-to-end performance is largely shaped by encoder-side allocation, while the decoder acts as a fixed conditional reconstruction operator. In particular, the framework introduces a split representation that separates ROI content preserved explicitly at high fidelity from RONI content delegated to completion. The latter is encoded as a compact set of localized image patches arranged on a grid and accompanied by spatial side information. This design is especially relevant under bandwidth constraints because it enables controllable operating points by varying the amount of explicitly preserved content and the amount and placement of transmitted evidence.

Within this baseline, two main challenges emerge, which also define the main directions along which prior work can be organized. The first challenge concerns **ROI selection**, i.e., how candidate regions or segments are identified, ranked, and selected for high-fidelity preservation. The second challenge concerns the **sampling of RONI evidence**, i.e., how sparse patches are sampled and distributed under a strict budget to constrain the subsequent completion. These decisions jointly determine the effective fidelity–rate trade-off and motivate the structure of the following sections, which review methods for semantic ranking and for budgeted evidence sampling (including informativeness measures and diversity mechanisms) in SPIFF-like architectures. On this basis, Sec. 2.3 formalizes selective fidelity in terms of fidelity notions, budget models, and failure modes of sparse communication.



(a) The SPIFF Semantic Encoder Architecture (b) The SPIFF Semantic Decoder Architecture

Figure 1: SPIFF semantic encoder and decoder architecture. The encoder isolates task-relevant content (ROI) and constructs a compact patch grid from the remaining regions (RONI); the decoder recomposes a partial observation and performs generative reconstruction. Adapted from [1].

2.2 Beyond SPIFF: Positioning and Scope

SPIFF is adopted as the architectural baseline of this thesis. The goal of this subsection is to contextualize this choice with respect to adjacent research directions. The comparison is used only to clarify the scope. It is organized along three axes:

- **Allocation unit:** explicit regions/patches versus implicit latents or channel symbols;
- **Budget model:** bitrate/coverage proxies versus channel uses/Signal-to-Noise Ratio (SNR)/latency;
- **Decoder assumption:** fixed reconstruction/completion module versus encoder–decoder co-optimization.

End-to-end learned image compression. End-to-end learned codecs optimize rate–distortion and allocate bits implicitly in latent space [4]. They typically expose global quality/bitrate controls. They do not usually provide an explicit ROI/RONI split or an explicit notion of sparse visual evidence used to condition a separate completion stage. This differs from SPIFF, where allocation is explicit, and evidence sampling can be studied directly under a fixed completion module.

Explicit semantic-structure transmission. Semantics-first schemes transmit semantic structure explicitly (e.g., masks) and target downstream inference under variable budgets. Dual-mode transmission for segmentation is a representative example [5]. In this family, appearance is auxiliary and can be omitted or coarsened depending on conditions. The re-

sulting setting differs from SPIFF, which focuses on selective visual preservation plus sparse evidence to constrain generative reconstruction.

Perception-oriented low-rate reconstruction. At very low rates, reconstruction can be dominated by learned priors. Perceptual plausibility may be favored over pixel accuracy, consistent with the perception–distortion trade-off [6]. This regime highlights the risk of semantic drift when conditioning is weak. SPIFF addresses this risk by preserving selected regions and by transmitting localized evidence to anchor completion [1].

Joint source–channel transmission. In wireless settings, joint source–channel coding maps images directly to channel symbols, as exemplified by DeepJSCC [7]. Budgets are expressed in channel uses and SNR, and the encoder and decoder are trained jointly for a channel model. These systems can be robust to SNR variation. However, they are less interpretable and expose fewer explicit, user-controlled allocation knobs than SPIFF-like pipelines.

In summary, SPIFF is used here because it provides an explicit and interpretable allocation interface under a fixed decoder, which matches the project constraints and enables systematic analysis of encoder-side policies. The remainder of this chapter, therefore, focuses on SPIFF-like pipelines and on the two encoder challenges studied in this thesis: semantic ranking for ROI selection and budgeted RONI evidence sampling.

2.3 Problem Setting: Selective Fidelity for Semantic Image Transfer

Selective fidelity semantic transmission departs from the assumption that all image regions should be conveyed with uniform accuracy. Instead, it formalizes the idea that, under resource constraints, communication resources should be concentrated on content that is most relevant to the intended semantics or task (ROI), while the remaining regions (RONI) can be represented more coarsely and, if needed, reconstructed at the receiver [1, 5]. This section introduces the fidelity notions underlying selective transmission, reviews budget models used to enforce resource constraints, and outlines the main failure modes of sparse communication, motivating the need for principled evidence selection.

2.3.1 Pixel fidelity vs semantic fidelity under resource constraints

Conventional image transmission is commonly framed as a pixel-centric reconstruction problem, in which the objective is to minimize a distortion measure defined directly in the image space. Under strict bandwidth, this viewpoint leads to global quality degradation that is largely indiscriminate with respect to semantic relevance. In contrast, semantic image transfer prioritizes the preservation of information that supports perception or downstream understanding, even when this implies non-uniform pixel accuracy across the scene [1]. From this perspective, fidelity is not evaluated solely by how closely the decoded pixels match the original image, but also by whether the reconstructed content maintains the task-relevant structure and meaning.

SPIFF embodies this semantic-centric viewpoint by explicitly separating high-relevance content from the remaining regions and allocating different levels of transmission effort accordingly [1]. A conceptually related direction is provided by mask-based dual-mode transmission schemes for segmentation tasks, where the transmitted representation is explicitly designed to support a task-level semantic objective rather than uniform pixel reconstruction [5]. Despite the different architectural choices, both this line of work and selective-fidelity frameworks share a common principle: under resource constraints, communication resources should be allocated according to semantic relevance, balancing the preservation of critical content against the cost of conveying the full visual signal [1].

2.3.2 Budget models: bitrate vs coverage/patch budget

Resource constraints can be expressed through different, yet complementary, budget models. In communication-oriented settings, the most direct constraint is the available bitrate, which limits the amount of information that can be transmitted per image. Selective fidelity methods address this constraint by restricting the explicitly transmitted content to a subset of the image (e.g., an ROI and sparse RONI evidence) and by compressing each component within the overall rate budget [1]. Mask-based representations also naturally support budgeted transmission, as they enable a separation between structured semantic information and appearance information, with a mode selection that can be adapted to the task [5].

For algorithmic design and empirical analysis, rate constraints are often re-expressed as

a *coverage* or *patch* budget, i.e., a constraint on the total number of patches or on the fraction of RONI area that can be preserved. This abstraction is particularly convenient in pipelines that encode sparse evidence as a set of local observations later recomposed at the receiver, as in SPIFF [1]. Under a patch budget, the core algorithmic question is where to place a limited number of observations to maximize their utility for reconstruction.

In this thesis, evidence placement is studied through a set of encoder-side selection strategies under a strict patch/coverage budget. Progressive strategies such as farthest point sampling (FPS) promote dispersion by iteratively selecting points that maximize the minimum distance to the current set [8]. Determinantal point processes (DPPs) formalize relevance–diversity trade-offs through negatively correlated subset distributions [9]. In addition, a score-guided covering strategy (SGC) is considered to combine cue-driven relevance with probabilistic patch sampling under budget. These strategies are evaluated as abstract placement baselines rather than as end-to-end communication schemes. They provide a controlled interface to compare dispersion, relevance, and robustness properties under the same nominal constraint.

2.3.3 Failure modes of sparse transmission and the need for principled evidence

Sparse transmission introduces characteristic failure modes that do not arise when the entire image is communicated. First, when large regions are not explicitly transmitted, the receiver-side completion problem becomes inherently ambiguous: multiple plausible reconstructions may be consistent with the preserved content, leading to variability and potential semantic drift. Second, if the sampled evidence is spatially clustered or biased toward visually uninformative regions, the completion model may lack the constraints needed to preserve global structure, resulting in geometric inconsistencies or unnatural textures. Third, a mismatch between the preserved evidence and the true scene layout can create discontinuities at the boundaries between transmitted and synthesized regions, degrading perceptual coherence.

These failure modes motivate encoder-side policies for sampling sparse evidence that is both informative and well distributed under budget, as reviewed in Sec. 2.5 and Sec. 2.6.

2.4 ROI Selection and Semantic Ranking

ROI selection in SPIFF-like selective-fidelity pipelines identifies candidate regions and assigns the region(s) to be preserved at high fidelity [1]. This encoder-side step determines which content is transmitted explicitly as ROI and which is delegated to RONI evidence and generative reconstruction. Operationally, it comprises two stages:

- **Semantic candidate generation (ROI extraction):** generation of candidate Regions of Interest (ROIs);
- **Semantic ranking and ROI assignment:** assignment of relative importance scores to candidates to select a consistent ROI (or a set of ROIs).

The following discussion reviews representative approaches for ROI extraction and for semantic ranking, focusing on two main methodologies: task-agnostic prioritization based on semantic consistency, and task-oriented prioritization guided by task-specific attribution signals derived from a downstream objective.

2.4.1 Segmentation/ROI extraction as a precursor to selective fidelity

Selective fidelity requires a spatial decomposition into candidate units on which allocation can operate. In SPIFF, this role is fulfilled by a segmentation backend that produces masks and bounding boxes used to define the ROI and to support subsequent encoder decisions [1]. More generally, ROIs can be obtained through open-vocabulary detection followed by promptable segmentation; the Grounded SAM [10] paradigm is a representative example that combines an open-world detector with Segment Anything [11] to generate label-consistent masks from textual queries (as discussed in [1]). Once candidates are available, minimum-area filtering and overlap handling are commonly applied to remove unstable micro-segments and near-duplicates, improving downstream ranking consistency.

2.4.2 Task-agnostic ranking: semantic consistency and region prioritization

Task-agnostic ranking prioritizes semantically central regions without assuming a specific downstream objective. A common strategy embeds each candidate's ROI into a transferable representation space and scores candidates by semantic consistency with a reference image

representation. CLIP provides a widely adopted visual encoder whose embeddings enable similarity-based scoring [12]. Each ROI is mapped to an embedding through the CLIP image encoder, and relevance scores are obtained from image-to-image similarity (e.g., cosine similarity) between the ROI embedding and a reference embedding computed from the full image. This criterion is label-set agnostic and supports consistent encoder-side ROI determination [1].

2.4.3 Task-oriented ranking: task-specific attribution signals and model-driven evidence

Task-oriented ranking prioritizes regions according to their expected utility for a specified visual downstream task. Model-driven attribution can operationalize this idea by producing a spatial relevance signal that quantifies the sensitivity of a task score to localized perturbations. In this thesis, face recognition is adopted as the reference downstream task; in this context, attribution approaches such as Canonical Saliency Maps illustrate how evidence used by deep face models can be localized and interpreted [13]. In addition, identity is commonly represented through discriminative embeddings; ArcFace is a standard formulation that yields separable face embeddings via an angular-margin objective [14]. Within a selective-fidelity pipeline, task-oriented ranking can be implemented by computing a face-relevance attribution signal through occlusion sensitivity in the embedding space and aggregating this signal over each candidate ROI mask, yielding segment scores aligned with the downstream objective and enabling ROI assignment for explicit transmission.

2.4.4 Emerging direction: prompt-based semantic ranking

More recently, prompt-based vision-language models have emerged as flexible multimodal reasoners with strong image-conditioned instruction-following capabilities [15, 16]. Although not originally designed for ROI ranking, these models suggest a possible route toward semantic prioritization by eliciting image–text-conditioned relevance judgments through prompting. In the present work, this direction is considered only as an exploratory extension and is not part of the main comparative study.

2.4.5 Practical issues: overlap handling and ranking stability

Candidate generation often produces partially redundant ROIs (e.g., near-identical masks from overlapping detections). Without explicit handling, redundancy can destabilize semantic ranking by spreading high scores across duplicates and increasing sensitivity to minor segmentation variations. Overlap-aware suppression addresses this issue by removing or down-weighting candidates with high overlap. Soft-NMS is a representative refinement that penalizes scores as a function of overlap rather than applying hard removal [17]. In mask-based settings, analogous filtering can be implemented via mask overlap criteria (e.g., IoU and area-ratio checks) to retain a single representative among near-duplicates.

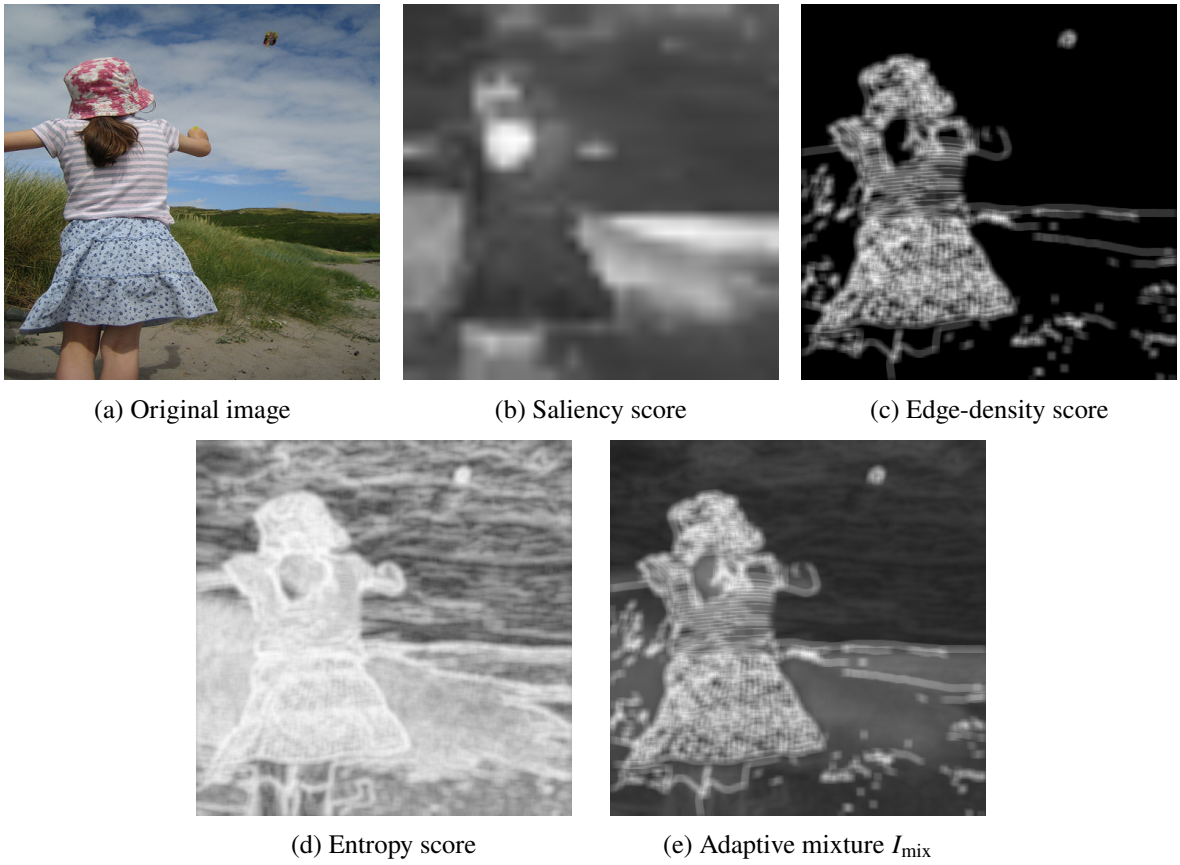
Ranking stability also depends on how continuous scores are mapped to discrete importance levels and on how sparse or low-contrast cases are handled. In selective-fidelity pipelines, enforcing that at least one candidate attains the maximum importance level helps prevent degenerate allocations in which the ROI becomes empty. Together with conservative duplicate suppression and minimum-area filtering, such top-rank enforcement yields more predictable ROI assignment across images and reduces failure cases. Overall, these design choices act as practical stabilizers that make semantic ranking a reliable interface for encoder-side allocation in SPIFF-like pipelines.

2.5 RONI Evidence Sampling: Informativeness Measures

RONI evidence sampling can be formulated as the estimation of a spatial informativeness measure $I(u)$ over the RONI domain Ω_{roni}, subsequently discretized into a budgeted set of K patches (or an equivalent coverage budget). Cue scores indicate where evidence is potentially useful, whereas discretization under budget and mechanisms enforcing dispersion/diversity are addressed in Sec. 2.6. Fig. 2 reports an illustrative example of the cue scores discussed in this section and of their fusion into an adaptive mixture I_{mix}; the individual cue families are detailed in Sec. 2.5.2, 2.5.3, and 2.5.4, while their fusion is discussed in Sec. 2.5.5.

2.5.1 Why sparse evidence can anchor generative reconstruction

Receiver-side reconstruction from partial observations is ill-posed, as multiple plausible scenes can match the preserved content. Sparse RONI evidence mitigates this ambiguity



(a) Original image (b) Saliency score (c) Edge-density score

(d) Entropy score (e) Adaptive mixture I_{mix}

Figure 2: Illustrative cue scores for RONI informativeness and their fusion into an adaptive mixture I_{mix} (brighter indicates higher informativeness). The cue visualizations are computed within the present work and reported for explanatory purposes.

by providing localized constraints that anchor synthesis and reduce semantic or geometric drift. The encoder-side objective is therefore to maximize the utility of each transmitted observation under budget while avoiding overly clustered evidence configurations.

2.5.2 Texture/uncertainty cues: local entropy

Texture-oriented cues target regions whose appearance is locally heterogeneous and thus difficult to reconstruct reliably without direct observation. A common proxy for local heterogeneity is *local entropy*, computed on a neighborhood of pixels $\mathcal{N}(u)$ by estimating a discrete distribution of values and measuring its uncertainty:

$$H(u) = -\sum_i p_i(u) \log p_i(u),$$

i.e., the Shannon entropy of a discrete distribution [18]. Here, $p_i(u)$ denotes the empirical histogram of quantized intensities (or luminance) within $\mathcal{N}(u)$, normalized such that $\sum_i p_i(u) = 1$. High entropy is usually associated with fine-grained textures (e.g., foliage, gravel, patterned surfaces) and clutter, while low entropy corresponds to smooth or slowly varying regions (e.g., sky, walls). In patch-based selection, $H(u)$ is typically averaged over each candidate patch to yield a patch score, biasing the budget toward areas where the reconstruction model is more likely to produce inconsistent micro-structure if left unconstrained. At the same time, entropy may over-emphasize noise, compression artifacts, or high-frequency patterns that are not semantically meaningful; as a consequence, texture cues are most effective when complemented by cues that encode geometric structure or perceptual prominence.

In the illustrative example of Fig. 2d, entropy remains relatively high over most textured and structured regions, whereas lower values appear mainly in the sky, whose more uniform appearance produces lower local uncertainty.

2.5.3 Structure cues: edges and edge density

Structure cues aim to preserve geometric constraints such as contours, boundaries, and linear features, which strongly influence perceptual coherence and spatial alignment in reconstructed images. Classical edge detection provides a principled way to extract such constraints. In particular, the Canny detector was designed to balance good detection, accurate localization, and suppression of multiple edge responses, yielding thin and well-localized edges under controlled noise sensitivity [19]. To convert edges into an informativeness signal compatible with patch-based transmission, a standard approach is to compute a local *edge density* map, i.e., the proportion (or count) of edge pixels within a neighborhood (equivalently, a local average of the binary edge map). This signal can then be pooled over candidate patches to obtain patch-level scores. Edge density highlights regions rich in contours, corners, and junctions, where sparse observations can stabilize completion by constraining layout and structure that would otherwise be underdetermined. However, edge-based cues may underrepresent homogeneous regions whose appearance is important for realism, and can become sensitive under low contrast or blur; these limitations further motivate the use of multiple, complementary cue families.

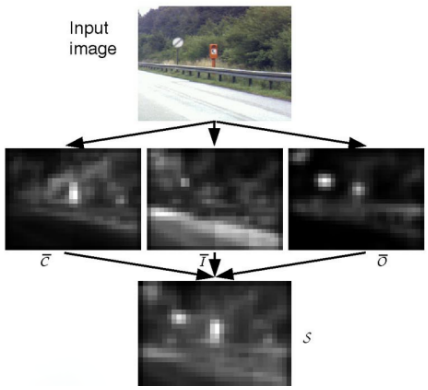


Figure 3: Construction of feature conspicuity maps (color $\bar{c}$, intensity $\bar{I}$, orientation $\bar{o}$) and their fusion into the saliency output S. Adapted from [2].

Fig. 2c shows that higher edge-density values occur around the child and along the main scene boundaries, where contours and intensity transitions are more frequent.

2.5.4 Attention cues: classical and deep saliency

Attention-oriented cues prioritize regions that are likely to be perceptually prominent, under the assumption that reconstruction errors in such regions are disproportionately noticeable. Classical bottom-up saliency models implement this idea through multi-scale center–surround contrasts and feature competition across channels. The formulation by Itti *et al.* is a representative example, integrating intensity, color opponency, and orientation information into a unified saliency signal that emphasizes visually distinctive locations [2]. In selective-fidelity transmission, such signals can be used to allocate sparse evidence to RONI regions that attract visual attention, thereby improving perceived plausibility even when large portions of the image are synthesized. Fig. 3 illustrates the feature-channel conspicuity maps and their fusion into the saliency output S, which provides an attention-oriented informativeness signal for RONI evidence sampling. In the Itti formulation, these conspicuity maps summarize three complementary feature families, intensity, color opponency, and orientation, combined through multi-scale center–surround contrasts; a more detailed description of the adopted formulation is provided in Sec. 3.3.2. In this context, bottom-up saliency provides a task-agnostic attention cue that complements texture- and structure-based informativeness signals.

As shown in Fig. 2b, the highest saliency values are concentrated around the child, corresponding to the most visually prominent region in the scene.

2.5.5 Adaptive mixtures of heterogeneous cues

No single cue family is expected to dominate uniformly across images: texture cues are beneficial in highly patterned regions, structure cues constrain geometry, and attention cues reflect perceptual or model-driven relevance. Consequently, RONI evidence sampling often relies on combining heterogeneous cue scores into a single sampling signal that balances these complementary objectives. A common and interpretable strategy is to use a convex mixture,

$$I_{\text{mix}}(u) = \sum_k \lambda_k I_k(u), \quad u \in \Omega_{\text{roni}}, \quad \lambda_k \geq 0, \ \sum_k \lambda_k = 1,$$

where each I_k is normalized (e.g., to $[0, 1]$ or to unit mass) before fusion. The resulting score field is typically converted into a sampling distribution $p(u) \propto I_{\text{mix}}(u)$, and patch scores are obtained by pooling $p(u)$ over candidate patches.

After fusion, the map in Fig. 2e emphasizes the child while still retaining part of the surrounding structural and textured context. This example illustrates how cue fusion can preserve a balanced emphasis between subject prominence and contextual structure.

In budgeted patch sampling, adaptive weighting is particularly relevant because cue scores can be highly peaked or scene-dependent; without adaptation, sampling may collapse onto a small region, reducing spatial coverage and weakening the anchoring effect of sparse evidence. By modulating cue contributions based on simple statistics (e.g., score concentration or entropy), the mixture can trade off local informativeness against spatial spread, yielding evidence sets that are more robust across scene types and more stable under strict budgets. While cue fusion controls informativeness, it does not by itself guarantee spatial dispersion; explicit diversity/coverage mechanisms are therefore required under strict budgets (Sec. 2.6). In SPIFF-like architectures, such robustness is essential because the generative completion stage is conditioned on sparse observations and can be sensitive to the exact evidence configuration [1].

2.6 Patch Sampling Under Budget: Coverage, Diversity, and Selection Strategies

Patch sampling transforms a continuous informativeness signal into a discrete subset of transmitted patches under a strict budget. The goal is to sample patches that are informative, spatially well distributed, and stable under limited budgets.

2.6.1 From informativeness measures to discrete patch sets

Let M denote the RONI mask, i.e., the region eligible for evidence transmission, and let $I(u)$ be an informativeness measure defined over pixel locations u (e.g., derived from entropy, edge density, saliency, or combinations of these cues). The RONI is discretized into a finite candidate set $\mathcal{P} = \{p_i\}_{i=1}^{N}$ of patches of fixed size (or from a small set of sizes). Each candidate patch p_i is characterized by a center location x_i and an aggregated score s_i, obtained by pooling cue values within the patch support, e.g.,

$$s_i = \mathrm{Agg}_{u \in p_i} I(u),$$

where $\mathrm{Agg}(\cdot)$ denotes an average or a sum. The budget can be expressed either as a cardinality constraint $|S| \leq K$ (at most K patches) or as a coverage constraint $\sum_{p_i \in S} \mathrm{area}(p_i) \leq A_{\max}$ (at most a fraction of RONI area), where $S \subseteq \mathcal{P}$ is the sampled subset.

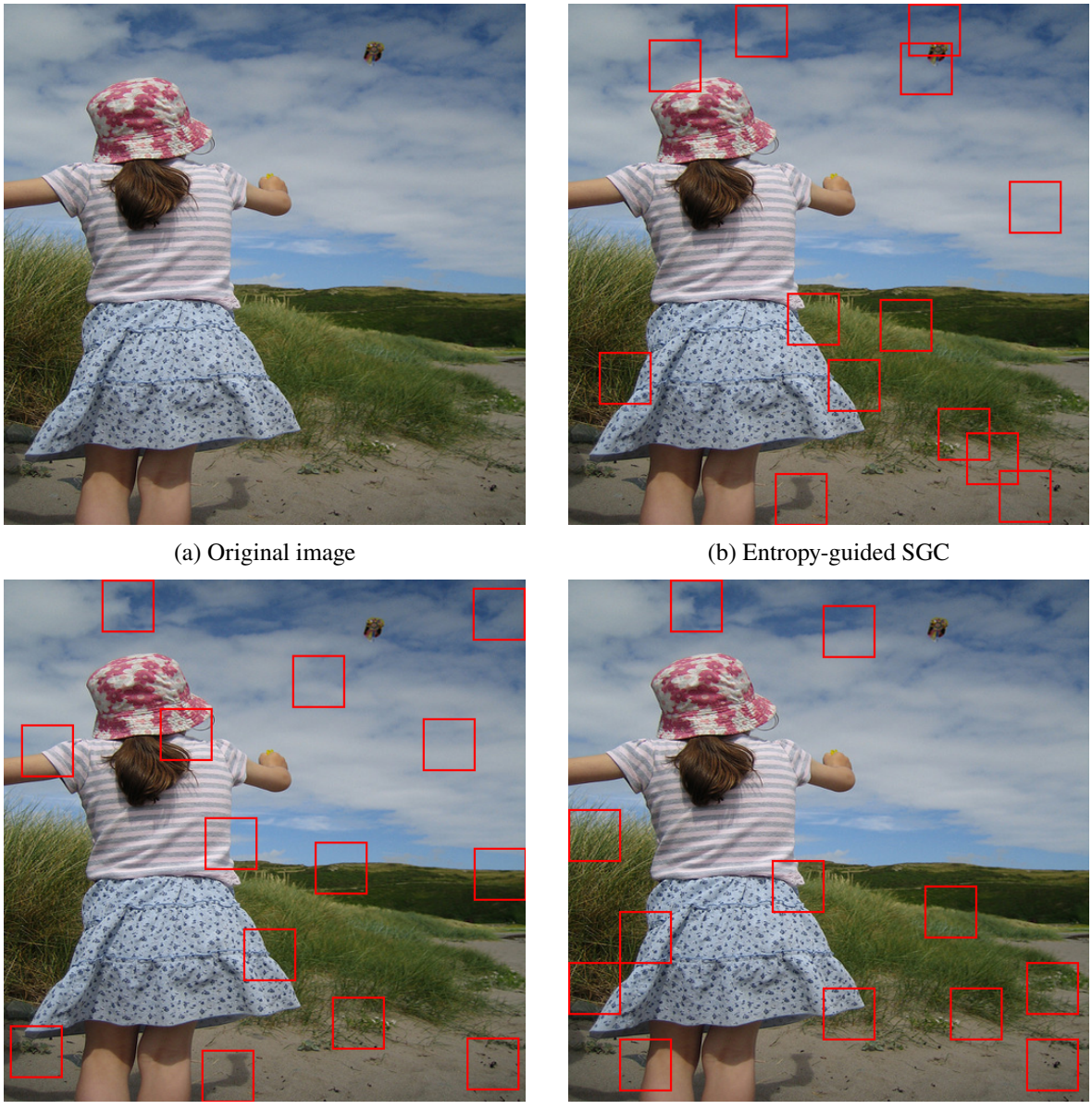
In this thesis, patch selection strategies are evaluated through the three selector families implemented and analyzed in the methodology:

- **SGC** (Score-Guided Covering with density regularization);
- **FPS** (Farthest Point Sampling; evaluated in *deterministic* and *probabilistic* variants for coverage-driven sampling);
- **DPP** (Determinantal Point Processes; evaluated both as *deterministic* MAP selection and as *probabilistic* DPP-based sampling for relevance–diversity).

These families represent complementary ways to trade off cue relevance (from $\{s_i\}$) and spatial dispersion/coverage under the same strict budget, while enabling a direct comparison

between deterministic and probabilistic evidence configurations. For completeness, two classical references are also recalled (score-only Top-K and blue-noise Poisson-disk sampling), although they are *not included* in the experimental pipeline of this thesis.

Fig. 4 illustrates cue-to-patch discretization under a fixed budget on an entropy-derived score field, contrasting representative sampling policies implemented in the present work.



(a) Original image

(b) Entropy-guided SGC

(c) Entropy-guided FPS (deterministic)

(d) Entropy-guided DPP (deterministic)

Figure 4: Patch sampling under a fixed budget on an entropy-derived cue, generated by the proposed implementation. Red boxes denote sampled patches produced by SGC and by two deterministic alternatives (FPS and MAP-DPP). The figure is illustrative and not a performance comparison.

Patch sampling can be viewed as a relevance–diversity optimization problem: relevance is

captured by the scores $\{s_i\}$, while diversity is enforced through spatial repulsion or similarity kernels over patch features. This abstraction provides a common interface for score-guided probabilistic covering, progressive coverage-driven sampling (FPS), and determinantal point processes (DPP).

2.6.2 Score-guided probabilistic covering with density regularization (SGC)

A class of probabilistic patch selection strategies samples patch centers from a distribution derived from the informativeness measure. In this setting, cue scores guide exploration toward high-utility regions, while a density or residual-coverage factor discourages repeated sampling in already covered areas. Such approaches provide a simple mechanism to trade relevance against spatial spread under strict budgets and to generate multiple plausible evidence configurations for robustness analysis.

2.6.3 Progressive and coverage-driven sampling: farthest point strategies (FPS)

Progressive sampling strategies select points (or patch centers) iteratively to maximize coverage of the admissible region. The farthest point strategy selects, at each iteration, the candidate whose distance to the current set is maximal, thereby maximizing the minimum distance between selected points and promoting uniform spatial dispersion. This principle was studied for progressive image sampling, where each new sample is placed at the farthest location from previously selected samples to improve coverage as the sampling set grows [8]. In the patch-based setting, the same idea can be applied by treating patch centers as sampling points and restricting candidates to lie within M, yielding a simple deterministic procedure that produces well-spread evidence under increasing budgets.

Farthest point strategies have two properties that are advantageous for budgeted transmission. First, they are inherently *progressive*: a solution for budget K contains the solution for any smaller budget $K' < K$, which enables monotone rate–quality operating points. Second, they directly control spatial coverage through a geometric criterion, making them effective when the main failure mode is leaving large RONI areas unconstrained. Their main limitation is that dispersion is enforced independently of patch relevance unless the candidate set or the distance function is modulated by the cue signal (e.g., by restricting the search to high-score candidates or by using a weighted distance), which motivates hybrid relevance–coverage

variants in practice.

2.6.4 Determinantal point processes for relevance–diversity trade-offs (DPP)

Determinantal point processes (DPPs) provide a probabilistic framework for subset selection that favors diversity through negative correlations among selected items. For a ground set of candidates $\mathcal{P}$, a DPP defines a distribution over subsets S such that $\mathbb{P}(S) \propto \det(L_S)$, where L is a positive semidefinite kernel matrix, and L_S is the principal submatrix indexed by S [9]. Intuitively, the determinant increases when selected items are mutually dissimilar under the kernel, thereby encouraging diverse selections. Relevance can be incorporated by factorizing the kernel as $L_{ij} = q_i q_j k_{ij}$, where q_i encodes patch quality (e.g., derived from cue scores) and k_{ij} encodes similarity (e.g., spatial proximity or feature similarity). Under this construction, high-quality patches are preferred, but redundant patches are penalized if they are too similar.

DPPs are attractive for patch selection because they unify relevance and diversity in a single model and allow both sampling-based and optimization-based selection. In particular, the maximum a posteriori (MAP) subset under a DPP provides a deterministic selection that approximates the most diverse high-quality set, while random sampling from the DPP yields probabilistic selections whose expected properties can be controlled through the kernel design [9]. The main practical challenge is computational: exact sampling and MAP inference can be expensive for large candidate sets, motivating approximate methods (e.g., greedy MAP approximations and low-rank kernel constructions) when the candidate patch grid is dense.

2.6.5 Classical references (not included in experiments)

Blue-noise baseline (Poisson-disk). Poisson-disk sampling is a canonical blue-noise mechanism that enforces dispersion by requiring a minimum separation r between sampled centers [20]. Efficient constructions of maximal Poisson-disk sets have been proposed to improve coverage and scalability [21]. Its role largely overlaps with dispersion/coverage-driven placement as represented by FPS in this thesis.

Score-only reference (Top-K). A common reference is to sample the top-K patches by score $\{s_i\}$. This relevance-only strategy is not included in the experimental pipeline of this

thesis, as it typically clusters evidence when $I(u)$ is peaked, and does not explicitly control spatial coverage or diversity.

This baseline is recalled for completeness, but it is *not included* in the experimental pipeline of this thesis.

2.6.6 Deterministic vs probabilistic selection and robustness considerations

Under a fixed budget, patch sampling can be performed deterministically (e.g., deterministic FPS, MAP-DPP) or probabilistically (e.g., SGC, probabilistic FPS, and DPP sampling). Deterministic strategies offer reproducibility and simplify controlled ablations because the sampled set is uniquely determined by the input and the policy. However, determinism may amplify biases of the cue signal, for example, by repeatedly sampling patches in the same visually dominant areas while neglecting alternative plausible anchors that could improve reconstruction in different scenes. Sampling-based strategies mitigate this issue by exploring multiple evidence configurations and by producing diversified patch sets even when the cue signal is highly peaked; this can improve average performance and provide a natural mechanism to quantify uncertainty.

In SPIFF-like pipelines, robustness is relevant because the receiver-side generative reconstruction can be sensitive to the exact evidence configuration, especially under very sparse budgets [1]. For sampling-based selectors, performance should therefore be characterized not only by the mean outcome but also by variability across random seeds, so that stability can be assessed in addition to fidelity. Conversely, deterministic policies can be evaluated for robustness by analyzing their sensitivity to perturbations of the cue signal, to changes in patch size, or to small variations in budget. These considerations motivate sampling mechanisms with the following properties:

- explicit enforcement of dispersion (coverage);
- guidance by relevance;
- robust behavior under randomness.

Such properties are jointly addressed by coverage-driven sampling strategies (e.g., farthest-point selection) [8], relevance–diversity models such as DPPs [9], and score-guided probabilistic covering schemes with explicit coverage regularization (as considered in this thesis).

This perspective suggests that patch selection strategies should be compared not only in terms of reconstruction fidelity, but also in terms of the geometry and stability of the sampled evidence under a common budget.

2.7 Decoder-Side Reconstruction as Context: Generative Inpainting and Conditioning

The receiver reconstructs a complete image from partial observations through conditional generative inpainting. Reconstruction quality depends on the conditioning signals available (preserved regions, sparse evidence, and spatial constraints) and becomes sensitive to evidence configuration at low budgets.

2.7.1 Inpainting from partial observations: constraints and ambiguity

Reconstruction from partial observations is inherently ambiguous, as the same preserved content can be consistent with multiple plausible reconstructions in unobserved regions. Generative inpainting mitigates this ill-posedness by exploiting a learned prior over natural images and producing reconstructions that satisfy the observed-pixel constraints under the adopted conditioning signals [22]. In the SPIFF reference framework, the fixed receiver-side reconstruction module is instantiated with Stable Diffusion 2, adopted as the latent diffusion model (LDM) for conditional image completion [1]. Diffusion-based models provide an effective mechanism for conditional generation by progressively denoising toward an image while enforcing consistency with the available context [3]. Under sparse evidence, reconstruction quality becomes sensitive to which pixels are preserved, linking reconstruction stability to encoder-side ROI selection and RONI evidence sampling [1].

Figure 5 provides the architectural context of the latent diffusion decoder underlying Stable Diffusion 2. Figure 6 then illustrates this effect on an example from the implemented pipeline, by comparing the original image to the reconstruction obtained under sparse evidence with deterministic FPS guided by the adaptive mixture informativeness measure.

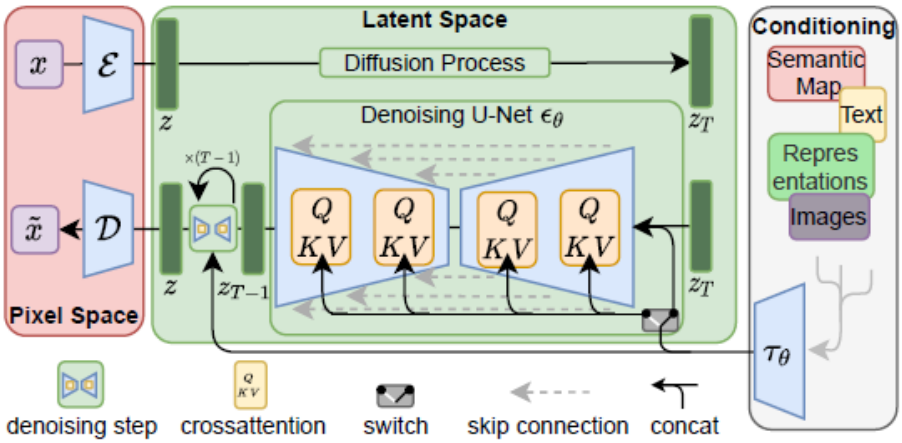
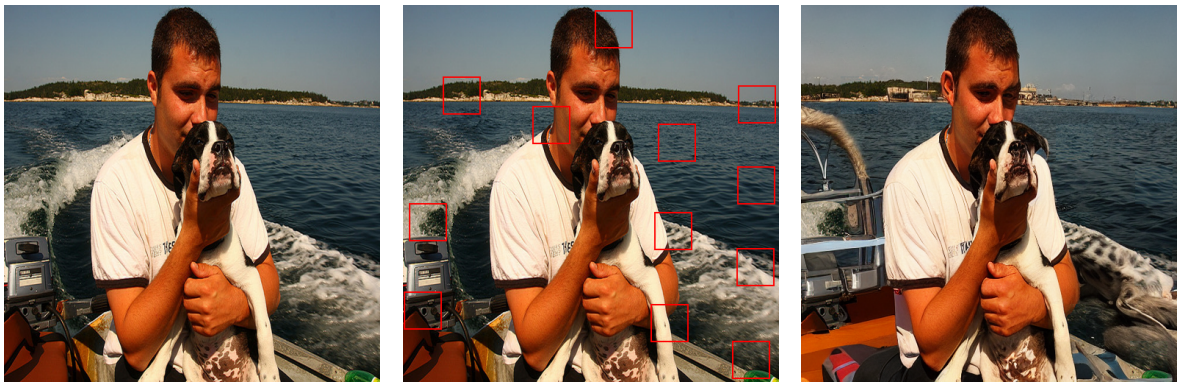


Figure 5: Latent diffusion architecture underlying Stable Diffusion 2. The input image is encoded into a latent space, processed by a denoising U-Net, and decoded back to pixel space, with conditioning injected through concatenation or cross-attention. Adapted from [3].



(a) Original image.

(b) Original image with sampled RONI patches highlighted, selected by deterministic FPS under the adaptive mixture informativeness measure.

(c) Reconstruction of the missing RONI content obtained by conditioning the decoder on the sampled patches as sparse anchors for completion.

Figure 6: Qualitative example of RONI reconstruction in a SPIFF-like pipeline. From left to right: original image, original image with the sampled RONI patches selected for transmission, and reconstructed image in which the missing RONI content is completed by the decoder using the sampled patches as sparse conditioning evidence.

2.7.2 Conditioning sources: masks, patch grids, and optional text guidance

In SPIFF, receiver-side completion is conditioned on a recomposed partial observation obtained by integrating transmitted ROI content with sparse RONI evidence and spatial side information [1]. Masks and metadata specify which pixels are observed and where transmitted patches are located, converting the bitstream into structured constraints for the inpainting model. From a modeling viewpoint, this corresponds to constrained generation, where ob-

served regions impose hard pixel constraints, and the remaining content is synthesized to be consistent with them [3]. When available, auxiliary semantic guidance (e.g., a caption) can further bias synthesis toward text-consistent content, although its impact depends on the informativeness of the text and its alignment with the preserved visual evidence [1].

2.7.3 System trade-offs: receiver compute vs link rate

Selective-fidelity architectures exchange communication resources for receiver-side computation by transmitting only high-utility visual information and delegating the remainder to generative reconstruction [1]. The operating point is governed by the amount of explicitly transmitted content (ROI extent and evidence budget) and by the complexity of the generative model. Increasing transmitted evidence typically reduces ambiguity and improves stability, whereas relying more heavily on inpainting reduces bitrate at the cost of higher compute demand and potentially greater variability in reconstructed regions. Within this context, encoder-side policies that prioritize informative and well-distributed evidence are central to achieving favorable trade-offs without modifying the decoder [1].

2.8 Evaluation Protocols and Metrics for Selective Fidelity

Selective-fidelity architectures require evaluation protocols that cover both encoder decisions and end-to-end reconstruction under a budget. Evaluation is organized around two components:

- **Encoder-side agreement:** consistency of semantic ranking (ROI selection) against manual annotations;
- **End-to-end reconstruction:** perceptual and semantic fidelity of the decoded image under a fixed decoder, for patch sampling strategies compared under the same nominal budget.

Both components are evaluated through dedicated metrics. Encoder-side agreement is reported for the semantic ranking stage. Fidelity, patch-set geometry, and runtime indicators are reported for the end-to-end reconstruction stage. They are used to interpret outcomes in terms of redundancy, dispersion, and computational cost.

2.8.1 Encoder-side agreement for semantic ranking

Encoder-side semantic ranking is evaluated at the category level. Predicted ranks are compared against manually annotated ground truth. A rank value of 0 denotes an *ignored* state. Pairs with $G(s, \ell) = 0$ or $P(s, \ell) = 0$ are excluded from all metrics. Agreement is computed for ranks $r \in \{1, 2, 3, 4\}$ using a one-vs-rest decomposition. Precision, recall, and F_β are reported per rank. Results are also aggregated with rank-dependent weights. Both weighted macro and weighted micro summaries are reported. A ground-truth to prediction transition table is also accumulated over ranks $\{1, 2, 3, 4\}$. It highlights systematic confusions between rank levels.

2.8.2 Perceptual fidelity metrics

Perceptual fidelity targets human-aligned similarity between an original image and its reconstruction. This is relevant when completion modifies pixel-level details but preserves plausible appearance. Learned perceptual metrics compare images in deep feature space. Learned Perceptual Image Patch Similarity (LPIPS) compares corresponding deep features extracted at multiple layers of a visual network, normalizes them in the channel dimension, and aggregates their discrepancies across spatial locations and layers into a perceptual distance [23]. In this thesis, the LPIPS distance is mapped to a similarity-oriented score as $\mathrm{PF} = 1 - \mathrm{LPIPS}$. The score is clipped to $[0, 1]$ for numerical stability. This mapping supports comparisons across evidence placements. Pixel-wise measures can be overly sensitive to benign variations introduced by inpainting.

2.8.3 Semantic fidelity metrics

Semantic fidelity targets preservation of high-level content and meaning. Exact pixel correspondence is not required. A common approach embeds images into a semantic representation space. Reconstructions are scored via embedding similarity. Semantic fidelity is quantified through CLIP embedding similarity. It is mapped to $[0, 1]$ as $\mathrm{SF} = (\mathrm{sim} + 1)/2$, where sim is cosine similarity between normalized embeddings. This follows the CLIPScore perspective [24]. The metric complements perceptual fidelity. It emphasizes semantic consistency across reconstructions with different texture realizations.

2.8.4 Patch-set geometry: overlap and dispersion indicators

Reconstruction depends on sparse evidence. It depends on the evidence layout as well as the amount. Geometry indicators are therefore reported. Redundancy is summarized via patch overlap. It is defined as the fraction of patch-covered pixels that are covered more than once. Spatial spread is summarized via the mean pairwise Euclidean distance between patch centers. This measures the dispersion of anchors over the RONI. For probabilistic selectors, geometry indicators are summarized across random seeds. Mean and standard deviation are reported. This characterizes stability under the same nominal budget.

2.8.5 Runtime and complexity indicators

Selective fidelity introduces computational trade-offs across the transmitter and receiver. Encoder-side policies incur costs for cue computation, scoring, and subset selection. Decoder-side costs are dominated by generative inference. Runtime is therefore reported for both stages when available. Encoder runtime includes semantic ranking and patch sampling. Decoder runtime includes completion inference time. Within-sample ranking of PF and SF is also used in comparisons. It reduces sample-dependent scale effects. It supports aggregated comparisons through summary tables and performance-profile style curves.

2.8.6 Limitations and interpretation of PF/SF and geometry indicators

Learned similarity measures are appropriate when generative reconstruction modifies pixel-level details. They provide compact, comparable scores across configurations. They are not sufficient to characterize faithfulness in a sparse-conditioning regime. Different artifacts can yield similar PF/SF values. Geometry indicators are therefore reported alongside PF/SF. They support attributing fidelity changes to evidence redundancy and spatial coverage rather than to incidental completion effects.

Perceptual fidelity (PF). LPIPS correlates with human judgments in many natural-image settings. It is not a calibrated perceptual error [23]. Its scale depends on the chosen backbone and training protocol. Similar LPIPS values can correspond to different artifacts. Boundary discontinuities and texture inconsistencies are examples. LPIPS can also be insensitive

to some semantic substitutions when local appearance remains plausible. PF is therefore interpreted comparatively. It is not an absolute correctness measure.

Semantic fidelity (SF). CLIP similarity captures semantic alignment in a transferable embedding space. It can be biased toward salient objects [12, 24]. It can under-penalize layout errors when global semantics remain consistent. Visually plausible hallucinations can also obtain high similarity. SF is therefore interpreted as semantic consistency. It is complementary to PF. It does not guarantee faithfulness.

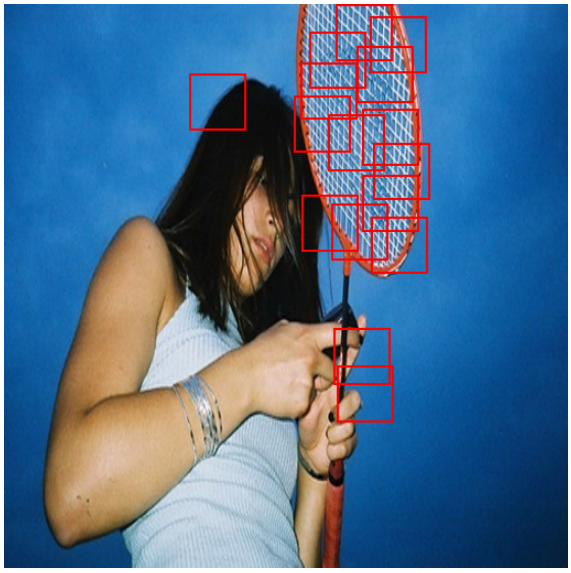
Why geometry metrics are necessary. With sparse conditioning, reconstruction depends on evidence layout and evidence quantity. Overlap and dispersion provide explanatory variables. High overlap suggests redundancy and wasted budget. Low dispersion indicates clustered evidence. Clustered evidence can leave large regions weakly constrained. Geometry metrics support attributing gains to improved evidence utility and coverage. They reduce reliance on incidental effects of the generative model.

Fig. 7 provides an illustrative comparison between a clustered evidence configuration and a more spatially dispersed one under a common nominal budget.

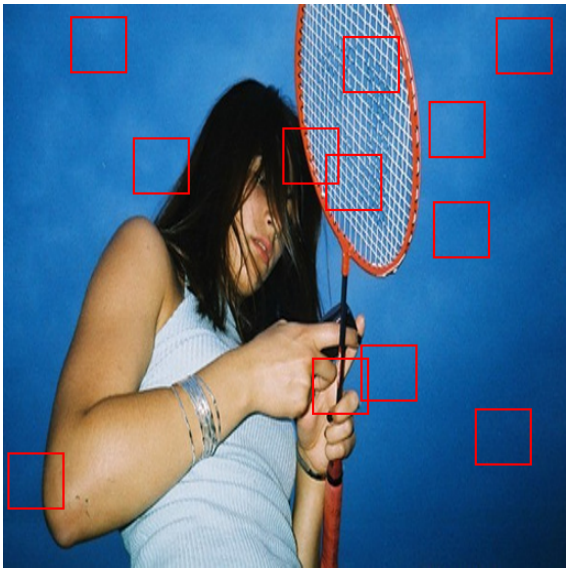
Robustness and aggregation. For probabilistic selectors, mean performance is insufficient. Variability across seeds is required. PF/SF should therefore be reported together with dispersion statistics across seeds. Within-sample ranking and normalized scores improve robustness across heterogeneous images.

2.9 Network Scenarios: Heterogeneous Terrestrial and Non-Terrestrial Links

Selective-fidelity semantic transfer targets settings in which communication resources are variable and often scarce, and in which receivers may have heterogeneous capabilities [1]. This section provides network-side context for the design choices considered in this thesis. The focus is on heterogeneous terrestrial and non-terrestrial links. Only aspects that motivate encoder-side allocation policies are discussed.



(a) Clustered evidence configuration for the illustrated sample, obtained with SGC guided by edge density. Patch overlap: 35.38%, mean center distance: 128.07 px



(b) More spatially dispersed evidence configuration for the illustrated sample, obtained with probabilistic FPS guided by entropy. Patch overlap: 1.95%, mean center distance: 258.05 px

Figure 7: Illustrative comparison between two sample-specific evidence layouts under a common nominal budget. The reported patch overlap and mean center distance values refer to the shown sample and to the corresponding selector–cue combinations. In this example, SGC guided by edge density produces a more clustered configuration, with higher overlap and more sampled patches needed to satisfy the target coverage budget, whereas probabilistic FPS guided by entropy yields a more spatially dispersed layout with lower overlap.

Fig. 8 illustrates a representative heterogeneous delivery scenario combining terrestrial access, non-terrestrial relaying, and distribution toward end users.

2.9.1 Heterogeneous links under time-varying channels

Heterogeneous delivery arises when the end-to-end path spans access technologies with different propagation conditions and resource availability. In terrestrial cellular links, variability is driven by mobility, interference, and scheduling. In non-terrestrial networks (NTN), especially LEO satellite segments, additional factors are relevant. These include long, time-varying propagation delays, Doppler shifts, and intermittency due to satellite motion and beam-coverage dynamics [25].

Channel and load variability are addressed through link-adaptation and reliability mechanisms. Examples include modulation and coding selection, retransmissions, and scheduling. In Non-Terrestrial Networks (NTN), longer round-trip times and stronger temporal varia-

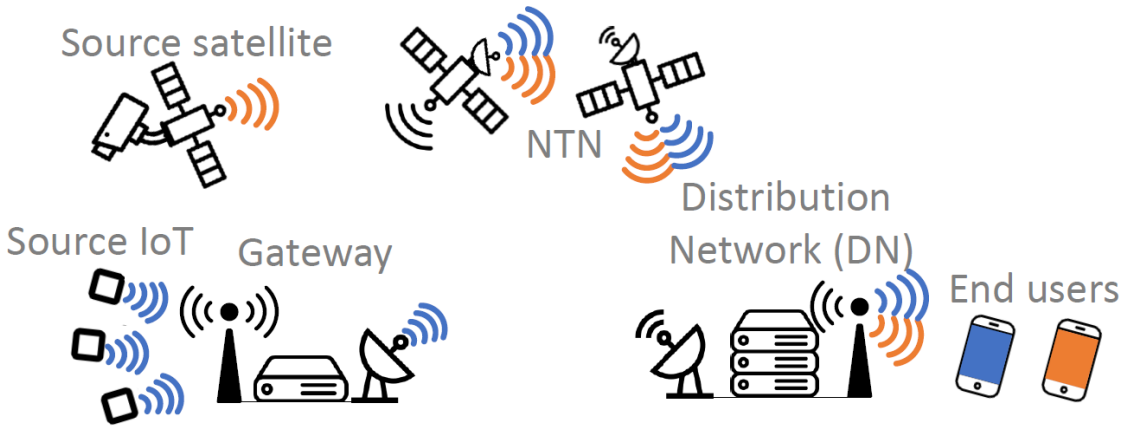


Figure 8: Illustrative heterogeneous terrestrial/NTN delivery scenario, including source acquisition, gateway access, NTN relaying, and end-user delivery.

tions affect these mechanisms and their effective operating points [25]. As a consequence, achievable bitrate and latency can vary over time and across users.

2.9.2 Implications for selective fidelity and semantic transfer

Under limited resources, transmitting the full visual signal at high fidelity can be inefficient. Latency constraints can further restrict payload size. Semantic communication addresses these regimes by prioritizing task-relevant meaning over uniform pixel fidelity [26]. In satellite-assisted networks, end-to-end performance also depends on network-layer decisions. Routing can be affected by the availability of semantic encoders and matched knowledge bases along the path [27].

Selective-fidelity architectures, as exemplified by SPIFF, implement this principle through explicit allocation [1]. High-relevance content (ROI) is preserved at higher fidelity, while the remaining regions (RONI) are conveyed through sparse evidence and completed at the receiver. Channel heterogeneity motivates encoder policies that preserve essential information and adjust RONI evidence sampling under a varying budget.

2.9.3 Design implications for encoder-side policies

In heterogeneous terrestrial/NTN links, channel variability translates into fluctuations of the application-layer budget. This affects both the amount of transmitted content and the effectiveness of sparse evidence delivery. Three implications are relevant for the encoder-side decision points studied in this thesis.

Budget-to-policy translation. A rate or payload constraint must be mapped to explicit allocation parameters. These include ROI extent/quality and a RONI evidence budget K. This mapping is treated at the level of operating points. Different values of K represent different evidence budgets under the same allocation interface.

Progressive operating points. Adaptation benefits from nested solutions across budgets. Progressive patch selection strategies support monotone operating points as K increases. This is useful under fluctuating capacity [8].

Robustness and overhead. In NTN, long and variable Round-Trip-Time (RTT) reduces the effectiveness of feedback and retransmissions [25]. Evidence placement should therefore avoid highly clustered anchors under sparse budgets. Diversity and coverage mechanisms are relevant in this regime [9]. Side-information overhead is also non-negligible. Masks, patch indices/coordinates, and optional guidance consume budget that competes with appearance evidence. Compact parameterizations and consistent overhead assumptions are therefore required.

2.10 Summary and Research Gaps

Building on the two challenges discussed above (semantic ranking for ROI selection and RONI evidence sampling), the following gaps summarize the main limitations in robustness and comparability that motivate the present work.

2.10.1 Gaps in semantic ranking under redundant and overlapping candidates

ROI selection is performed over candidate sets that are often redundant and ambiguous. Multiple segments can plausibly explain the same visual content at different granularities or under different labels [1]. Task-agnostic semantic ranking can be implemented by scoring candidate ROIs through similarity in a transferable embedding space. CLIP embeddings provide a standard instance of this approach [12]. Task-oriented semantic ranking can be based on attribution or saliency signals defined with respect to a downstream model [13]. When the downstream objective is face identity, the target signal can be expressed in a discriminative embedding space such as ArcFace [14]. Prompt-based vision-language

models also suggest a possible alternative route for semantic prioritization, but their role in selective-fidelity ROI ranking remains underexplored and has not yet been systematically compared against established task-agnostic and task-oriented approaches [15, 16].

2.10.2 Gaps in RONI evidence sampling under strict budgets

When RONI content is not transmitted exhaustively, reconstruction quality depends on whether sparse evidence provides effective anchors for completion [1]. Common cue families include texture/uncertainty measures (e.g., local entropy), structure cues (e.g., edges and edge density), and attention cues (e.g., classical saliency), which capture complementary notions of informativeness [18, 19, 2]. Under strict budgets, two limitations are central. First, cue scores alone do not specify how evidence should be spatially distributed to avoid large unconstrained areas; discretization and dispersion mechanisms therefore become essential [8, 9]. Second, cue effectiveness is scene dependent: texture measures may overemphasize noise-like frequencies, edge cues may underrepresent low-contrast but visually relevant regions, and attention measures may concentrate evidence into a small portion of the image. These limitations motivate evidence sampling strategies that combine heterogeneous informativeness measures with explicit diversity/coverage control, especially when completion is sensitive to the exact evidence configuration [1].

2.10.3 Need for systematic comparisons across informativeness measures and patch selection strategies

Many works evaluate cues and selection schemes in isolation, making it difficult to separate the contribution of the informativeness signal from that of discretization and diversity control. In selective-fidelity settings, this separation is critical. For a fixed informativeness measure, different patch selection strategies can produce markedly different geometric properties; conversely, for a fixed strategy, different measures can induce distinct failure modes. This motivates evaluation protocols that compare informativeness measures and patch selection strategies under a common budget and a fixed decoder, while reporting both reconstruction metrics and patch-geometry indicators so that fidelity differences can be interpreted in terms of redundancy and dispersion.

2.10.4 Thesis contributions and expected impact on PF/SF trade-offs

Within the SPIFF baseline, the decoder is treated as fixed, and improvements are pursued through encoder-side policies [1]. The present work contributes to both challenges. On the ROI side, it develops task-agnostic and task-oriented semantic ranking mechanisms based on transferable embeddings and face-relevance attribution, together with practical top-rank enforcement [12, 13]. It also includes a preliminary prompt-based vision-language ranking study [15, 16]. On the RONI side, it investigates evidence sampling based on heterogeneous informativeness measures and compares multiple patch selection strategies under strict budgets [18, 19, 2, 8, 9]. The expected outcome is an improved operating region in the perceptual–semantic trade-off space, measured through perceptual and semantic fidelity metrics (PF/SF) under fixed decoding conditions [23, 24]. Reconstruction metrics are complemented with patch-geometry and runtime indicators. This supports attributing gains to improved evidence utility and dispersion rather than to increased computational cost.

Semantic ranking is evaluated separately at the encoder. Agreement is reported through per-rank and aggregated indicators, so that improvements can be assessed independently of reconstruction fidelity.

3 Methodology

3.1 System overview

This chapter builds on the SPIFF reference framework introduced in Section 1 and specifies the concrete *encoder-side* instantiation adopted in this work. The receiver-side reconstruction is treated as a fixed black-box stage; therefore, methodological choices are restricted to how semantic relevance is estimated at the transmitter and how a compact set of observations is sampled under a rate constraint.

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the transmitter produces two forms of guidance that drive the subsequent stages of the SPIFF pipeline:

- an *ROI prior* derived from semantic candidates, used to determine which content is preserved explicitly as ROI via semantic ranking;
- a *sparse RONI evidence set* sampled from the remaining pixels, used to anchor generative reconstruction at the receiver.

Accordingly, the methodological core of the system is organized into the following encoder modules:

- **Semantic ranking pipeline** (Section 3.2), which extracts semantic candidates with a label-conditioned segmenter and assigns discrete relevance ranks on a common scale, enabling ROI assignment and category-level rank export.
- **RONI informativeness measures and patch sampling strategy** (Section 3.3), which computes an informativeness measure over the RONI and samples a limited set of localized patches, together with the minimal spatial side information required for spatial recomposition at the receiver.

In addition to these core encoder modules, an exploratory LLM-based ranking implementation is also considered and evaluated separately within the semantic-ranking study.

The experimental evaluation is conducted on a fixed image corpus. The following subsection summarizes the dataset and the splits used by the two encoder modules.

3.2 Semantic ranking pipeline

The semantic ranking stage assigns relevance scores and discrete ranks to the semantic candidates of the input image. Candidate regions are generated by a GroundingDINO+SAM [10] segmentation pipeline operating on a fixed set of labels. For each candidate mask, a segment-isolated view is constructed by retaining only the pixels within the mask and setting all remaining pixels to a constant background value. Candidate importance is then estimated via CLIP similarity between the original image embedding and the embedding of the segment-isolated view.

In parallel, a task-oriented ranking is computed on the same set of masks. An occlusion-sensitivity procedure generates a pixel-level face-relevance signal by measuring how masking localized patches affects the similarity of face embeddings. Candidate-level task scores are computed by aggregating this signal within each mask and mapped to discrete rank levels using metric-specific thresholds.

Overall, for each image, the main semantic-ranking pipeline outputs:

- a task-agnostic, CLIP-based semantic ranking;
- a task-oriented, face-relevance semantic ranking;

Both rankings are expressed on the same discrete scale of $K - 1$ non-zero levels and enforce the same top-rank constraint. In addition, a separate exploratory LLM-based ranking study is considered.

3.2.1 Segment ranking problem definition

Let $I \in \mathbb{R}^{H\times W\times 3}$ denote an input image and let $\mathcal{L}$ be the fixed semantic label set used by the segmentation backend. For each image, the segmenter returns a finite set of candidate regions

$$\mathcal{S} = \{s_i\}_{i=1}^{N}, \qquad s_i \triangleq (M_i, \ell_i, \beta_i, \sigma_i), \tag{1}$$

where $M_i \in \{0,1\}^{H\times W}$ is a binary mask, $\ell_i \in \mathcal{L}$ the predicted label, β_i the bounding box, and σ_i the detection confidence. Candidates with negligible support are removed by enforcing an area constraint $|M_i|/(HW) \geq \varepsilon_{\text{roi}}$, with ε_{roi} set in Table 1. Additionally, cross-label duplicate

suppression is applied to remove redundant overlapping candidates predicted under different labels.

The semantic ranking module assigns to each retained segment s_i a discrete importance level

$$r_i \in \{1, \ldots, K-1\}, \tag{2}$$

together with a consistent within-image ordering, where K is the configured number of rank levels and larger values correspond to higher importance. A top-rank constraint is enforced so that at least one segment attains the maximum level $K-1$.

Two complementary criteria are supported and computed on the same candidate set $\mathcal{S}$:

- **Task-agnostic (CLIP-based).** A score a_i is computed from the CLIP image-to-image similarity between the original image I and a segment-isolated view derived from M_i.
- **Task-oriented (face-relevance).** A score t_i is computed by aggregating a pixel-level face-relevance signal over the mask support M_i, where the signal is obtained via occlusion sensitivity of face-embedding similarity.

In both cases, the continuous scores $\{a_i\}$ and $\{t_i\}$ are mapped to the same discrete rank set $\{1, \ldots, K-1\}$ through metric-specific thresholding; the details of score construction, normalization, and discretization are given in the corresponding ranking subsections.

Default configuration and rationale. Unless otherwise stated, experiments adopt the default configuration reported in Table 1. These values reflect practical trade-offs: ε_{roi} filters out spurious micro-segments that are unlikely to be semantically stable; the maximum number of detections and backend threshold bound computational cost and reduce candidate noise; cross-label duplicate suppression uses a high intersection over union (IoU) threshold (with an area-ratio check) to remove only near-identical overlaps while preserving genuinely distinct regions. For semantic ranking, a small number of levels ($K = 5$) provides an interpretable discrete scale.

Discretization and top-rank enforcement. For either ranking criterion, let x_i denote the branch-specific continuous score assigned to candidate s_i (i.e., $x_i = a_i$ for task-agnostic CLIP scoring and $x_i = t_i$ for task-oriented face relevance). Discrete relevance ranks are obtained by

Symbol / item	Meaning	Default	Role in pipeline
ε_{roi}	minimum ROI area fraction	0.001	discard negligible masks before ranking
K	number of rank levels	5	discrete importance scale $\{1, \ldots, K-1\}$
–	backend acceptance threshold	0.35	minimum confidence for candidate generation by the segmentation backend
–	cross-label duplicate suppression (IoU)	0.9	remove near-identical overlaps predicted under different labels
–	cross-label duplicate suppression (area ratio)	0.5	additional overlap check to avoid suppressing distinct regions

Table 1: Default hyperparameters for semantic-candidate extraction and semantic ranking (project configuration).

thresholding x_i on the common scale $\{1, \ldots, K-1\}$. With $K = 5$, three ordered thresholds $\{\theta_{12}, \theta_{23}, \theta_{34}\}$ partition the score axis as

$$r_i(x_i) = \begin{cases} 1, & x_i < \theta_{12}, \\ 2, & \theta_{12} \le x_i < \theta_{23}, \\ 3, & \theta_{23} \le x_i < \theta_{34}, \\ K-1, & x_i \ge \theta_{34}. \end{cases}$$

Thresholds are defined as percentiles of the within-image score distribution over the retained candidates and were selected through dedicated fine-tuning; alternative cutoffs were empirically observed to reduce end-to-end performance and are therefore not adopted. For the task-agnostic branch, $\{\tau_{12}, \tau_{23}, \tau_{34}\}$ correspond to the (P15, P55, P97) cutoffs of $\{a_i\}$, while for the task-oriented branch $\{\upsilon_{12}, \upsilon_{23}, \upsilon_{34}\}$ are set to (P10, P45, P90) for the high-quantile mean task score t_i. Branch-specific threshold sets are reported in Table 2 for CLIP scores ($\theta \equiv \tau$) and in Table 3 for task scores ($\theta \equiv \upsilon$).

If no candidate reaches level $K-1$, the maximum-score candidate is promoted:

$$\text{if } \max_i r_i(x_i) < K-1 \text{ then } r_{i^\star}(x_{i^\star}) \leftarrow K-1, \quad i^\star \in \arg\max_i x_i.$$

3.2.2 Semantic candidates and ROI representation

Semantic candidates are provided by a configurable segmentation backend operating on the fixed label set $\mathcal{L}$. The segmentation is implemented as a two-stage pipeline: a zero-shot detector (GroundingDINO) first produces label-conditioned bounding boxes and confidence scores, and Segment Anything (SAM) then converts each detection into a pixel mask. The resulting masks are merged per label and are subsequently processed by cross-label duplicate

suppression. Each candidate is described by a predicted semantic label, a pixel mask, a bounding box, and an associated confidence score. Redundant proposals arising under different labels are reduced through *cross-label duplicate suppression*, which removes near-identical overlaps based on mask overlap criteria (intersection-over-union and an area-ratio check), retaining a single representative candidate.

For ranking and aggregation, each mask is converted into a binary ROI indicator:

$$M_i(y, x) = \mathbb{1}\big[\text{mask}(y, x) > 0\big].$$

Candidates are filtered by the minimum mask area fraction ε_{roi} using the normalized mask area defined in Section 3.2.1 (Table 1). The retained masks are shared by both ranking methodologies and provide a common spatial support for candidate-level scoring.

3.2.3 Task-agnostic scoring and semantic ranking (CLIP-based)

The task-agnostic methodology estimates candidate relevance through semantic self-consistency: a candidate is considered more relevant if, when isolated from the image, it preserves the global semantics captured by a CLIP image embedding.

Candidate-isolated view construction. Given a retained semantic-candidate mask M_i, a candidate-isolated view $\tilde{I}_i$ is generated by preserving pixels inside the mask and replacing all remaining pixels with a constant fill color. The image is first cropped around the mask support, and the replacement is applied within the cropped window. The crop size is controlled by the expansion parameter reported in Table 2. Let $(I_i^{\text{crop}}, M_i^{\text{crop}})$ denote the image and mask returned by the crop operator applied around the mask support, with expansion factor η_{crop} (Table 2). The candidate-isolated view is then obtained by background-filling pixels outside the mask support within the cropped window:

$$\tilde{I}_i \;=\; I_i^{\text{crop}} \odot M_i^{\text{crop}} \;+\; \mathbf{c} \odot \big(\mathbf{1} - M_i^{\text{crop}}\big),$$

where $\mathbf{c}$ is a constant RGB value and $\odot$ denotes element-wise multiplication.

Symbol / item	Meaning	Default	Role in pipeline
η_{crop}	crop expansion factor	0.0	context margin around the mask support
ε_{roi}	minimum ROI area fraction	0.001	discard negligible masks before scoring
K	number of rank levels	5	discrete scale $\{1, \ldots, K-1\}$
β	area-compensation weight	0.8	mitigate size bias in CLIP scoring via $\bar{a}_i = a_i - \beta \log(\rho_i + \epsilon_{\log})$
$\epsilon_{\log}$	log-stability constant	10^{-6}	avoid $\log(0)$ in area-compensated scoring
–	discretization strategy	`thresholds_img`	thresholding on CLIP image–image scores
$\{\tau_{12}, \tau_{23}, \tau_{34}\}$	CLIP img–img thresholds (P15/P55/P97)	{0.844, 0.96, 1.169}	map a_i to $\{1, \ldots, K-1\}$
–	enforce top-rank constraint	`true`	promote the top-scoring segment if needed

Table 2: Default hyperparameters for task-agnostic CLIP-based scoring and discretization.

CLIP similarity score. For each candidate-isolated view, a CLIP image-to-image similarity score is computed:

$$a_i \triangleq \text{CLIP}_{\text{img}}\big(I,\, \tilde{I}_i\big),$$

where $\text{CLIP}_{\text{img}}(\cdot, \cdot)$ denotes the project CLIP scoring routine applied to the original image and the corresponding candidate view.

Area-compensated CLIP score (segment-size balancing). CLIP image-to-image similarity tends to favor large candidates, as they preserve a larger portion of the global content. To characterize this size bias, an auxiliary area-compensated score is defined using the candidate area fraction $\rho_i \triangleq |M_i|/(H \times W)$, with H and W denoting the image height and width:

$$\bar{a}_i \triangleq a_i - \beta\, \log(\rho_i + \epsilon_{\log}),$$

where β controls the strength of the compensation and $\epsilon_{\log}$ is a small constant for numerical stability. This definition provides a size-normalized view of candidate relevance and is used to analyze the impact of segment area on CLIP-based scoring.

Discretization. Task-agnostic scores $\{a_i\}$ are mapped to discrete ranks using the common thresholding rule defined in Section 3.2.1, with $\theta \equiv \tau$ and thresholds $\{\tau_{12}, \tau_{23}, \tau_{34}\}$ reported in Table 2.

3.2.4 Task-oriented semantic ranking (face-relevance aggregation)

The task-oriented methodology estimates candidate relevance with respect to a face-recognition objective. A task-relevance heatmap is constructed once per image via occlusion sensitivity

of face-embedding similarity [13], and candidate scores are obtained by aggregating heatmap values within each semantic-candidate mask. Face detection, landmark localization, affine alignment, and face-embedding extraction are implemented using **InsightFace** [28].

Task-relevance heatmap (multi-face occlusion sensitivity). Let $I \in \mathbb{R}^{H\times W\times 3}$ be the input image. Face instances are obtained with **InsightFace**, which returns detected faces along with their bounding boxes and facial landmarks; the same module provides the face-recognition embedding model (`buffalo_s`). For each detected face, an affine normalization produces an aligned crop $F \in \mathbb{R}^{112\times 112\times 3}$ and the corresponding affine transform. A reference embedding e_0 is extracted from the unoccluded crop. Occlusion sensitivity is evaluated by sliding a square patch of side p with stride s over F. At each patch location u, pixels inside the patch are replaced by a constant baseline value, yielding an occluded crop F_u and an embedding e_u. The occlusion-induced drop is defined as

$$d(u) \triangleq 1 - \cos\big(e_u, e_0\big),$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity. Drops are accumulated over the covered pixels and averaged across overlaps, producing a dense per-face relevance map in the aligned coordinates. Each per-face map is then projected back to the original image plane via the inverse affine transform, yielding a set of full-resolution maps $\{H_f\}_{f=1}^{n_f}$.

The final multi-face heatmap is obtained by pixel-wise combination of the projected maps (max rule) followed by global min–max normalization and 8-bit quantization:

$$H(y,x) = \max_f H_f(y,x), \qquad H^{\mathrm{FR}} = \mathrm{uint8}\left(\mathrm{clip}\left(255 \cdot \frac{H - \min H}{\max H - \min H + \epsilon},\, 0,\, 255\right)\right),$$

with a small ϵ added for numerical stability. If no faces are detected, the task-relevance heatmap is set to zero everywhere.

Symbol / item	Meaning	Default	Role in pipeline
–	InsightFace `FaceAnalysis` model pack	`buffalo_s`	face detection/landmarks, alignment, and embedding extraction
p	occlusion patch size	16	spatial support of each perturbation
s	occlusion stride	8	sampling density of occlusions
–	baseline fill value	`mean`	patch replacement used during occlusion
–	multi-face fusion rule	`max`	pixel-wise aggregation of per-face maps
q	high-quantile level	0.9	retain top-10% heatmap values within each mask
K	number of rank levels	5	discrete scale $\{1, \ldots, K-1\}$
–	task score definition	`fr_seg_qmean`	high-quantile mean over the mask support
$\{\upsilon_{12}, \upsilon_{23}, \upsilon_{34}\}$	task thresholds (P10/P45/P90)	{0.0001, 0.0162, 0.2065}	discretization of task scores to $\{1, \ldots, K-1\}$
–	enforce top-rank constraint	`true`	promote the top-scoring candidate

Table 3: Default hyperparameters for task-oriented face-relevance scoring and discretization.

Candidate-level aggregation. Let $M_i \in \{0,1\}^{H\times W}$ be the mask of candidate s_i and let $\mathcal{P}_i = \{(y,x) : M_i(y,x) = 1\}$ denote its support. Heatmap values are normalized to $[0,1]$ as $h(y,x) = H^{\mathrm{FR}}(y,x)/255$. The task score is computed as a quantile-based mean over the mask support:

$$t_i \triangleq \frac{1}{|\mathcal{Q}_i|} \sum_{(y,x)\in\mathcal{Q}_i} h(y,x), \qquad \mathcal{Q}_i \triangleq \big\{(y,x) \in \mathcal{P}_i : \; h(y,x) \geq \mathrm{Quantile}_q\big(h(\mathcal{P}_i)\big)\big\},$$

where q is set by configuration (Table 3). This statistic emphasizes the most task-relevant pixels within each candidate. In addition to the high-quantile mean t_i (used as the task score in the default configuration), auxiliary statistics for logging/analysis are computed: the full-mask mean

$$\mu_i \triangleq \frac{1}{|\mathcal{P}_i|} \sum_{(y,x)\in\mathcal{P}_i} h(y,x),$$

and the non-zero fraction

$$\rho_i \triangleq \frac{1}{|\mathcal{P}_i|} \sum_{(y,x)\in\mathcal{P}_i} \mathbb{1}\big[h(y,x) > 0\big].$$

Discretization. Task-oriented scores $\{t_i\}$ are mapped to discrete levels using the common thresholding rule defined in Section 3.2.1, with $\theta \equiv \upsilon$ and thresholds $\{\upsilon_{12}, \upsilon_{23}, \upsilon_{34}\}$ reported in Table 3.

3.2.5 Category-level aggregation and SPIFF mapping

The semantic ranking module assigns discrete relevance ranks at *candidate* granularity. Downstream stages, however, consume a *label-level* prior (one value per semantic cat-

egory). Candidate ranks are therefore aggregated per label and exported in a compact SPIFF-compatible representation.

Rank domain and missing labels. For a fixed number of levels K (default $K = 5$), relevance ranks lie in $\{1, \ldots, K-1\}$, where larger values indicate higher relevance (by convention, rank 1 corresponds to low-importance/background-like content, while rank $K - 1$ corresponds to the main subject). A special value 0 denotes *missing / not present* labels; entries with rank 0 are treated as ignored (e.g., excluded from evaluation).

Category-level aggregation. Let $\mathcal{S} = \{s_i\}_{i=1}^{N}$ be the retained semantic candidates for an image, with labels $\ell_i \in \mathcal{L}$ and candidate-level ranks r_i (task-agnostic) and r_i^{task} (task-oriented). A single rank per label is obtained by max-reduction over all instances of the same category:

$$R(\ell) \triangleq \max_{i:\, \ell_i = \ell} r_i, \qquad R^{\text{task}}(\ell) \triangleq \max_{i:\, \ell_i = \ell} r_i^{\text{task}}, \qquad \ell \in \mathcal{L}.$$

If a label ℓ is not present among the retained candidates, it is assigned rank $R(\ell) = 0$ (and analogously $R^{\text{task}}(\ell) = 0$), so that missing categories are explicitly represented but ignored by evaluation protocols that exclude zeros.

SPIFF label keys and export. Labels are mapped to stable dictionary keys. For each image, the implementation exports two SPIFF maps (label → rank):

- a task-agnostic map $\ell \mapsto R(\ell)$ (CLIP-based methodology);
- a task-oriented map $\ell \mapsto R^{\text{task}}(\ell)$ (face-relevance methodology).

3.2.6 Exploratory LLM-based semantic category ranking

An additional exploratory ranking branch is implemented with a vision-language LLM. For each image, the model receives the original image together with the textual context and returns one discrete relevance rank for each detected semantic label. In the present work, the prompt uses semantic labels only and does not include segment-level numeric descriptors such as ROI fraction, detection score, or bounding-box coordinates. The implementation

uses the Gemini API with structured JSON outputs [29], and the default configured model is `gemini-2.5-flash` [30].

Detected-label set and prompt construction. Let $S = \{s_i\}_{i=1}^{N}$ denote the cached segment set for the image, and let $\hat{\mathcal{L}} \subseteq \mathcal{L}$ be the ordered set of semantic labels detected in that image after label normalization and duplicate removal. The prompt is constructed from four elements: the sample identifier, the textual caption associated with the image, the global semantic label universe $\mathcal{L}$, and the ordered list of detected labels $\hat{\mathcal{L}}$. The detected-label list is provided without any segment-level numeric information and without sorting by geometric or detector-based cues. This design makes the LLM rely only on the image, the caption, and the semantic identities of the detected categories. The prompt also specifies the common ranking scale, requires exactly one rank for each detected label under each ranking definition, and explicitly instructs the model to compute the two rankings independently.

Dual ranking formulation. A single structured generation step is used to obtain two category-level rankings at once. The prompt explicitly requests two independent assessments:

- **Generic semantic ranking:** category-level importance for general image understanding and reconstruction relevance;
- **Face-oriented semantic ranking:** category-level importance with respect to downstream face recognition, with emphasis on identity preservation whenever faces are relevant in the image.

The prompt further specifies that the two rankings must be produced independently and that the face-oriented ranking must not be derived from the generic one. Relevance is expressed on the common discrete scale $\{0, 1, 2, 3, 4\}$, where larger values indicate higher importance and rank 0 denotes absent or negligible relevance.

Example prompt instance. Figure 9 shows a representative sample image used to illustrate the instantiated prompt in the exploratory LLM-based ranking methodology. The model receives the image, the paired caption, the full semantic label universe, and the ordered list of detected labels $\hat{\mathcal{L}}$. For sample `000000000474`, the detected labels are `baseball_glove`, `person`, and `face`.

Figure 9: Representative sample image used to illustrate the instantiated prompt in the exploratory LLM-based semantic category ranking methodology.

```
Sample id: 000000000474

Caption: A young boy kneeling down to catch a baseball.

Global semantic label universe: airplane, apple, backpack, banana,
baseball_bat, baseball_glove, bear, bench, bicycle, bird, boat,
book, bottle, bowl, bus, cake, car, carrot, cat, cell_phone, chair,
clock, cow, cup, dining_table, dog, donut, fire_hydrant, fork,
frisbee, giraffe, handbag, horse, hot_dog, kite, knife, laptop,
motorcycle, orange, parking_meter, person, face, pizza, potted_plant,
sandwich, scissors, sheep, skateboard, skis, snowboard, spoon,
sports_ball, stop_sign, suitcase, surfboard, teddy_bear, tennis_racket,
tie, toilet, toothbrush, traffic_light, train, truck, umbrella,
vase, wine_glass, zebra

You will receive the image and the list of DETECTED segment labels
below. Output must be JSON matching the provided schema with
exactly two top-level keys: free and face. Each of "Generic
semantic ranking" and "Face-oriented semantic ranking" must provide
exactly one rank per detected label.

Ranking scale: integers in {0,1,2,3,4}: 4=critical, 3=high,
2=medium, 1=low, 0=ignore.

Compute the two rankings as independent assessments: - Generic
semantic ranking: generic reconstruction importance (perceived
```

```
fidelity). - face-oriented semantic ranking: downstream FACE
RECOGNITION; prioritize identity preservation if relevant in the
image. Do NOT derive one ranking from the other; do NOT reuse
the free ranking to decide face. Compute generic semantic ranking
first, then compute face-oriented semantic ranking as a separate
assessment.

Detected segments: - baseball_glove - person - face
```

Category-level rank completion and stabilization. For each image, the model returns one discrete relevance rank for the detected labels under the two ranking methodologies. Let $\tilde{R}^{\text{gen}} : \hat{\mathcal{L}} \to \{0,1,2,3,4\}$ and $\tilde{R}^{\text{fr}} : \hat{\mathcal{L}} \to \{0,1,2,3,4\}$ denote the resulting present-label maps for the generic and face-oriented rankings, respectively. If a detected label is not explicitly assigned by the model, it is given a default base rank equal to 1.

The present-label maps are then expanded to the full semantic label universe $\mathcal{L}$ by assigning rank 0 to every label not detected in the image. Denoting the final maps by

$$R^{\text{gen}}_{\text{LLM}} : \mathcal{L} \to \{0,1,2,3,4\}, \qquad R^{\text{fr}}_{\text{LLM}} : \mathcal{L} \to \{0,1,2,3,4\},$$

the completion rule is

$$R^{b}_{\text{LLM}}(\ell) = \begin{cases} \tilde{R}^{b}(\ell), & \ell \in \hat{\mathcal{L}}, \\ 0, & \ell \notin \hat{\mathcal{L}}, \end{cases} \qquad b \in \{\text{gen}, \text{fr}\}.$$

As in the main ranking pipeline, a top-rank constraint is enforced separately for each branch. If no detected label is assigned rank 4, the highest-ranked detected label(s) are promoted to rank 4. This guarantees that each branch contains at least one maximally relevant detected category.

3.3 RONI informativeness measures and patch selection strategies

Following the reference pipeline introduced in Section 3.1, once the ROI region(s) have been selected and preserved, the remaining part of the image is treated as RONI. Under a strict

rate constraint, the encoder does not transmit the RONI exhaustively; instead, it conveys a compact set of localized visual *evidence* (sampled patches) that can anchor the generative reconstruction performed at the receiver. This design implements an unequal bit-allocation principle: high fidelity is reserved for semantically dominant content (ROI), whereas the RONI is represented through sparse but informative observations.

3.3.1 Problem formulation for RONI patch sampling

Let $I \in \mathbb{R}^{H\times W\times 3}$ denote the input image. The RONI domain is specified by a mask $M_{\text{roni}} \in \{0, 255\}^{H\times W}$, where pixels with value 255 identify admissible locations. Evidence is conveyed by sampling square patches of fixed side length s, whose centers are constrained to lie in this domain and whose spatial support is required to fit within the image bounds. Patch sampling is performed iteratively: at each iteration, a new patch is added, and the set of covered pixels is updated accordingly.

Formally, let

$$\Omega_{\text{roni}} = \{(y, x) : M_{\text{roni}}(y, x) = 255\},$$

and let $\Omega_t \subseteq \Omega_{\text{roni}}$ denote the subset of RONI pixels still *uncovered* after sampling t patches. The achieved coverage is defined as

$$\text{cov}(t) \;=\; 1 - \frac{|\Omega_t|}{|\Omega_{\text{roni}}|}. \tag{3}$$

The sampling loop terminates when $\text{cov}(t) \geq \rho$, where $\rho \in (0, 1]$ is the prescribed evidence budget. In the experimental setting considered in this thesis, $\rho = 0.15$, meaning that at most 15% of the eligible area is explicitly revealed through transmitted patches.

The choice $\rho = 0.15$ instantiates a strict rate constraint while preserving sufficient spatial anchors to stabilize diffusion-based inpainting. Enforcing a small coverage budget:

- limits the number of transmitted patches and, consequently, the bitstream size;
- promotes evidence sampling on regions that are most informative for reconstruction;
- preserves a clear separation between *explicit* ROI transmission and *implicit* RONI reconstruction.

Importantly, the budget is enforced as an *area* constraint (coverage) rather than as a fixed patch count, so the resulting number of sampled patches adapts to image resolution, RONI extent, and patch size.

Both the evidence budget ρ and the patch size are treated as configurable hyperparameters. In particular, the patch side length is controlled through a dimensionless ratio r with respect to the input resolution. Let $D = \min(H, W)$ and define the largest odd size not exceeding D as

$$D_{\text{odd}} = \begin{cases} D, & D \text{ odd}, \\ D - 1, & D \text{ even}. \end{cases}$$

Given a dimensionless ratio $r \in (0, 1]$, the nominal patch side is

$$s \;=\; 1 + \lfloor r\,(D_{\text{odd}} - 1) \rfloor\,, \qquad s \leftarrow s - \mathbf{1}[s \text{ even}], \qquad s = \max(s, 1). \tag{4}$$

This convention enforces an odd side length and makes s scale linearly with the image's minimum dimension, enabling symmetric centering.

Unless otherwise stated, all experiments are conducted under a fixed reference configuration, with evidence budget $\rho = 0.15$ and patch-size ratio $r = 0.10$. The value $r = 0.10$ sets the patch scale to approximately 10% of the image minimum dimension, providing a compromise between sufficiently informative local context for inpainting and a sparse evidence budget that allows spatially distributed anchors under the coverage constraint. Additional method-specific hyperparameters (e.g., window sizes for score computation and selector parameters) are introduced and motivated in the corresponding subsections. To ensure reproducibility of probabilistic components in sampling-based procedures, all experiments are conducted with a fixed random seed.

3.3.2 RONI informativeness maps

Given the RONI domain Ω_{roni} and the associated residual mask M_t defined in Section 3.3.1, the encoder assigns to each pixel a non-negative *informativeness* (or *score*) value, yielding a score field $S \in \mathbb{R}_{\geq 0}^{H \times W}$. This field quantifies the value of each location as contextual evidence for reconstruction and guides RONI patch sampling under the coverage constraint. Unless stated otherwise, cues that operate on grayscale intensity use an 8-bit grayscale image

$G \in \{0, \ldots, 255\}^{H \times W}$ obtained from the RGB input, whereas saliency is computed on an RGB image normalized to $[0, 1]$. For consistency across cues, each non-negative score field is stored in a common 8-bit range via max-rescaling,

$$\tilde{S}(u) = 255 \, \frac{S(u)}{\max_v S(v)} \in [0, 255], \qquad \text{for } \max_v S(v) > 0. \tag{5}$$

Three single-cue score fields are considered: local entropy, edge density, and saliency, along with an adaptive mixture that combines them. These scores are subsequently converted into sampling distributions over admissible patch centers to drive RONI patch sampling under the coverage constraint.

Table 4 summarizes the key hyperparameters and implementation conventions, reporting the default values used throughout the experiments unless stated otherwise.

Symbol / item	Meaning	Default	Implementation / notes
w_{ent}	entropy half-window size	5	window is $(2w_{\text{ent}} + 1)^2$
w_{edge}	edge-density half-window size	5	uniform filter over $(2w_{\text{edge}} + 1)^2$
k_{blur}	Gaussian blur kernel size	3	forced odd if even
σ	Canny hysteresis band around median	0.33	thresholds from $m = \text{median}(G)$
L	pyramid scales (saliency)	9	dyadic Gaussian pyramid
c	center levels (saliency)	$\{2, 3, 4\}$	center–surround pairs
d	center–surround offsets	$\{3, 4\}$	surround level $\ell_s = c + d$ with $\ell_s < L$
θ	Gabor orientations	$\{0°, 45°, 90°, 135°\}$	orientation conspicuity
$\lambda(\ell)$	Gabor wavelength (per scale)	$7 \cdot 1.5^{\ell}$	ℓ is the pyramid scale
$\sigma_{\text{gab}}(\ell)$	Gabor std (per scale)	$0.56\,\lambda(\ell)$	used to set kernel extent
$k(\ell)$	Gabor kernel size (odd)	$\text{clip}(\lceil 6\sigma_{\text{gab}}(\ell) \rceil, 7, 65)$	forced odd, bounded in $[7, 65]$
γ_{gab}	Gabor aspect ratio	0.5	fixed in filter definition
ψ	Gabor phase offset	0	fixed in filter definition
ε	numerical stabilizer	10^{-9}	used in normalizations (e.g., saliency)
γ	exponent for anti-peaky mixing weights	1	–
$[0, 255]$	common score range	–	max-rescaling for visualization/logging
dtype	stored score type	`uint32`	in-memory `uint32`; saved as `uint8` PNG for visualization

Table 4: Key parameters and implementation-level conventions for RONI informativeness measures (score fields) and derived sampling distributions.

Local entropy. Local entropy prioritizes RONI regions exhibiting local heterogeneity (texture). For each pixel u, a local (discrete) Shannon entropy score [18] is computed over a square neighborhood $\mathcal{N}_{w_{\text{ent}}}(u)$ of side $(2w_{\text{ent}} + 1)$, where $w_{\text{ent}} \in \mathbb{N}$ is the half-window size. In practice, entropy is implemented via a rank-based entropy filter with a binary square footprint. Let $p_{u,w_{\text{ent}}}(g)$ be the empirical distribution of gray levels $g \in \{0, \ldots, 255\}$ within $\mathcal{N}_{w_{\text{ent}}}(u)$;

the score is

$$S_{\text{ent}}(u) \;=\; -\sum_{g} p_{u,w_{\text{ent}}}(g)\, \log_2 p_{u,w_{\text{ent}}}(g). \tag{6}$$

Edge density. Edge density emphasizes RONI regions rich in local structural cues, which can better constrain geometry during inpainting. The converted image is smoothed with a Gaussian filter (kernel size k_{blur}) before edge extraction. A mild pre-smoothing reduces noise-induced spurious edges while preserving fine structures. A binary edge map $E \in \{0,1\}^{H\times W}$ is then obtained via the Canny detector [19]. The Canny thresholds are computed from the median intensity $m = \text{median}(G)$ as

$$t_{\text{low}} = \max\big(0, (1-\sigma)m\big), \qquad t_{\text{high}} = \min\big(255, (1+\sigma)m\big), \tag{7}$$

where σ controls the Canny hysteresis band around m. Local edge density is computed by counting edge pixels within a square neighborhood $\mathcal{N}_{w_{\text{edge}}}(u)$ of side $(2w_{\text{edge}}+1)$:

$$S_{\text{edge}}(u) \;=\; \sum_{v\in\mathcal{N}_{w_{\text{edge}}}(u)} E(v). \tag{8}$$

Saliency. Saliency prioritizes visually conspicuous regions through a bottom-up, center–surround mechanism inspired by Itti et al. [2] Starting from the RGB image normalized to $[0,1]$, dyadic Gaussian pyramids with L scales are built by repeated Gaussian smoothing and $2\times$ downsampling. Three feature families are extracted.

- **Intensity.** The intensity channel is defined as $\mathcal{I} = (R+G+B)/3$ and is represented across scales by its pyramid $P_{\mathcal{I}}(\ell)$.
- **Color opponency.** Color cues are computed under an intensity gate $\mathbf{1}[\mathcal{I} > 0.1\,\mathcal{I}_{\max}]$ to suppress low-energy pixels. After gating, channels are normalized by $(\mathcal{I}+\varepsilon)$ and combined into broad opponent components R, G, B, Y, each represented by a pyramid $P_X(\ell)$.
- **Orientation.** Orientation cues are obtained by Gabor filtering the intensity pyramid at $\theta \in \{0°, 45°, 90°, 135°\}$. For each scale s, the Gabor parameters $(\lambda(s), \sigma_{\text{gab}}(s), k(s), \gamma_{\text{gab}}, \psi)$ follow Table 4. Each response is energy-normalized per scale by division by its standard

deviation (plus ε).

Center–surround feature maps are then computed uniformly across feature channels. Let $P_X(\ell)$ denote pyramid level ℓ of channel X. For center levels $c \in \{2, 3, 4\}$ and offsets $d \in \{3, 4\}$ (with $s = c + d < L$), the center–surround map is

$$F^X_{c,s} = | P_X(c) - \uparrow P_X(s) |,$$

where $\uparrow$ denotes bilinear interpolation to the spatial shape of $P_X(c)$. Color feature maps follow the same formulation under a cross-opponent scheme (RG and BY).

Each feature map is normalized by an Itti-style operator that promotes sparse responses: the map is rescaled to $[0, 1]$, local maxima are extracted over a 3×3 neighborhood (reflect padding), and the map is reweighted according to the squared difference between the strongest maximum and the mean of the remaining maxima. Normalized maps are resized to a common reference scale (level 4) and fused by summation followed by normalization to form intensity and color conspicuity maps C_I and C_C. For orientation, conspicuity maps are formed independently per θ (across-scale fusion and normalization), then averaged across orientations and normalized to obtain C_O. The final saliency map is computed as

$$S_{\text{sal}} = \tfrac{1}{3}(C_I + C_C + C_O).$$

Adaptive mixture. In addition to single-cue scores, an adaptive *mix* strategy is defined to combine the three cues at the level of sampling probabilities. Let p_{ent}, p_{edge}, and p_{sal} denote the per-pixel sampling distributions derived from the entropy, edge-density, and saliency scores, respectively. All three distributions are defined over the full $H \times W$ grid. Mixing weights are estimated on RONI-only values, whereas the residual mask enforces final admissibility during patch sampling.

The mixed distribution is defined as the convex combination

$$p_{\text{mix}}(u) = \lambda_{\text{ent}}\, p_{\text{ent}}(u) + \lambda_{\text{edge}}\, p_{\text{edge}}(u) + \lambda_{\text{sal}}\, p_{\text{sal}}(u), \qquad \lambda_k \geq 0,\ \sum_k \lambda_k = 1. \tag{9}$$

Weights are adapted per image by estimating the *concentration* of each cue over the RONI. Given the initial RONI mask $M_0 \in \{0, 255\}^{H \times W}$, all score values outside Ω_{roni} are first set to

zero, and a Gini coefficient $g_k \in [0, 1]$ is computed from the remaining RONI values, using only strictly positive entries. A high g_k indicates a peaky (highly concentrated) cue, whereas a low g_k corresponds to a more spread-out cue. Anti-peaky weights are then assigned as

$$\lambda_k \propto (1 - g_k)^\gamma, \qquad k \in \{\text{ent}, \text{edge}, \text{sal}\}, \tag{10}$$

with exponent $\gamma \geq 0$, and normalized to sum to one.

Sampling distributions from score fields. For patch-center sampling, each non-negative score field is converted into a discrete probability distribution by restricting its support to admissible locations and normalizing the remaining mass. Let $\Omega_{\text{adm}}^{(t)}$ denote the set of admissible centers at iteration t, defined by the current residual mask (uncovered pixels) and by the border-feasibility constraint required to extract a full patch. The sampling distribution is

$$p_t(u) = \frac{S(u)}{\sum_{v \in \Omega_{\text{adm}}^{(t)}} S(v)} \quad \text{for } u \in \Omega_{\text{adm}}^{(t)}, \qquad p_t(u) = 0 \text{ otherwise.}$$

If $\sum_{v \in \Omega_{\text{adm}}^{(t)}} S(v) = 0$, sampling falls back to the uniform distribution over $\Omega_{\text{adm}}^{(t)}$.

3.3.3 Patch representation and transmitted side information

Each sampled evidence element corresponds to a square patch of side length s. It is therefore sufficient to identify each patch by

$$p_t = (y_t, x_t),$$

where (y_t, x_t) denotes its top-left corner in the original image coordinates, while s is treated as a global parameter. This is equivalent to specifying its axis-aligned bounding box.

Border feasibility is enforced differently depending on the patch selection strategy. In score-guided covering (SGC), if a sampled center would yield an out-of-bounds patch, it is clamped to the nearest feasible location before extraction, and the resulting patch is represented by its top-left corner. In FPS and DPP-based strategies, centers are instead restricted to an interior domain that excludes a margin of $h = \lfloor s/2 \rfloor$, so feasibility holds by construction.

For transmission, patch crops are packed into a compact *patch grid*, while the list of patch

locations $\{p_t\}$, together with the common side length s, is sent as side information to allow spatial recomposition at the receiver.

3.3.4 Patch sampling algorithms

Given an informativeness score field S (Section 3.3.2) and an evidence budget ρ (Section 3.3.1), patch sampling iteratively chooses patch centers over the RONI domain and updates the residual mask until the coverage criterion is met. Notation and feasibility conventions for patch extraction (mask values, patch side s, and border handling) follow Section 3.3.3.

Termination and implementation conventions. Sampling stops once the target coverage $\mathrm{cov}(t) \geq \rho$ is reached (Section 3.3.1) or when a safety cap $T_{\max}$ is exceeded, with coverage evaluated on the residual mask. SGC and FPS update the residual mask after each accepted patch, whereas DPP accrues coverage via a union-of-windows test and applies a single mask update at the end. Border feasibility follows the selector-specific convention described in Section 3.3.3.

Default parameters. Table 5 summarizes the key hyperparameters and implementation conventions for patch sampling, reporting the default values used throughout the experiments unless stated otherwise.

Symbol / item	Meaning	Default	Notes
`patch_size_ratio`	patch side as ratio of $\min(H, W)$	0.1	s forced odd
ρ	target RONI coverage	0.15	termination threshold
$T_{\max}$	safety iteration cap (SGC/FPS)	500	prevents long degenerate runs
ε	stabilizer (SGC)	10^{-8}	added before score normalization
α	spatial exponent (FPS)	1.0	dispersion strength in d^{α}
λ	semantic weight (FPS)	1.0	influence of normalized $\bar{S}$
T_{fps}	temperature (prob. FPS)	0.7	sampling $\propto A^{1/T_{\text{fps}}}$
Δ	candidate stride (DPP)	$\max(1, \lfloor s/2 \rfloor)$	regular interior grid
η	oversampling factor (DPP)	1.5	ordered list length $\lceil \eta\, k \rceil$
T_{dpp}	temperature (prob. DPP)	0.7	on weights and/or gain sampling
r_{cand}	candidate sampling ratio (prob. DPP)	1.0	fraction of grid candidates retained
$N_{\max}$	max candidates (prob. DPP)	50000	computational cap
p, σ	quality exponent / kernel scale	adaptive	from Gini + autocorrelation length

Table 5: Default hyperparameters and conventions for patch sampling algorithms. Adaptive parameters are estimated per image from global score-field statistics.

Score-guided probabilistic covering with density regularization (SGC). SGC samples patch centers from a distribution that combines informativeness and residual availability. At iteration t, the score field is restricted to the currently uncovered domain and stabilized:

$$S_t(u) = (S(u) + \varepsilon)\, \mathbf{1}[M_t(u) = 255]. \tag{11}$$

A score-based distribution p_t^{score} is obtained by normalizing S_t over admissible pixels; if the admissible mass is zero, the distribution falls back to uniform sampling on the admissible set (Section 3.3.2).

To encourage exploration of regions that can still contribute to the uncovered area, a *local residual density* term is computed by box filtering the current residual mask with a window of side s; after rescaling and normalization, this yields a density distribution p_t^{dens} obtained by normalizing the local-density field. In practice, the loop is entered only while uncovered pixels exist, so the normalization is well-defined. The final sampling distribution is the product distribution

$$p_t(u) \;\propto\; p_t^{\text{score}}(u)\, p_t^{\text{dens}}(u), \tag{12}$$

followed by renormalization. A center $c_t \sim p_t$ is sampled, required to lie in the admissible domain, clamped if needed for feasibility, and then instantiated as a patch. The residual mask

is updated by removing the patch footprint and coverage is evaluated via (3). The complete SGC procedure is summarized in Algorithm 1.

Algorithm 1 SGC: score-guided covering with density regularization

Input: score field S; initial mask M_0; target ρ; patch side s; cap $T_{\max}$
Stabilize the score field with ε
$M \leftarrow M_0$; $\quad t \leftarrow 0$
while $\mathrm{cov}(t) < \rho$ **and** $t < T_{\max}$ **do**
    Build p^{score} by restricting $(S + \varepsilon)$ to pixels with $M = 255$ and normalizing
    Build p^{dens} by box filtering M with window side s, then normalizing
    Form $p \propto p^{\mathrm{score}} \odot p^{\mathrm{dens}}$ and renormalize
    Sample $c \sim p$ in the residual mask; clamp c to be border-feasible if needed
    Remove the patch footprint from M
    $t \leftarrow t + 1$
end while

Farthest-point sampling with semantic weighting (FPS). FPS promotes spatial dispersion using farthest-point sampling [8] through distances to previously sampled centers, while allowing overlap. Let $\mathcal{A}$ denote the border-feasible domain. Given sampled centers $\{c_k\}_{k \le t}$, define the distance field

$$d_t(u) = \min_{k \le t} \|u - c_k\|_2, \qquad u \in \mathcal{A}, \tag{13}$$

updated by $d_{t+1}(u) = \min(d_t(u), \|u - c_{t+1}\|_2)$. The score field is min–max normalized on $\mathcal{A}$ to obtain $\bar{S} \in [0, 1]$, and dispersion and informativeness are combined through

$$A_t(u) = d_t(u)^{\alpha} \left(1 + \lambda \, \bar{S}(u)\right), \qquad u \in \mathcal{A}, \tag{14}$$

where α controls the spatial repulsion and λ the semantic weighting.

Two variants are used:

- **Deterministic.** The distance field is initialized from a seed given by the highest-score admissible pixel (used only to initialize d_0). Each iteration selects $c_{t+1} = \arg\max_{u \in \mathcal{A}} A_t(u)$.
- **Probabilistic.** The seed is sampled from a metric-derived distribution restricted to $\mathcal{A}$ (uniform fallback if degenerate). Subsequent centers are sampled from $\mathbb{P}(c_{t+1} = u) \propto A_t(u)^{1/T_{\mathrm{fps}}}$, rejecting invalid draws.

Each accepted center is instantiated as a patch, the residual mask is updated by removing its footprint, and coverage is evaluated via (3). The overall FPS selection loop is summarized in Algorithm 2.

Algorithm 2 FPS: adaptive farthest-point selection

Input: score field S; mask M_0; target ρ; patch side s; cap $T_{\max}$; α, λ; T_{fps}
Define $\mathcal{A}$ as admissible pixels excluding a margin of $h = \lfloor s/2 \rfloor$
Initialize seed c_0 from the metric-restricted RONI domain (prob.) or from the highest-score RONI pixel (det.), and set $d(u) \leftarrow \|u - c_0\|_2$
$M \leftarrow M_0; \quad t \leftarrow 0$
while $\mathrm{cov}(t) < \rho$ **and** $t < T_{\max}$ **and** valid candidates remain **do**
    Compute $A(u) = d(u)^{\alpha}\,(1 + \lambda\,\bar{S}(u))$ for $u \in \mathcal{A}$
    Choose c as $\arg\max_{u \in \mathcal{A}} A(u)$ (det.) or sample c with probability $\propto A(u)^{1/T_{\text{fps}}}$ (prob.)
    Remove the patch footprint from M
    Update $d(u) \leftarrow \min(d(u), \|u - c\|_2)$ for $u \in \mathcal{A}$
    $t \leftarrow t + 1$
end while

Determinantal point process selection (DPP). DPP selection enforces diversity through a repulsive kernel while favoring high-quality candidates [9]. Candidate centers are generated on a regular interior grid with stride Δ and restricted to admissible pixels.

Quality and diversity. Let $C = \{c_i\}_{i=1}^{N}$ denote the candidate centers and $z_i \in [0,1]^2$ their normalized coordinates. Diversity is modeled by an RBF kernel

$$K_{ij} = \exp\left(-\frac{\|z_i - z_j\|_2^2}{2\sigma^2}\right), \tag{15}$$

with a small diagonal stabilization term for numerical robustness. Quality weights are derived from candidate scores $s_i = S(c_i)$ via rank normalization: if scores are non-constant, candidates are ranked and mapped to $\tilde{r}_i \in [0,1]$, then

$$q_i = \tilde{r}_i^{\,p}. \tag{16}$$

Both p and σ are set adaptively from global score-field statistics (concentration via a Gini coefficient on positive values, and spatial scale via a correlation-length estimate from the 2D

autocorrelation). The DPP is defined as an L-ensemble,

$$L = \mathrm{diag}(q)\, K\, \mathrm{diag}(q). \tag{17}$$

Coverage-driven selection. An area-based target cardinality is estimated as

$$k_{\mathrm{target}} = \left\lceil \frac{\rho\, \|M_0\|_0}{s^2} \right\rceil. \tag{18}$$

An oversampled ordered list of candidates is produced by maximizing incremental log-determinant gain (greedy MAP), and then converted into a patch set by retaining only candidates that add new covered pixels. Coverage accrual is computed using a symmetric half-open window around the center,

$$\mathcal{W}(c_i) = [y_i - h,\ y_i + h) \times [x_i - h,\ x_i + h), \tag{19}$$

clipped to image bounds. This convention corresponds to a $(2h) \times (2h)$ window; for odd s, $2h = s - 1$, so the marginal-coverage test is slightly conservative with respect to the nominal $s \times s$ patch footprint. A boolean coverage map is updated by union over accepted windows, and the procedure stops when the covered fraction reaches ρ; the residual mask is then updated once by setting to 0 all covered pixels.

Probabilistic DPP. A probabilistic variant introduces randomness at two stages. First, candidates are sub-sampled without replacement from the stride grid with probability proportional to a metric-derived distribution evaluated at candidate locations; a temperature-controlled power transform is applied to candidate weights, and the candidate pool is capped by $N_{\max}$ while ensuring a minimum size proportional to k_{target}. Second, selection proceeds iteratively by sampling from a temperature-controlled softmax over incremental log-determinant gains. As in the deterministic case, candidates that do not add new pixels under (19) are discarded, and the procedure terminates once the coverage target is met (or once a safety cap is reached). The full DPP-based patch selection procedure is summarized in Algorithm 3.

Algorithm 3 DPP: relevance–diversity patch selection

Input: score field S; mask M_0; target ρ; patch side s; stride Δ; oversampling factor η; temperature T_{dpp}; candidate ratio r_{cand}; candidate cap $N_{\max}$
Define $\mathcal{A}$ as admissible pixels excluding a margin of $h = \lfloor s/2 \rfloor$
Build candidate centers $C = \{c_i\} \subset \mathcal{A}$ on a regular grid with stride Δ
Estimate $k_{\text{target}} \leftarrow \lceil \rho \, \|M_0\|_0 / s^2 \rceil$
For the probabilistic variant, replace C with a metric-weighted sub-sample controlled by r_{cand} and T_{dpp}, capped by $N_{\max}$ and with minimum size proportional to k_{target}
Estimate adaptive p and σ from global score-field statistics
Compute rank-normalized qualities q_i from $S(c_i)$ and form $L = \operatorname{diag}(q)\, K \operatorname{diag}(q)$ with $K_{ij} = \exp\!\big(-\|z_i - z_j\|_2^2 / (2\sigma^2)\big)$
Define $W(c_i) = [y_i - h, y_i + h) \times [x_i - h, x_i + h)$ clipped to image bounds
Initialize coverage map $U \leftarrow \varnothing$; $\quad M \leftarrow M_0$
while $\operatorname{cov}(U) < \rho$ **and** candidates remain **do**
    Build an ordered batch B from the remaining candidates:
     greedy log-det maximization with size $\lceil \eta \, k_{\text{target}} \rceil$ (det.),
     or a temperature-controlled softmax over incremental log-det gains using the residual uncovered area (prob.)
    for each $c \in B$ in order **do**
        if $W(c)$ adds new covered pixels to U **then**
            Accept c and update $U \leftarrow U \cup W(c)$
        end if
        Remove c from further consideration
        if $\operatorname{cov}(U) \geq \rho$ **then**
            break
        end if
    end for
end while
Set $M(u) \leftarrow 0$ for all $u \in U$

4 Results

4.1 Experimental setup and evaluation protocol

This chapter reports two complementary evaluation tracks targeting the coupled encoder challenges studied in this work: semantic ranking for ROI assignment, and RONI evidence placement through patch selection with end-to-end reconstruction. All experiments operate under common feasibility constraints and a shared evidence budget, while the generative decoder is kept fixed so that reconstruction differences can be attributed to encoder-side policies.

4.1.1 Dataset and evaluation subsets

All images used in this thesis are drawn from COCO-Stuff [31]. Each image is paired with a textual caption describing its visual content, generated with the BLIP model [32]. In the end-to-end reconstruction pipeline, this caption is used as auxiliary text conditioning for decoder-side inpainting. In the exploratory LLM-based ranking study, it is also used explicitly as part of the prompt context together with the input image and the detected semantic labels.

Since one of the main semantic-ranking criteria is explicitly task-oriented (face recognition), the dataset is further filtered to retain only samples for which the face-relevance computation is well-defined.

Three subsets are used across the experimental evaluation:

- a subset of 500 images for *RONI informativeness measures and patch sampling*, to evaluate patch selection strategies while keeping end-to-end reconstruction experiments computationally tractable;
- a subset of 780 images for the *semantic-ranking pipeline*, all satisfying the face-presence constraint, to enable a statistically more stable comparison of task-agnostic and task-oriented criteria under the face-recognition objective;
- a reduced subset of 20 images for the exploratory *LLM-based ranking study*, which is therefore interpreted as a preliminary extension rather than as part of the main comparative evaluation.

4.1.2 Experimental splits and common constraints

Two evaluation tracks are considered.

Semantic ranking. The main ranking pipeline is evaluated on the semantic-ranking subset introduced above. The exploratory LLM-based ranking study is evaluated separately on its dedicated reduced subset, using the same category-level evaluation definition but interpreted independently because of the much smaller sample size.

Patch selection and end-to-end reconstruction. Patch selection and reconstruction are evaluated under two regimes: a fixed-seed setting with complete outputs for all configurations, used for dataset-level comparisons under a common random realization of the probabilistic selectors, and a separate multi-seed analysis on $N = 50$ images evaluated under 5 seeds, restricted to probabilistic variants. Across all configurations, the preserved ROI is fixed, patch sampling is restricted to the admissible RONI, and all methods operate under the same coverage budget.

4.1.3 Semantic-ranking evaluation protocol

Agreement between automatic rank assignments and manual annotations is evaluated at the *category level* against a face-driven ground truth, using a common JSON representation in which each sample identifier maps to a dictionary `label` $\rightarrow$ `rank`. The same protocol is used for the two main semantic-ranking methodologies and, separately, for the exploratory LLM-based category-ranking study on its reduced subset.

Inputs and evaluation domain. Let $G(s, \ell)$ and $P(s, \ell)$ denote the ground-truth and predicted rank for sample s and label ℓ, respectively. Ranks are assumed to lie in $\{0, 1, 2, 3, 4\}$, where rank 0 denotes an *ignore* state. Since ground truth and predictions are defined on the same detected-label set, rank-0 entries correspond to labels outside the evaluated set and are not included in metric computation. Therefore, all reported semantic-ranking metrics are computed over ranks $r \in \{1, 2, 3, 4\}$ only.

Per-rank one-vs-rest evaluation. For each rank $r \in \{1, 2, 3, 4\}$, metrics are computed through a one-vs-rest decomposition. For every evaluated pair (s, ℓ), define

$$y_r(s, \ell) = \mathbb{1}[G(s, \ell) = r], \qquad \hat{y}_r(s, \ell) = \mathbb{1}[P(s, \ell) = r],$$

and accumulate $(\mathrm{TP}_r, \mathrm{FP}_r, \mathrm{FN}_r, \mathrm{TN}_r)$ over the dataset. Precision, recall, and F_β are then computed as

$$P_r = \frac{\mathrm{TP}_r}{\mathrm{TP}_r + \mathrm{FP}_r}, \qquad R_r = \frac{\mathrm{TP}_r}{\mathrm{TP}_r + \mathrm{FN}_r}, \qquad F_{\beta,r} = \frac{(1 + \beta^2)\, P_r\, R_r}{\beta^2 P_r + R_r},$$

with safe handling of zero denominators. In all experiments reported in this chapter, $\beta = 1$.

Rank-weighted aggregation. To emphasize errors on higher-importance ranks, per-rank results are aggregated using non-negative rank-importance weights $\{w_r\}_{r=1}^{4}$ (default: $w_1 = 1, w_2 = 2, w_3 = 3, w_4 = 4$). Two complementary summaries are reported.

- **Rank-weighted macro.** The final macro score is obtained as a weighted average across ranks:

  $$P_{\text{macro}} = \frac{\sum_{r=1}^{4} w_r\, P_r}{\sum_{r=1}^{4} w_r}, \qquad R_{\text{macro}} = \frac{\sum_{r=1}^{4} w_r\, R_r}{\sum_{r=1}^{4} w_r}, \qquad F_{\beta,\text{macro}} = \frac{\sum_{r=1}^{4} w_r\, F_{\beta,r}}{\sum_{r=1}^{4} w_r}.$$

  This view preserves rank-specific behavior and highlights failures concentrated on particular positions of the ranking.

- **Rank-weighted micro.** Weighted counts are first pooled across ranks:

  $$TP_w = \sum_{r=1}^{4} w_r\, TP_r, \qquad FP_w = \sum_{r=1}^{4} w_r\, FP_r, \qquad FN_w = \sum_{r=1}^{4} w_r\, FN_r.$$

  Precision and recall are then computed from these global weighted totals:

  $$P_{\text{micro}} = \frac{TP_w}{TP_w + FP_w}, \qquad R_{\text{micro}} = \frac{TP_w}{TP_w + FN_w},$$

and the corresponding micro F_β score is

$$F_{\beta,\text{micro}} = \frac{(1+\beta^2)\, P_{\text{micro}}\, R_{\text{micro}}}{\beta^2 P_{\text{micro}} + R_{\text{micro}}}.$$

This view summarizes the overall weighted error mass and is more influenced by ranks with larger associated weights.

Transition analysis. In addition to scalar metrics, a rank transition table is accumulated by counting occurrences of $(G(s,\ell), P(s,\ell))$ for $G, P \in \{1,2,3,4\}$. This view highlights systematic confusions between adjacent or distant rank levels.

4.1.4 Patch sampling strategies and reconstruction evaluation protocol

The second evaluation track concerns RONI patch sampling strategies and end-to-end reconstruction. Results are reported over all combinations of informativeness measure (entropy, edge density, saliency, and the adaptive mix) and patch selection strategy (SGC, FPS deterministic/probabilistic, and DPP deterministic/probabilistic). Each configuration produces a reconstructed version of the same sample through the full encoder–decoder pipeline. In parallel, map-level diagnostics are collected for each informativeness measure to characterize the concentration and spatial scale of the corresponding score field.

Reconstruction-quality metrics. For each reconstructed image, two complementary fidelity measures are reported:

- **Perceptual fidelity (PF).** A similarity score derived from LPIPS by converting distance into a bounded fidelity value,

  $$\text{PF} = 1 - \text{LPIPS},$$

  with optional clipping to $[0,1]$ for numerical stability.

- **Semantic fidelity (SF).** A CLIP-based similarity score mapped from cosine similarity to $[0,1]$,

  $$\text{SF} = \frac{\text{CLIP} + 1}{2},$$

  again clipped to $[0,1]$ when needed.

In addition, the inference time required for generative reconstruction is recorded to contextualize fidelity gains against computational cost.

Informativeness-measure diagnostics. Each informativeness measure m produces a score field $S^{(m)} \in \mathbb{R}^{H\times W}$ on the full image domain $\Omega = [H] \times [W]$. Four diagnostics are used to summarize dispersion, concentration, effective support, and spatial scale:

- **Score dispersion (coefficient of variation).** Let μ_S and σ_S be the mean and standard deviation of $\{S(u)\}_{u\in\Omega}$. Dispersion is summarized as

$$\mathrm{CV}_S = \frac{\sigma_S}{\mu_S + 10^{-12}}.$$

- **Score concentration (Gini coefficient).** Let $v = \{S(u) : u \in \Omega,\, S(u) > 0\}$ and let $n_+ = |v|$. Let $v_{(1)} \le \cdots \le v_{(n_+)}$ be the sorted values and $c_i = \sum_{j=1}^{i} v_{(j)}$. Concentration is summarized as

$$G_S = \frac{n_+ + 1 - 2\,\sum_{i=1}^{n_+} \frac{c_i}{c_{n_+}+10^{-12}}}{n_+},$$

  with $G_S = 0$ if $n_+ = 0$.

- **Effective support at half mass.** Let p be the sampling distribution induced by the informativeness measure on Ω, with $\sum_{u\in\Omega} p(u) = 1$. Let $p_{(1)} \ge \cdots \ge p_{(n)}$ denote its values sorted in descending order, with $n = |\Omega|$. Define

$$k^\star = \min\left\{ k : \sum_{j=1}^{k} p_{(j)} \ge \tfrac{1}{2} \right\},$$

  and report the normalized support size $k^\star/n$.

- **Spatial scale (correlation length, px).** Spatial smoothness is summarized by a correlation-length estimate derived from the normalized 2D autocorrelation of the score field. Let d_x and d_y be the first half-maximum crossing distances from the center along the central row and column. The reported scale is

$$\ell_{\mathrm{corr}} = \sqrt{d_x d_y}.$$

Patch-set geometry and robustness. Each sampled patch set is further summarized through patch overlap, defined as the fraction of patch-covered pixels that are covered more than once, and mean center distance (MCD), defined as the mean pairwise Euclidean distance between patch centers. Selection-side execution times are also tracked, and the ROI area fraction is retained to support stratified analyses by ROI dominance.

Probabilistic variants are evaluated under multiple random seeds. Patch-geometry and execution-time variability are summarized through mean, standard deviation, and seed coefficient of variation across seeds for each selector–measure configuration, whereas PF and SF are compared using seed-averaged values as the default input for cross-configuration analyses.

Comparative summaries. Since absolute fidelity can vary substantially across samples, configurations are also compared within each sample by ranking PF and SF separately (rank 1 is best). Two complementary normalized views are reported: gap-to-best, i.e., the difference from the best configuration on the same sample, and centered deviation, i.e., the deviation from the sample-wise mean performance across configurations. These statistics are computed both within selector families and globally across all selector–measure combinations.

At the dataset level, results are summarized through tables of mean rank, rank standard deviation, mean gap-to-best, and win rate, as well as heatmaps of mean ranks and deviations and performance profiles based on within-ε distance from the sample-wise best. To relate evidence geometry to reconstruction quality, PF and SF are also analyzed against patch overlap and MCD through correlation summaries, scatter plots, and per-sample diagnostic heatmaps.

4.2 Segment ranking pipeline results

This section reports agreement between the ground truth and the ranks produced by the two main semantic-ranking methodologies on the common evaluated subset, following the protocol defined in Section 4.1.3.

4.2.1 Overall agreement summary

The evaluation uses $\beta = 1$ and weights $w_1 = 1$, $w_2 = 2$, $w_3 = 3$, $w_4 = 4$. A symmetric choice $\beta = 1$ is adopted to weight precision and recall equally, since both false inclusions

Table 6: Rank-weighted precision, recall, and F_1 for the two ranking methodologies ($\beta = 1$).

Method	Prec.	Rec.	F_1
Task-agnostic (rank-weighted macro)	0.286	0.284	0.283
Task-agnostic (rank-weighted micro)	0.290	0.289	0.290
Task-oriented (rank-weighted macro)	0.838	0.774	0.741
Task-oriented (rank-weighted micro)	0.842	0.767	0.803

and false exclusions affect the encoder allocation under a fixed evidence budget. Rank weights increase linearly with r to reflect the ordinal semantics of the scale: higher ranks correspond to progressively more task-critical content, and mismatches at high ranks should contribute more to the aggregate score than mismatches at low ranks. This choice preserves interpretability while enforcing a monotone importance profile across ranks.

For the main ranking pipeline, the annotated split yields the same set of 780 samples in both the ground truth and the predictions. Each sample contributes a variable number of ranked segments, determined by the semantic segmentation output and by the subset of segment labels that are mapped to the evaluation vocabulary. After excluding rank-0 entries, agreement is computed over a total of 2872 label-level comparisons.

Figure 10 summarizes overall agreement through rank-weighted F_1. The task-agnostic methodology achieves $F_1 = 0.283$ (rank-weighted macro) and $F_1 = 0.290$ (rank-weighted micro). The task-oriented methodology achieves $F_1 = 0.741$ (rank-weighted macro) and $F_1 = 0.803$ (rank-weighted micro).

Table 6 reports the corresponding rank-weighted precision, recall, and F_1 values for the two methodologies under both aggregation schemes.

In the task-agnostic methodology, rank-weighted precision and recall are both low and closely matched, indicating substantial false inclusions and false exclusions with respect to the face-driven reference. In the task-oriented methodology, precision is consistently higher than recall: high-rank assignments are usually correct, while a non-negligible portion of relevant labels is missed, consistent with the systematic under-ranking of person-associated but face-distant segments (notably GT rank 2) discussed in Section 4.2.4. As defined in Section 4.1.3, rank-weighted macro averages per-rank one-vs-rest scores and therefore does not, in general, coincide with the harmonic mean of the reported macro precision and macro recall.

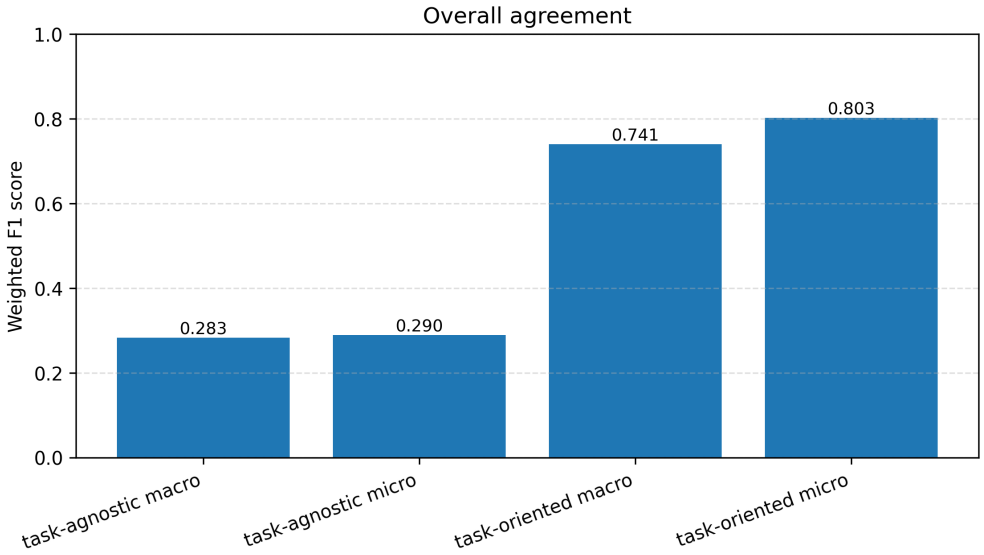


Figure 10: Overall agreement in the task-agnostic and task-oriented methodologies, reported as weighted macro and weighted micro F_1 scores ($\beta = 1$).

The observed gap between the two methodologies is expected. The manual ground truth was constructed under a face-recognition objective. Ranks, therefore, encode task relevance with respect to faces, and higher agreement is expected when the same objective is used at inference time. Accordingly, the task-oriented methodology is aligned with the reference by design, as it ranks segments using a face-relevance score derived from the face-recognition objective. The task-agnostic methodology is intentionally generic and is not intended to detect task-specific relevance. Low agreement against a face-driven reference thus reflects objective mismatch rather than a pipeline malfunction.

In this setting, the task-agnostic results act as a baseline, quantifying the agreement that can be achieved without task supervision. The task-oriented results quantify the gain induced by task alignment.

4.2.2 Per-rank one-vs-rest metrics

Figure 11 reports per-rank F_1 for both methodologies.

In the task-agnostic methodology, F_1 remains low across ranks. The CLIP criterion is based on semantic self-consistency: a segment is promoted when its isolated view preserves the global CLIP embedding of the original image. This setup is intrinsically biased toward regions that retain substantial scene content and context (typically larger masks), while small

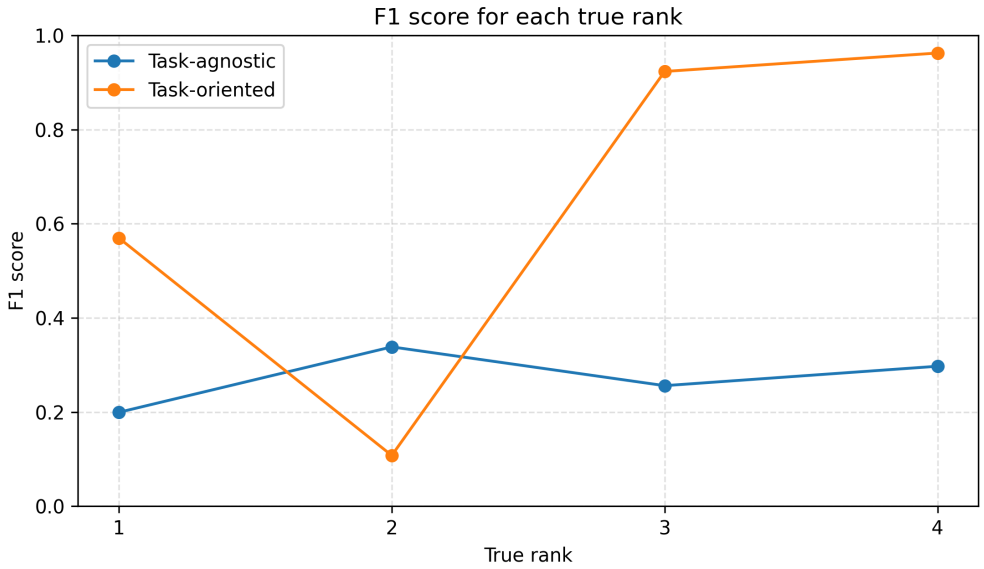


Figure 11: F_1 score for each true rank in the task-agnostic and task-oriented methodologies.

but task-critical regions (e.g., faces) explain only a limited portion of the global appearance and are therefore disadvantaged. The post-processing controlled by γ, $s_{\text{corr}} = s - \gamma \log(\rho + \varepsilon)$, compensates this size effect and improves agreement up to an intermediate value, after which over-correction degrades performance.

Rank 2 is the strongest level because it aligns best with what the generic CLIP score tends to surface after area compensation: moderately large, visually salient foreground regions (often person-associated objects) that still preserve substantial global semantics, without specifically encoding face relevance. Even at this level, agreement remains limited.

In the task-oriented methodology, ranks 3 and 4 show consistently high F_1. These levels correspond to segments with strong face relevance, for which the adopted criterion is well matched to the reference. Segment scores are obtained by aggregating face-relevance values over each candidate region; consequently, regions that overlap the face or lie in its immediate vicinity accumulate high face evidence and are reliably promoted to the top ranks.

By construction, the same mechanism is less effective for segments that are person-associated but only weakly coupled to the face region (e.g., worn or carried objects, clothing parts, or contextually person-related items that are spatially separated). For these regions, the face-relevance values are weak and sparse, so the region-level aggregation returns a low score, and the segment is consistently under-ranked. This primarily affects rank 2, which, in the manual reference, is intended to capture such person-related but face-distant evidence,

and explains its low recall.

Rank 1 is consequently overloaded: many segments that are not strongly face-driven receive similarly low aggregated face evidence and collapse into the lowest evaluated level, increasing heterogeneity and reducing precision relative to ranks 3 and 4. The resulting structured failure mode, characterized by frequent transitions from GT rank 2 to predicted rank 1, is further detailed in Section 4.2.4.

4.2.3 Rank-weighted aggregation analysis

Figure 10 contrasts rank-weighted macro and rank-weighted micro aggregation.

For the task-agnostic methodology, rank-weighted macro and rank-weighted micro are nearly identical (Figure 10), indicating that errors are not dominated by a single rank but are distributed across the scale.

For the task-oriented methodology, rank-weighted micro F_1 is higher than rank-weighted macro F_1 (Figure 10). This reflects a mixed regime: ranks 3 and 4 are predicted reliably and carry larger weights, so they contribute strongly to the pooled counts, whereas the systematic failure at rank 2 penalizes the per-rank averaging more severely.

4.2.4 Rank transitions (GT → Pred)

Rank transitions are analyzed through ground-truth to prediction confusion heatmaps over ranks $\{1, 2, 3, 4\}$. Rows correspond to GT ranks and columns to predicted ranks. Figures 12–13 visualize row-normalized transitions (percentages), with counts reported in-cell.

Task-agnostic methodology. Transitions are broadly spread, and diagonal mass is limited (Figure 12), indicating weak rank-by-rank agreement with the reference. A clear accumulation appears in the predicted rank-2 column, indicating a tendency to assign intermediate ranks when the continuous CLIP scores are discretized.

Two systematic drifts can be observed. First, many GT rank-1 labels are promoted to higher predicted ranks. This occurs when objects that are salient with respect to the caption or to global scene semantics are not face-relevant: they are assigned a low rank in the face-driven reference, yet they score high under a generic CLIP similarity criterion. Second, a substantial portion of high GT ranks is mapped to predicted rank 2. Segments that are highly

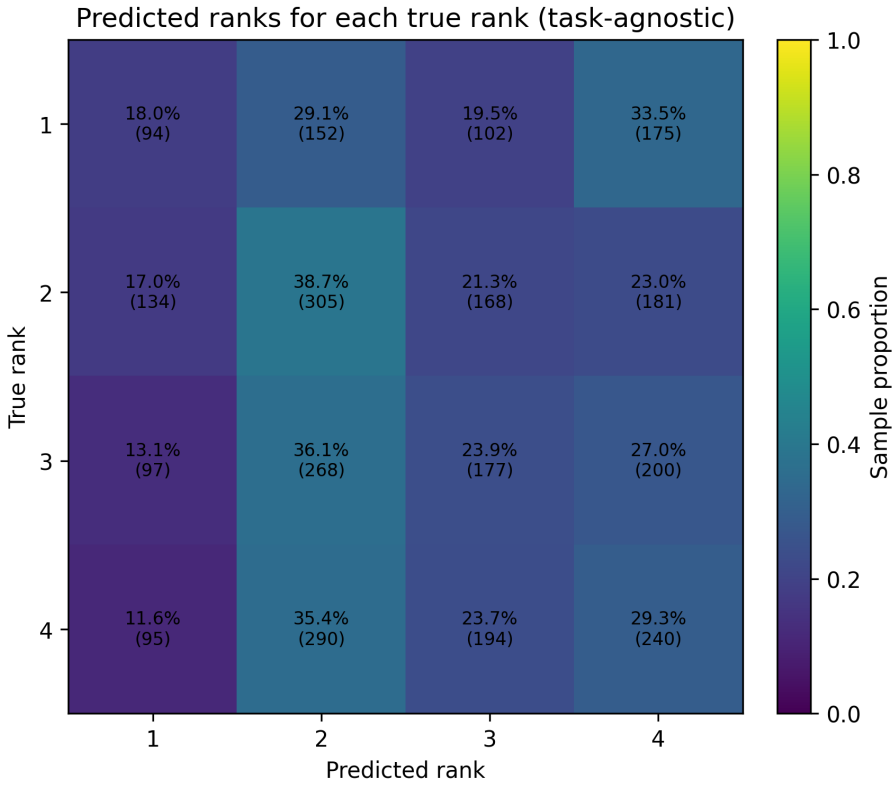


Figure 12: Distribution of predicted ranks for each true rank in the task-agnostic methodology, with row-wise normalization and in-cell counts.

face-relevant are not necessarily those that maximize generic semantic self-consistency with the whole image; they can therefore receive intermediate CLIP scores and be placed in the middle level.

Task-oriented methodology. Transitions concentrate on the diagonal for ranks 3 and 4 (Figure 13), indicating that high-relevance segments are ranked consistently. The dominant off-diagonal pattern concerns rank 2: most GT rank-2 labels are mapped to predicted rank 1, revealing a structured and repeatable failure mode.

This behavior matches the intended meaning of rank 2 in the ground truth reference. Rank 2 is used for foreground objects that are associated with the depicted person but are not directly supported by the face region, such as worn items, carried objects, or other person-related elements that may be spatially separated from the face.

The task-oriented criterion aggregates face evidence over each candidate region. It is therefore most reliable for segments that overlap the face or lie close to it, where face evidence is strong and spatially dense. For person-associated but face-distant regions, face

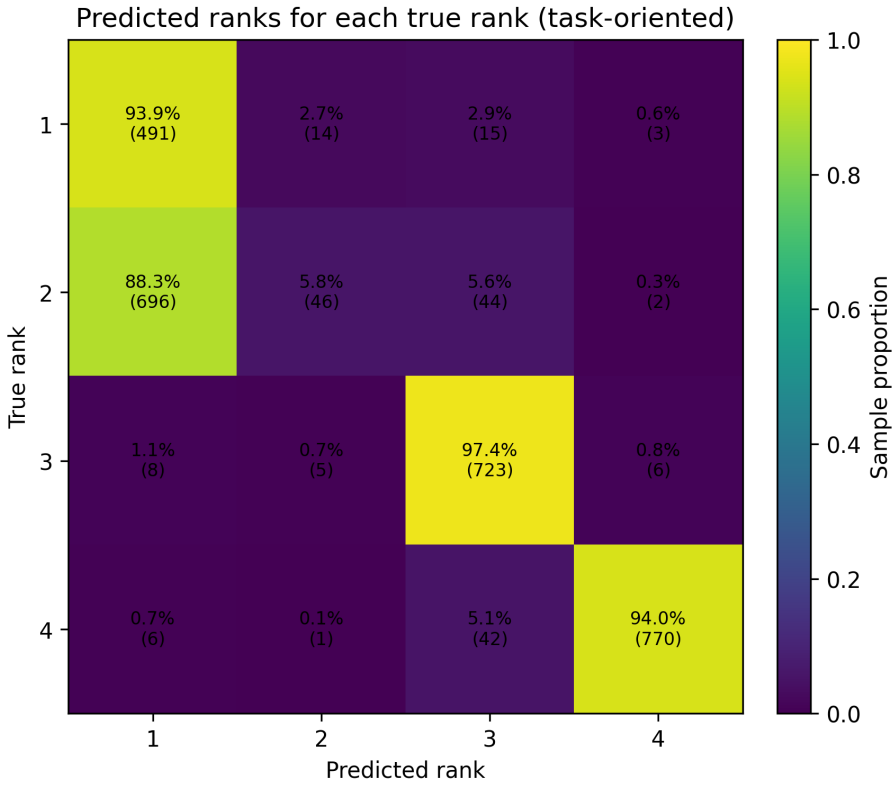


Figure 13: Distribution of predicted ranks for each true rank in the task-oriented methodology, with row-wise normalization and in-cell counts.

evidence is weak and sparse, yielding low aggregated scores and systematic under-ranking. This explains the frequent demotions from GT rank 2 to predicted rank 1.

4.2.5 Exploratory LLM-based ranking results

An exploratory evaluation of the LLM-based ranking methodology was conducted on the reduced subset introduced in Section 4.1.2, using the same category-level protocol defined in Section 4.1.3.

Overall agreement. Table 7 reports rank-weighted precision, recall, and F_1 for the two LLM-based outputs. The face-oriented ranking achieves substantially higher agreement than the generic one under both aggregation schemes. In particular, the face-oriented output reaches $F_1 = 0.559$ under rank-weighted macro and $F_1 = 0.753$ under rank-weighted micro, while the generic output reaches $F_1 = 0.280$ and $F_1 = 0.346$, respectively.

Relative to the main ranking results discussed above, the generic LLM output is close to the task-agnostic CLIP methodology in terms of aggregate performance, especially under

Table 7: Rank-weighted precision, recall, and F_1 for the exploratory LLM-based ranking outputs ($\beta = 1$).

Method	Prec.	Rec.	F_1
Generic semantic ranking (rank-weighted macro)	0.254	0.343	0.280
Generic semantic ranking (rank-weighted micro)	0.320	0.378	0.346
Face-oriented semantic ranking (rank-weighted macro)	0.637	0.624	0.559
Face-oriented semantic ranking (rank-weighted micro)	0.767	0.739	0.753

Table 8: Precision, recall, and F_1 computed separately for each rank in the exploratory LLM-based rankings ($\beta = 1$).

	Generic semantic ranking			Face-oriented semantic ranking		
Rank	Prec.	Rec.	F_1	Prec.	Rec.	F_1
1	0.667	0.500	0.571	0.167	1.000	0.286
2	0.000	0.000	0.000	0.000	0.000	0.000
3	0.143	0.111	0.125	1.000	0.412	0.583
4	0.361	0.650	0.464	0.800	1.000	0.889

rank-weighted macro (0.280 vs. 0.283), while its micro score is slightly higher (0.346 vs. 0.290). The face-oriented LLM output remains below the main task-oriented methodology (0.559/0.753 vs. 0.741/0.803 for macro/micro F_1), but exhibits the same qualitative advantage over the corresponding generic variant.

Per-rank one-vs-rest metrics. Table 8 reports per-rank precision, recall, and F_1. The generic output is comparatively strongest at rank 1 and rank 4, but remains weak overall because rank 2 is never recovered and rank 3 is predicted unreliably. The face-oriented output is more structured: rank 4 is identified very reliably, with $P = 0.800$, $R = 1.000$, and $F_1 = 0.889$, while rank 3 also achieves high precision ($P = 1.000$) with more moderate recall ($R = 0.412$). In both outputs, rank 2 is the weakest level, with zero precision and recall.

This pattern is consistent with the main comparison discussed earlier. As in the task-oriented methodology, the face-oriented LLM output is strongest on the upper part of the scale, while intermediate relevance remains more difficult to recover. By contrast, the generic LLM output resembles the task-agnostic methodology in that agreement remains limited across ranks and does not yield a stable recovery of the intermediate levels.

Table 9: Ground-truth to prediction transition counts for the exploratory LLM-based ranking outputs. Rows correspond to ground-truth ranks and columns to predicted ranks.

	Generic semantic ranking				Face-oriented semantic ranking			
GT	Pred 1	Pred 2	Pred 3	Pred 4	Pred 1	Pred 2	Pred 3	Pred 4
1	4	1	1	2	1	0	0	0
2	0	0	6	5	3	0	0	0
3	0	0	2	16	2	3	7	5
4	2	0	5	13	0	0	0	20

Rank-weighted aggregation analysis. For both LLM outputs, rank-weighted micro exceeds rank-weighted macro. In the generic case, the gap remains moderate (0.346 vs. 0.280), indicating limited agreement distributed across the scale. In the face-oriented case, the gap is larger (0.753 vs. 0.559), showing that performance is concentrated on the upper ranks, where both reliability and weights are higher.

This behavior mirrors the pattern already observed for the main task-oriented methodology, where stronger recovery of high ranks benefits rank-weighted micro more than rank-weighted macro.

Rank transitions (GT → Pred). Table 9 reports the ground-truth to prediction transition counts for the two LLM-based outputs. The generic ranking shows a marked upward drift toward high predicted ranks: all GT rank-2 labels are reassigned to predicted ranks 3 or 4, and most GT rank-3 labels are promoted to predicted rank 4. The face-oriented ranking is considerably better aligned with the upper part of the scale: all GT rank-4 labels remain at predicted rank 4, and most GT rank-3 labels are mapped to predicted ranks 3 or 4.

The dominant residual failure mode in the face-oriented output concerns GT rank 2, which is consistently mapped to predicted rank 1. This is the same structured pattern already observed for the main task-oriented methodology, although here it is measured on a much smaller subset. Conversely, the generic output is closer to the task-agnostic case in that it does not preserve a stable intermediate level and tends to shift relevance upward.

Overall, the exploratory LLM results follow the same qualitative ordering already observed in the main ranking comparison: the face-oriented formulation is clearly more consistent with the face-driven reference than the generic one. At the same time, the two LLM outputs also reproduce the main structural tendencies of the two baseline methodologies: the

generic formulation remains relatively close to the task-agnostic regime, whereas the face-oriented formulation shows stronger top-rank recovery together with a residual weakness on intermediate relevance.

4.3 Informativeness-measure diagnostics (map-level)

This section reports dataset-level summaries of the four global diagnostics introduced in Section 4.1.4 for the RONI informativeness score fields (entropy, edge density, saliency, and the adaptive mix). For each sample and informativeness measure, a score field is computed, and the following indicators are recorded:

- **Score dispersion:** coefficient of variation (CV_S).
- **Score concentration:** Gini coefficient (G_S).
- **Effective support at half mass:** $A_{50} = k^\star/n$.
- **Spatial scale:** correlation length (ℓ_{corr}, in pixels).

Table 10 aggregates these diagnostics as mean±std across samples.

Cue	CV_S	G_S	A_{50}	ℓ_{corr} (px)
entropy	0.25 ± 0.12	0.136 ± 0.058	0.404 ± 0.043	36.7 ± 22.2
edge	1.48 ± 0.72	0.316 ± 0.039	0.149 ± 0.079	25.7 ± 18.6
saliency	0.36 ± 0.12	0.192 ± 0.059	0.363 ± 0.044	54.8 ± 20.7
mix	0.53 ± 0.24	0.264 ± 0.084	0.306 ± 0.066	31.9 ± 20.2

Table 10: Mean ± standard deviation of the map-level diagnostics across the dataset for the four informativeness measures.

Concentration and effective support. Edge density is the most concentrated cue: it exhibits the highest dispersion and Gini coefficient, and the smallest half-mass support. The low A_{50} indicates that half of the sampling mass is typically concentrated on a relatively small fraction of pixels. This behavior is consistent with edge-driven score fields being dominated by thin, high-response structures (contours), which naturally produce peaky sampling distributions. Conversely, entropy is the most diffuse cue: it has the lowest CV_S and G_S, and the largest A_{50}, indicating that probability mass is spread over a wider portion of the RONI. The

adaptive mix occupies an intermediate regime, reducing the extreme concentration observed under edge while remaining more selective than entropy.

Spatial scale. Saliency exhibits the largest correlation length, indicating smoother score fields with larger coherent regions. Edge density has the smallest spatial scale, consistent with its high-frequency, contour-like structure. Entropy and the mix yield intermediate correlation lengths. Across all cues, the relatively large standard deviations (especially for ℓ_{corr}) indicate that spatial scale is strongly scene-dependent: some images induce very compact activations, while others produce broader, smoother score patterns.

Connection with patch-placement behavior. Overall, the map-level trends provide a qualitative interpretation of the selector–cue interactions observed in the end-to-end evaluation. Highly concentrated cues (notably edge) can favor repeated sampling around a small set of dominant locations unless the selector enforces strong dispersion/diversity. In contrast, smoother cues (notably saliency) provide broader high-score areas that are more compatible with diversity-promoting mechanisms. This interpretation is consistent with the patch-geometry indicators and the within-family summaries reported in Section 4.4.5.

4.4 Patch selection and end-to-end reconstruction results

4.4.1 Compared configurations

Patch-selection strategies are evaluated over a full-factorial grid obtained by pairing an informativeness cue with a patch-selector variant, yielding $5 \times 4 = 20$ configurations (Table 11). Informativeness is computed on the RONI using `entropy`, `edge_density` (`edge`), `saliency`, or an `adaptive mixture`. Selector variants include a score-guided covering baseline (SGC), farthest-point sampling (FPS), and DPP-based selection, with FPS and DPP each considered in both deterministic and probabilistic form.

All configurations operate on the same preserved ROI and place evidence within the RONI defined by the segmentation mask. Two evaluation regimes are considered.

Primary evaluation (fixed-seed setting). All 20 configurations are evaluated on $N = 500$ samples under a fixed random seed, so that probabilistic methods are compared under the

Table 11: Compared components and configuration space. Each selector variant is evaluated with each informativeness cue, for a total of $5 \times 4 = 20$ *configurations.*

Component	Variants
Informativeness cue (`metric`)	`entropy`, `edge`, `saliency`, `mix`
Patch selector (`algorithm`)	`SGC` (probabilistic), `FPS` (deterministic), `FPS` (probabilistic), `DPP` (deterministic), `DPP` (probabilistic)

same random realization of patch sampling; seed-induced variability is analyzed separately in the multi-seed subset. Unless stated otherwise, dataset-level comparisons (PF/SF, within-sample ranks, geometry, and timing) are obtained by aggregating metrics across samples in this fixed-seed setting.

Seed-sensitivity analysis (multi-seed subset). Probabilistic selectors (SGC, probabilistic FPS, probabilistic DPP) are additionally evaluated on a smaller subset of $N = 50$ samples under 5 random seeds. For each (sample, configuration), metrics are summarized across seeds through the seed mean and the seed-induced variability, reported as standard deviation and coefficient of variation (CV). Dataset-level robustness indicators are obtained by averaging these per-sample variability measures over the subset. Deterministic variants (deterministic FPS, deterministic DPP) do not depend on the seed in their patch placement and are excluded from seed-sensitivity statistics.

4.4.2 Patch-set geometry: overlap and dispersion

Patch-set geometry is characterized through two complementary indicators computed from patch coordinates. Patch overlap (%) measures redundancy and is defined as the fraction of patch-covered pixels that are covered more than once. Mean center distance (MCD, px) measures dispersion and is defined as the average pairwise Euclidean distance between patch centers.

Figure 14 shows that redundancy is primarily driven by the selector family, with an additional interaction with the cue. Deterministic FPS yields near-zero overlap under all cues, indicating that farthest-point placement effectively avoids re-covering the same background regions under the adopted patch size and coverage budget. At the opposite end, the SGC baseline exhibits the largest overlap, especially under the `edge` cue, suggesting that score-

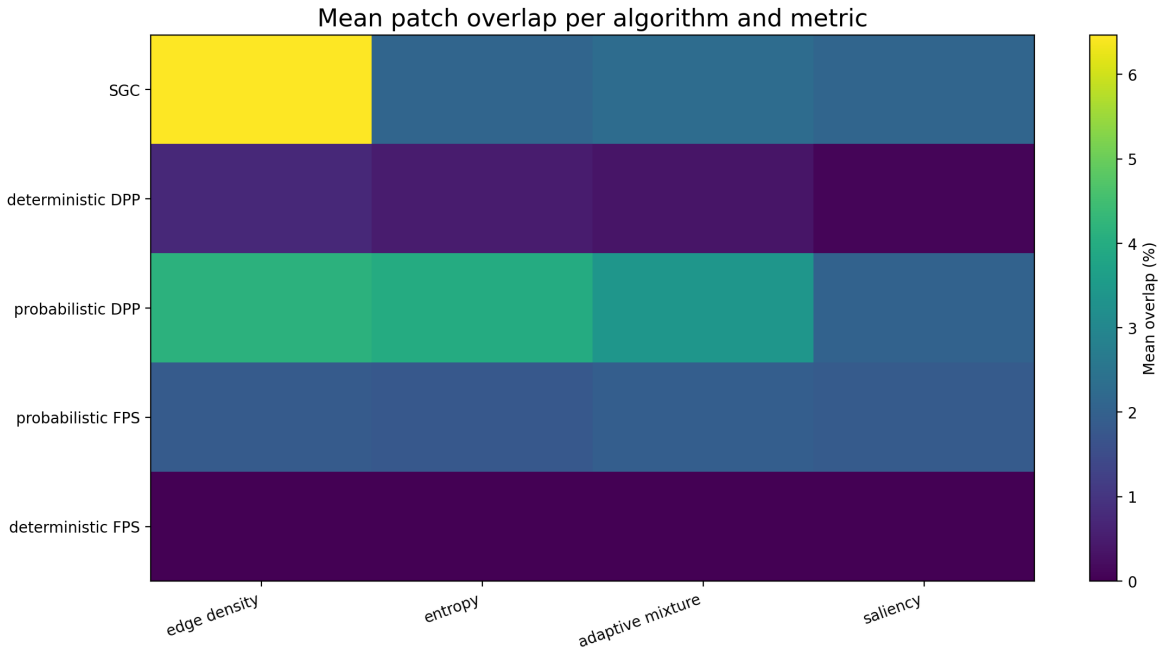


Figure 14: Mean patch overlap (%) per configuration. Lower values indicate less redundancy in the transmitted evidence.

guided sampling tends to concentrate patches on a small set of highly activated locations when the cue map is locally peaked. DPP-based variants reduce overlap compared to the baseline by explicitly promoting diversity; deterministic DPP is the most consistently overlap-suppressing among the non-FPS families, whereas probabilistic DPP remains more redundant, particularly under `edge` and `entropy`. Probabilistic FPS shows intermediate overlap, typically higher than deterministic FPS but lower and more stable than the baseline.

Figure 15 reports the corresponding dispersion patterns. Deterministic FPS achieves the largest MCD values across cues, confirming that it enforces the strongest spatial spread of evidence over the admissible RONI. Probabilistic FPS remains comparatively dispersed but forms a lower tier than deterministic FPS. DPP-based selectors exhibit moderate dispersion: their diversity mechanism promotes repulsion between selected patches, yet does not maximize dispersion as aggressively as farthest-point placement. The SGC shows the widest cue-dependent behavior, with markedly lower dispersion under edges (consistent with concentrated maps) and higher dispersion under smoother cues, such as `entropy` and `saliency`.

These geometric trends help interpret fidelity outcomes in Section 4.4.3. At a fixed nominal budget, lower overlap increases the effective covered area. At the same time, higher dispersion spreads evidence over a larger portion of the background, both of which are ben-

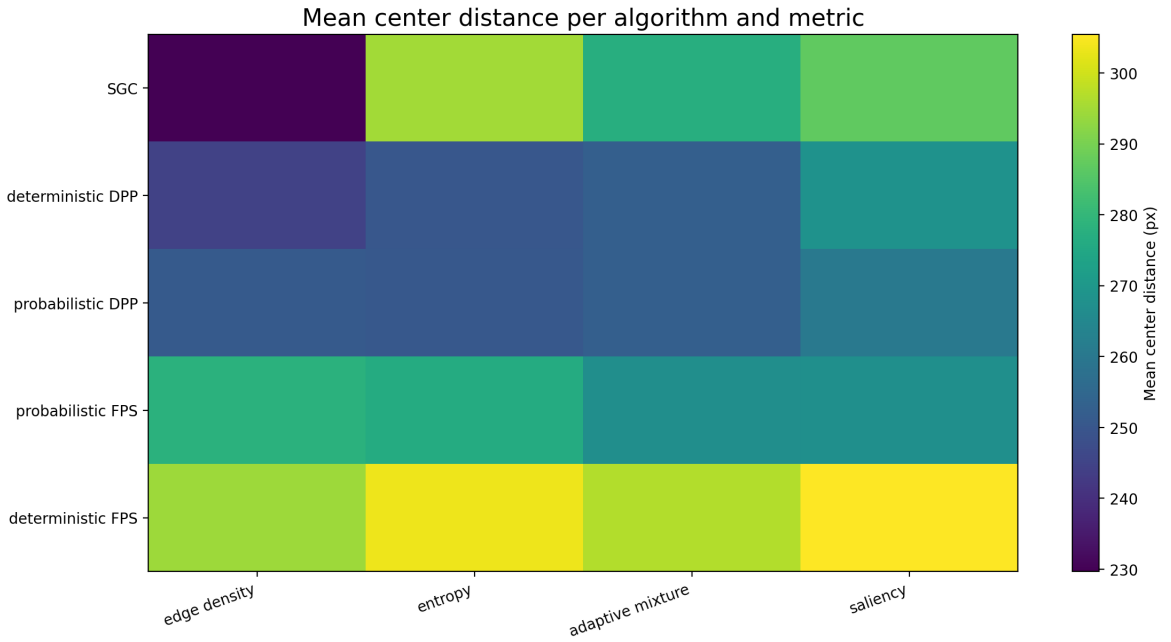


Figure 15: Mean patch-center distance (px) per configuration. Higher values indicate more spatially dispersed evidence.

eficial when the decoder must reconstruct the RONI from sparse conditioning. Accordingly, deterministic FPS combines near-zero redundancy with maximal dispersion and aligns with the strongest global PF/SF ranks. Conversely, the high overlap observed for the SGC under edge indicates that part of the budget is spent re-covering already conditioned regions, which is consistent with its weaker fidelity rankings under the same cue.

4.4.3 Reconstruction fidelity (PF and SF)

Reconstruction fidelity is evaluated end-to-end by running the fixed generative decoder conditioned on the patch evidence selected by each configuration in Table 11. This subsubsection is organized as follows. First, absolute reconstruction quality is reported through dataset-level summaries of perceptual fidelity (PF) and semantic fidelity (SF) (Tables 12 and 13). Then, relative comparisons are provided through within-sample rankings and performance profiles, which emphasize how consistently a configuration approaches the sample-wise best across diverse scenes.

PF quantifies perceptual similarity and is derived from LPIPS, mapped to a bounded fidelity score. SF quantifies semantic consistency and is derived from CLIP similarity, mapped to $[0, 1]$. Unless stated otherwise, PF and SF are reported under the primary fixed-

Table 12: Dataset-level PF summary under the primary fixed-seed setting ($seed = 42$), computed on the $N = 500$ samples with complete outputs for all 20 configurations. For each selector–cue pair, the first line reports mean ± standard deviation across samples, and the second reports the median with interquartile range $[Q_{25}, Q_{75}]$.

Selector variant	`entropy`	`edge`	`saliency`	`mix`
SGC (probabilistic)	0.612 ± 0.065 0.610 [0.569, 0.652]	0.592 ± 0.078 0.592 [0.551, 0.643]	0.616 ± 0.063 0.613 [0.569, 0.658]	0.616 ± 0.067 0.615 [0.569, 0.663]
FPS (deterministic)	0.629 ± 0.063 0.627 [0.584, 0.672]	0.631 ± 0.067 0.632 [0.581, 0.677]	0.627 ± 0.066 0.625 [0.582, 0.673]	0.631 ± 0.067 0.631 [0.583, 0.679]
FPS (probabilistic)	0.618 ± 0.064 0.617 [0.576, 0.657]	0.617 ± 0.063 0.616 [0.573, 0.661]	0.620 ± 0.063 0.621 [0.575, 0.659]	0.619 ± 0.065 0.619 [0.577, 0.663]
DPP (deterministic)	0.602 ± 0.075 0.596 [0.555, 0.649]	0.596 ± 0.069 0.591 [0.553, 0.644]	0.610 ± 0.067 0.606 [0.566, 0.656]	0.604 ± 0.071 0.598 [0.557, 0.651]
DPP (probabilistic)	0.603 ± 0.070 0.603 [0.557, 0.649]	0.606 ± 0.067 0.605 [0.558, 0.651]	0.609 ± 0.067 0.606 [0.564, 0.655]	0.606 ± 0.068 0.603 [0.562, 0.652]

Table 13: Dataset-level SF summary under the primary fixed-seed setting ($seed = 42$), computed on the $N = 500$ samples with complete outputs for all 20 configurations. For each selector–cue pair, the first line reports mean ± standard deviation across samples, and the second reports the median with interquartile range $[Q_{25}, Q_{75}]$.

Selector variant	`entropy`	`edge`	`saliency`	`mix`
SGC (probabilistic)	0.963 ± 0.022 0.967 [0.952, 0.978]	0.961 ± 0.021 0.966 [0.948, 0.978]	0.963 ± 0.021 0.968 [0.952, 0.978]	0.965 ± 0.020 0.968 [0.955, 0.979]
FPS (deterministic)	0.966 ± 0.020 0.971 [0.957, 0.980]	0.966 ± 0.020 0.971 [0.957, 0.980]	0.966 ± 0.021 0.971 [0.956, 0.980]	0.966 ± 0.019 0.970 [0.956, 0.980]
FPS (probabilistic)	0.963 ± 0.022 0.968 [0.956, 0.978]	0.962 ± 0.022 0.967 [0.952, 0.978]	0.963 ± 0.020 0.968 [0.953, 0.978]	0.963 ± 0.021 0.968 [0.955, 0.978]
DPP (deterministic)	0.963 ± 0.021 0.968 [0.952, 0.979]	0.961 ± 0.021 0.967 [0.951, 0.977]	0.962 ± 0.022 0.966 [0.951, 0.978]	0.964 ± 0.020 0.967 [0.953, 0.978]
DPP (probabilistic)	0.963 ± 0.021 0.967 [0.952, 0.978]	0.962 ± 0.022 0.967 [0.951, 0.977]	0.962 ± 0.021 0.966 [0.951, 0.978]	0.963 ± 0.021 0.968 [0.953, 0.978]

seed setting on $N = 500$ samples, and aggregated across samples to obtain dataset-level statistics. Seed-induced variability for probabilistic selectors is analyzed separately on the multi-seed subset ($N = 50$ samples) in Section 4.4.6.

Tables 12 and 13 show that PF exhibits a wider dynamic range than SF, while SF remains closer to saturation under the fixed decoder and the common evidence budget.

The tables also quantify absolute reconstruction quality under the primary fixed-seed setting. For PF, the configuration-wise means span [0.592, 0.631], indicating a non-negligible sensitivity to the evidence placement policy under a fixed coverage budget. Deterministic FPS provides the highest PF across cues, with the best mean attained under `edge` and `mix` (0.631). Probabilistic FPS yields intermediate PF values (means around 0.617–0.620), while SGC and DPP-based variants remain lower on average, with the weakest PF observed for SGC under `edge` (0.592). Across configurations, PF dispersion across scenes is substantial (std $\approx$ 0.063–0.078), suggesting that image-to-image variability is comparable to, or larger

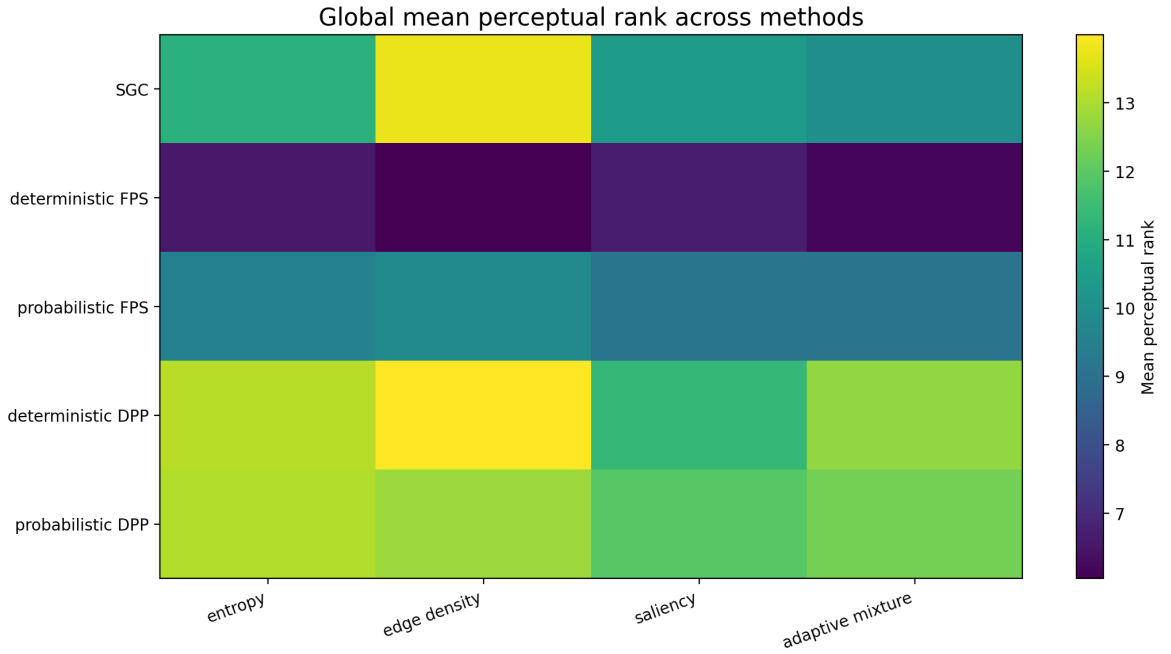


Figure 16: Global mean within-sample rank for PF (lower is better). Rows correspond to selector variants and columns correspond to cues.

than, several between-configuration gaps.

For SF, the dynamic range is markedly smaller: mean values lie in [0.961, 0.966] across all selector–cue pairs. Deterministic FPS remains consistently best, but the absolute differences are limited, coherently with saturation effects induced by the fixed decoder and the constrained evidence budget. SF variability across samples (std $\approx$ 0.019–0.022) is again non-negligible relative to the observed mean separations, motivating complementary relative analyses.

Median and interquartile ranges broadly track mean trends for both PF and SF, indicating that the above comparisons are not driven by a small number of outliers. In the following, within-sample rankings and performance profiles are therefore used to complement absolute summaries, highlighting how consistently each configuration approaches the sample-wise optimum across heterogeneous scenes.

Relative comparisons: ranks and performance profiles. Beyond absolute summaries, reconstruction fidelity is further analyzed through relative, within-sample comparisons. For each sample, configurations are ranked according to PF or SF, and ranks are then averaged across the $N = 500$ samples (rank 1 is best).

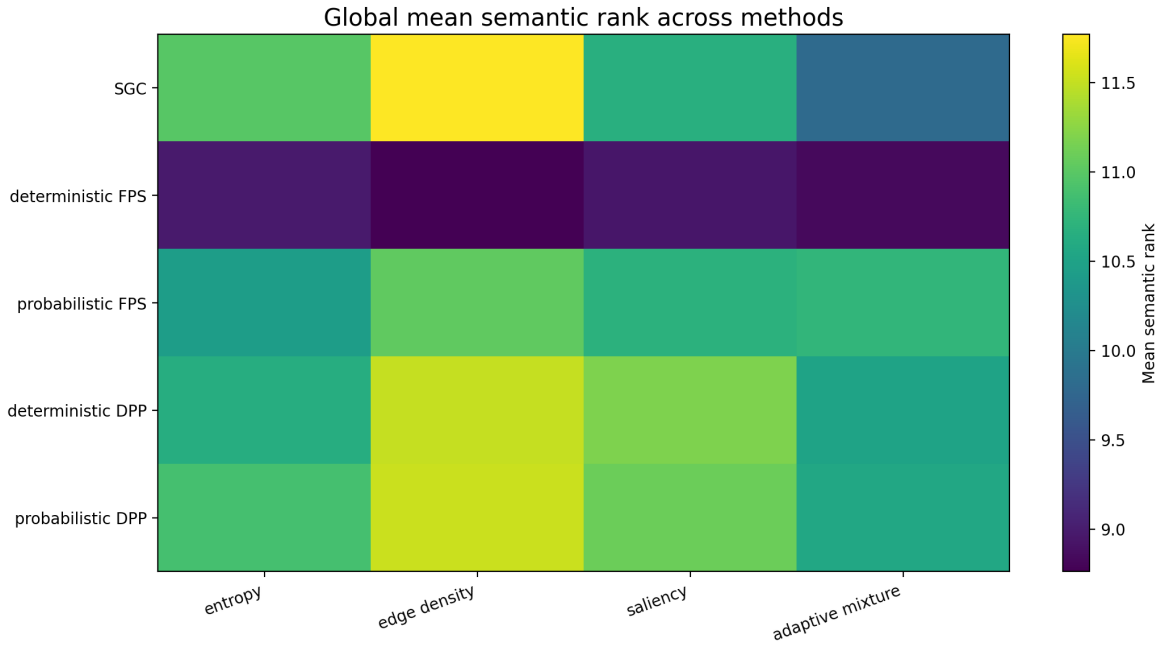


Figure 17: Global mean within-sample rank for SF (lower is better). Rows correspond to selector variants and columns correspond to cues.

Figure 16 shows that PF ranks are primarily separated by selector family. Deterministic FPS attains the best ranks under every cue, indicating that farthest-point placement yields the most reliable perceptual fidelity under a fixed evidence budget. Within deterministic FPS, `edge` and `mix` are the strongest cues. Probabilistic FPS forms a consistent second tier, with its best ranks obtained under `saliency`. In contrast, the baseline and DPP-based variants rank lower, and their performance degrades sharply under `edge`, suggesting that edge-driven evidence is beneficial only when the selector prevents spatial concentration.

For SF (Figure 17), rank contrasts are smaller, consistent with the reduced SF range in Table 13. Deterministic FPS remains best overall, while cue effects become more visible outside the FPS family. In particular, `edge` is consistently the weakest cue for the baseline and for both DPP variants, whereas `mix` yields the most competitive ranks among non-FPS families. This supports the view that, in a semantically saturated regime, cue design can partially compensate for weaker placement rules by providing more robust conditioning signals.

Performance profiles complement rank heatmaps by reporting, for each tolerance $\varepsilon \geq 0$, the fraction of samples for which a configuration attains a loss within ε of the sample-wise best. Equivalently, at a given ε, the curve value can be read as the empirical probability that

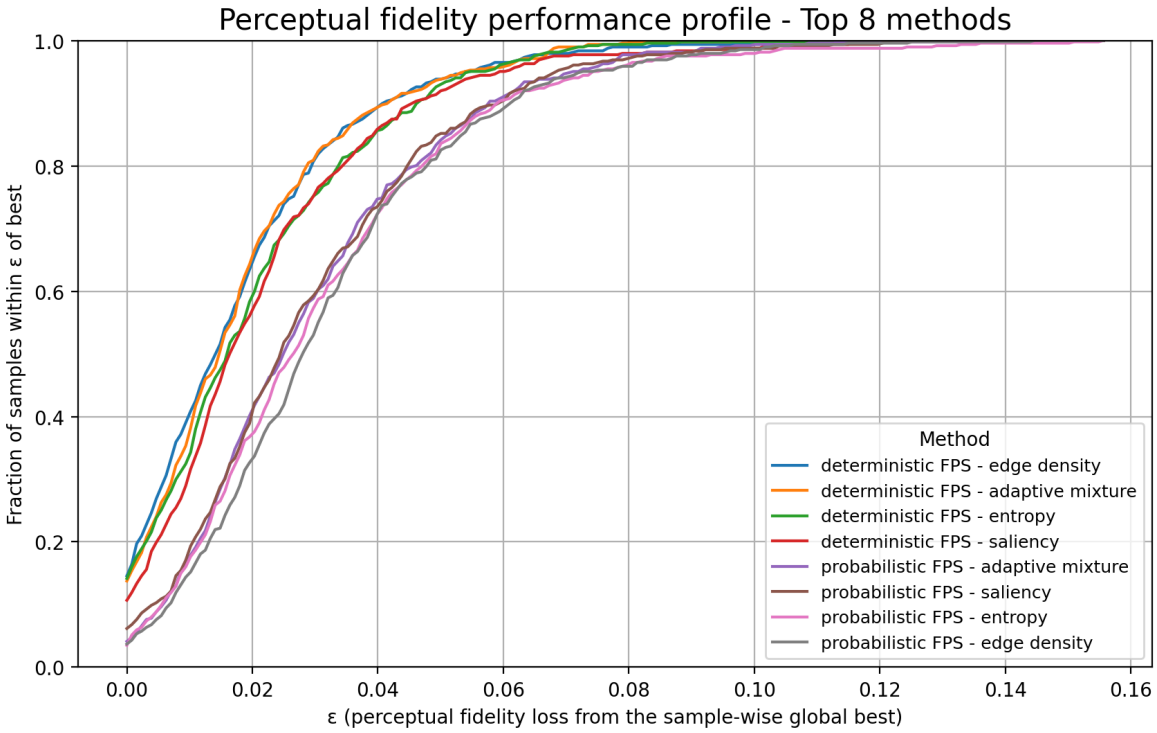


Figure 18: Top-8 PF performance profiles. Higher curves indicate a larger fraction of samples for which a configuration achieves PF values close to the best PF obtained on that sample.

the method is "ε-near-optimal" on a randomly drawn scene; higher curves therefore indicate more consistent near-best performance across heterogeneous samples.

For PF (Figure 18), the Top-8 set is entirely composed of FPS-based configurations (four deterministic FPS and four probabilistic FPS variants). Deterministic FPS dominates at small ε, confirming higher consistency in achieving near-optimal perceptual fidelity; deterministic FPS with `edge` provides the strongest profile in the tight-tolerance regime.

For SF (Figure 19), curves are tighter, and the Top-8 set includes additional `mix`-based configurations from the baseline and DPP families, consistent with weaker separations under SF. Overall, the profiles corroborate the heatmap trends: selector choice is the dominant factor for PF, while cue choice plays a comparatively larger role for SF.

4.4.4 Linking geometry to fidelity

Reconstruction outcomes are related to evidence geometry using global within-sample ranks (rank 1 is best). Geometry is summarized by mean patch overlap (%) and mean center distance (MCD, px), computed under the primary fixed-seed setting on the $N = 500$ samples.

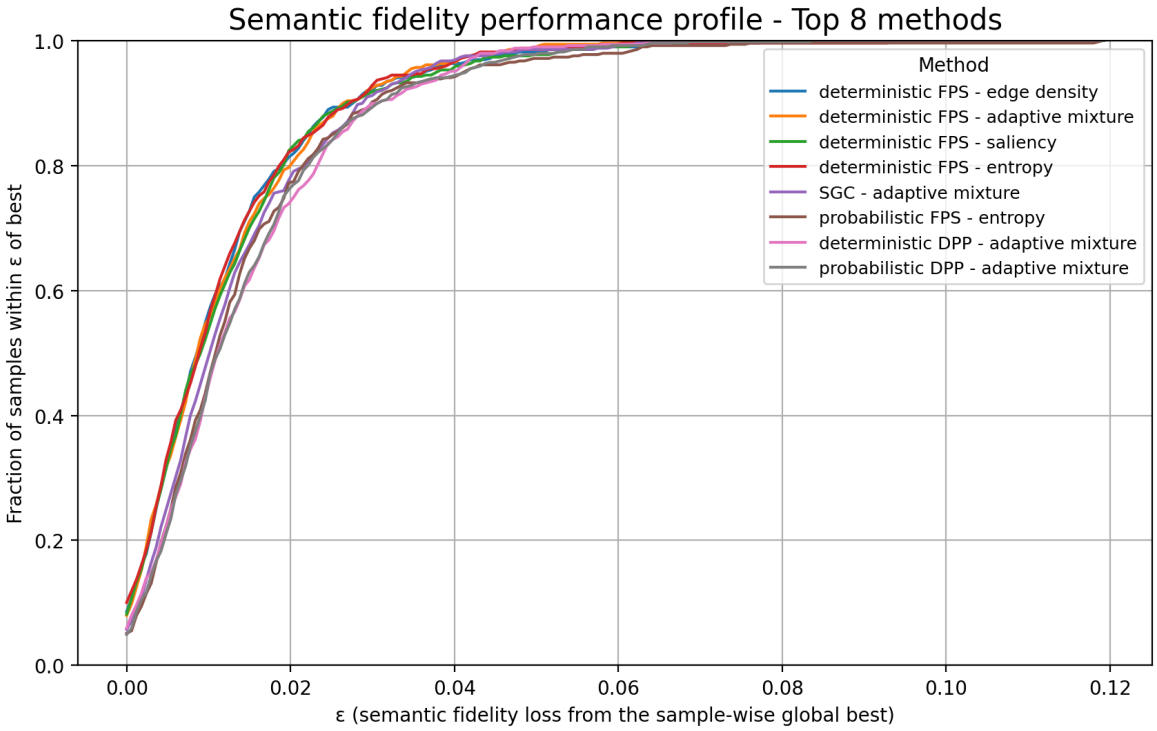


Figure 19: Top-8 SF performance profiles. Higher curves indicate a larger fraction of samples with SF close to the best value achieved for that sample.

Each point in the following scatter plots represents one configuration (selector variant × cue), and a least-squares trend line is included as a visual guide.

Figures 20 and 21 show a negative association between dispersion and rank: larger MCD values tend to correspond to lower (better) PF/SF ranks. Deterministic FPS forms the high-dispersion cluster and simultaneously attains the best ranks, supporting the interpretation that spatially spreading evidence is beneficial under a fixed coverage budget. Configurations with smaller dispersion, including the SGC and most DPP variants, concentrate at higher (worse) ranks. The association is visually tighter for PF than for SF, consistent with the larger PF dynamic range and the stronger separations previously observed for perceptual fidelity.

Figures 22 and 23 relate redundancy to fidelity rank and show a positive association: higher overlap tends to correspond to worse ranks. At fixed patch size and nominal budget, overlap reduces effective coverage of distinct background pixels, thereby limiting the spatial extent of the conditioning signal. This effect is most clearly visible for the baseline under the `edge` cue, which combines the largest overlap with the worst ranks, especially for SF. Conversely, deterministic FPS concentrates near-zero overlap and near-best ranks across cues.

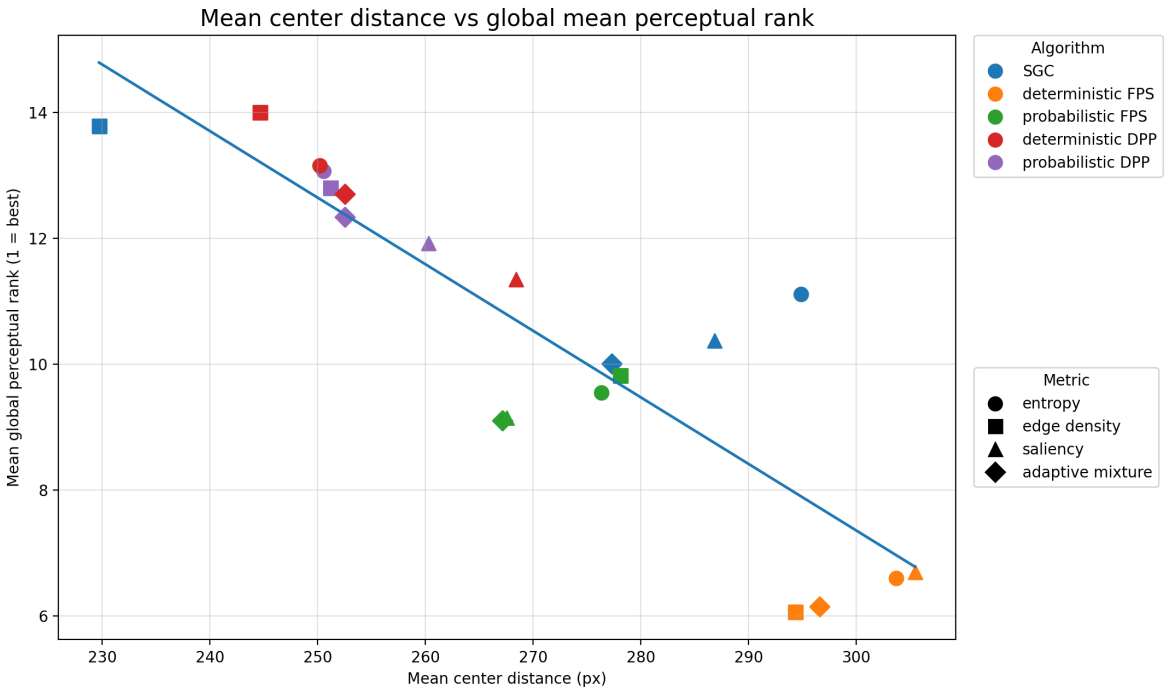


Figure 20: Mean center distance versus mean global PF rank (lower is better), with each point representing one configuration.

While the global trends are consistent, geometry is not sufficient to fully determine fidelity. Several configurations achieve moderate dispersion yet remain poorly ranked, indicating that cue-driven content (i.e., *what* is preserved) still matters in addition to placement (*where* it is preserved), particularly for SF. Similarly, some DPP configurations reduce overlap relative to the baseline but do not reach deterministic FPS ranks, suggesting that the strength of dispersion enforcement (maximal under farthest-point placement) is critical to translate diversity into fidelity gains.

Overall, two tendencies are supported under the evaluated budget: increasing dispersion (higher MCD) is associated with better global ranks, while increasing redundancy (higher overlap) is associated with worse global ranks. Both associations are stronger for PF than for SF, consistent with the reduced separations observed under semantic fidelity.

4.4.5 Dataset-level summaries

Dataset-level summaries aggregate within-sample comparisons across the $N = 500$ samples of the primary fixed-seed setting. Three indicators are reported: mean rank (average within-sample rank, rank 1 is best), win rate (fraction of samples where rank 1 is achieved), and mean

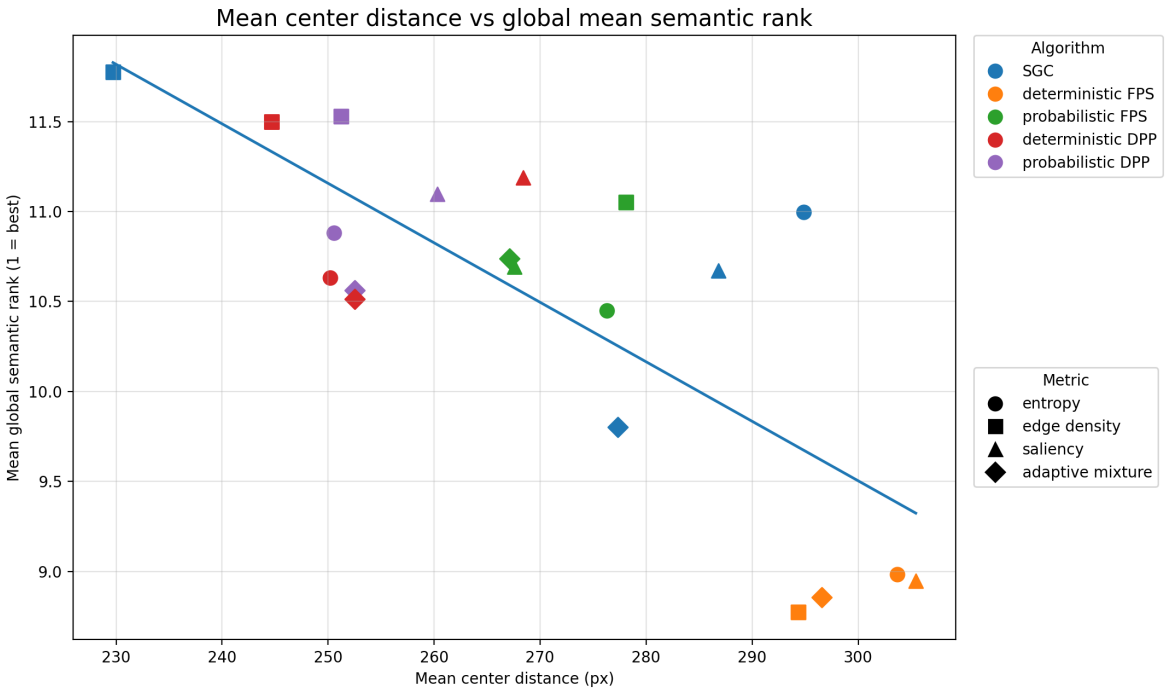


Figure 21: Mean center distance versus mean global SF rank (lower is better), with each point representing one configuration.

loss-to-best (average shortfall from the sample-wise best, $-\Delta_{\text{best}}$ as defined in Section 4.1.4). All indicators are computed separately for PF and SF.

Two aggregation views are considered. Within-family summaries isolate the effect of the cue by comparing cues under the same selector variant. Global summaries compare all 20 configurations jointly.

Within-family summaries. Table 14 reports, for each selector variant, the cue achieving the best average performance under PF and under SF. Cue preference depends on the objective: the SGC baseline favors `entropy` under PF but `saliency` under SF, while deterministic FPS favors `edge` under PF and `entropy` under SF. DPP-based variants consistently select `saliency`, whereas probabilistic FPS selects `adaptive mixture` for both objectives. Win rates remain well below 1, indicating that the within-family winner is not universal across scenes.

Global summaries. Tables 15 and 16 report the top-8 configurations under PF and under SF using global within-sample ranks. Under PF, the top positions are dominated by deterministic FPS across cues, confirming that selector choice is the main driver of perceptual fidelity;

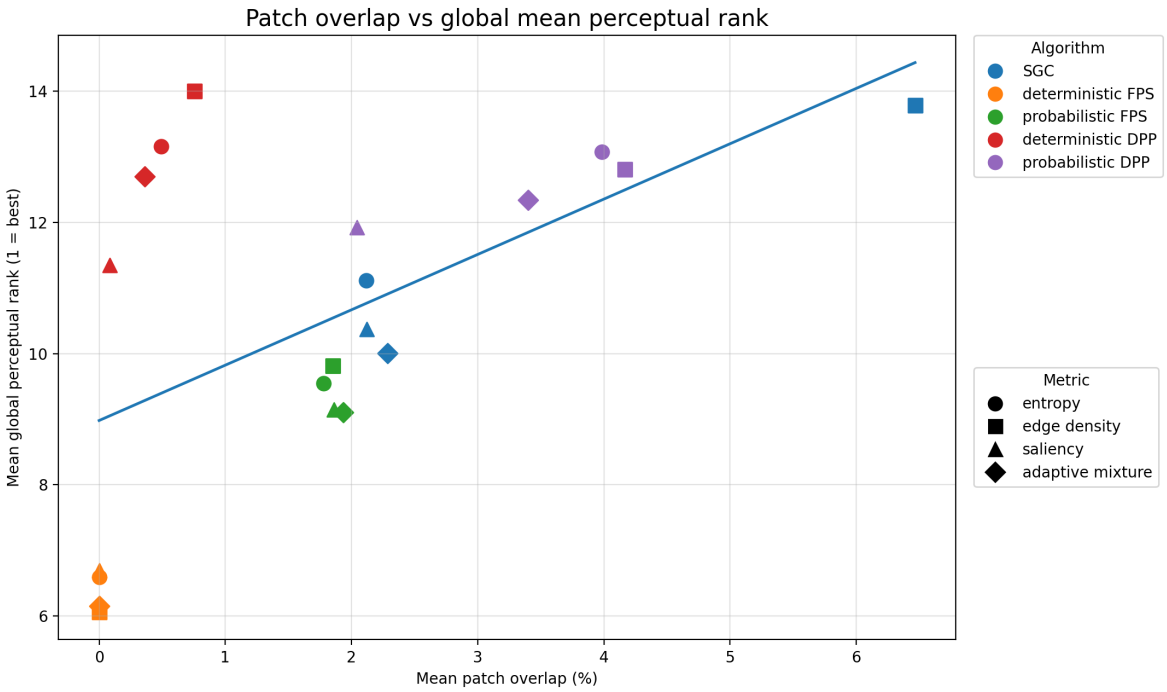


Figure 22: Mean overlap versus mean global PF rank (lower is better), with each point representing one configuration.

probabilistic FPS variants occupy the remaining top entries. Under SF, separations are smaller and additional families appear among the top configurations (notably deterministic DPP with `saliency`), indicating a comparatively stronger influence of cue design under a semantically saturated regime. In both objectives, win rates are far from 1, confirming that no single configuration dominates all scenes.

4.4.6 Seed robustness for probabilistic variants

Probabilistic selectors induce run-to-run variability through the random seed controlling patch placement. Robustness is assessed on the multi-seed subset comprising $N = 50$ samples evaluated under 5 seeds ($seed \in \{42, 43, 44, 45, 46\}$). Deterministic variants are seed-independent by construction and are therefore excluded from this analysis.

Geometry indicators. Patch-set geometry is computed for each seed and summarized through patch overlap (%) and mean center distance (MCD, px). For each (sample, configuration), seed variability is quantified via the standard deviation across seeds, σ_{seed}, and the

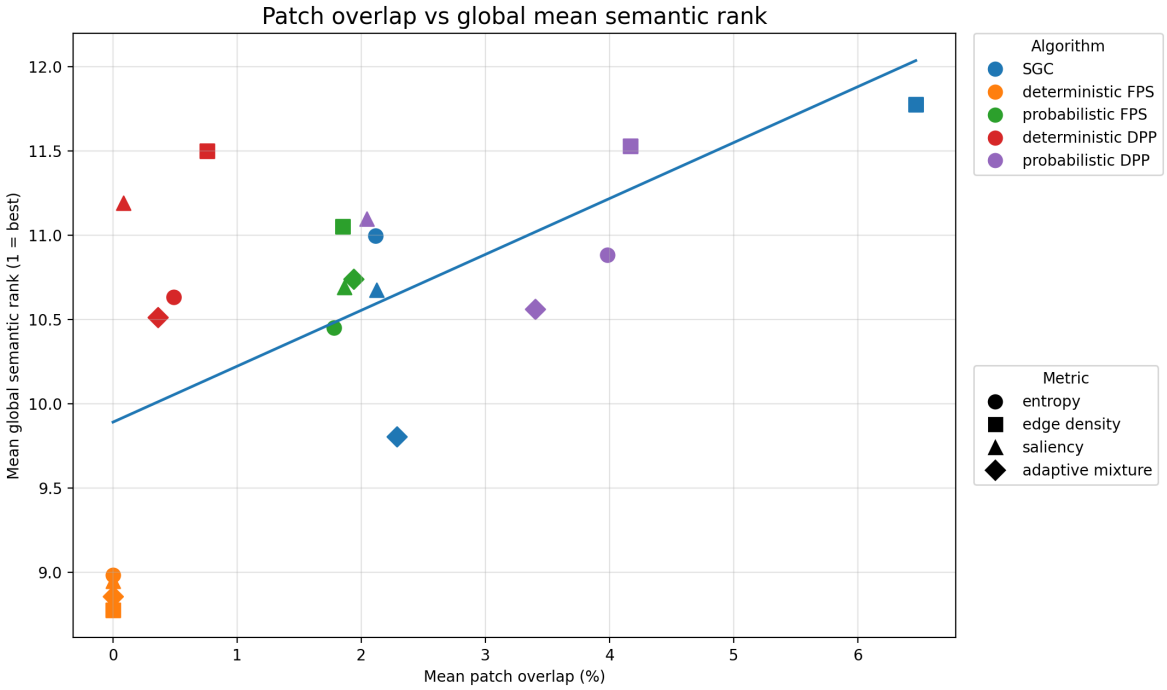


Figure 23: Mean overlap versus mean global SF rank (lower is better), with each point representing one configuration.

coefficient of variation (CV), defined as

$$\mathrm{CV} = 100 \cdot \frac{\sigma_{\mathrm{seed}}}{\mu_{\mathrm{seed}}},$$

where μ_{seed} is the mean of the metric across seeds. CV is therefore a relative (scale-normalized) measure of seed-induced variability, expressed in percent; it can become large when μ_{seed} is close to zero (e.g., for near-zero overlap). Dataset-level robustness values are obtained by averaging these per-sample quantities across the $N = 50$ samples. Table 17 reports the resulting statistics for the score-guided covering baseline (SGC), probabilistic FPS (`FPS`), and probabilistic DPP (`DPP_PROB`).

Mean overlap remains small in absolute terms, but exhibits the largest relative variability: overlap CV frequently exceeds 100% when the mean overlap is close to zero. This does not indicate large absolute fluctuations; rather, it reflects normalization by the seed mean. When overlap is near-zero, small seed-to-seed changes (e.g., between almost non-overlapping and mildly overlapping patch sets) yield large relative ratios. Accordingly, overlap robustness is best interpreted jointly through (mean, σ_{seed}, CV), with σ_{seed} providing the most direct measure of absolute seed-induced variation.

Table 14: Best cue within each selector variant under PF and SF in the primary fixed-seed setting ($N = 500$). Rank is reported as mean ± standard deviation. Win denotes the fraction of samples in which the cue is best within the same selector variant. Loss-to-best denotes the average fidelity gap from the best cue within that selector variant.

	PF (within-variant)				SF (within-variant)			
Selector variant	Best cue	Rank	Win	Loss	Best cue	Rank	Win	Loss
SGC (baseline)	`entropy`	2.02±1.06	0.42	0.0087	`saliency`	2.32±0.98	0.20	0.0049
DPP (probabilistic)	`saliency`	1.86±1.14	0.56	0.0052	`saliency`	2.38±1.18	0.34	0.0041
DPP (deterministic)	`saliency`	2.02±1.06	0.42	0.0089	`saliency`	2.06±1.11	0.44	0.0038
FPS (probabilistic)	`mix`	2.08±1.10	0.40	0.0062	`mix`	2.30±1.11	0.32	0.0031
FPS (deterministic)	`edge`	2.36±1.12	0.30	0.0110	`entropy`	2.34±1.12	0.32	0.0042

Table 15: Top-8 configurations under PF using global within-sample ranks (primary fixed-seed setting, $N = 500$). Each row corresponds to a full selector–cue configuration, ranked against all configurations. Rank is reported as mean ± standard deviation. Win denotes the fraction of samples in which a configuration is the best-performing one among all configurations. Loss-to-best denotes the average PF gap from the best configuration on the same sample.

Configuration	Rank (mean±std)	Win rate	Mean loss to best
deterministic FPS + `edge`	5.36±4.80	0.22	0.0130
deterministic FPS + `saliency`	5.60±4.50	0.18	0.0164
deterministic FPS + `mix`	5.74±4.72	0.16	0.0151
deterministic FPS + `entropy`	5.78±4.14	0.16	0.0140
probabilistic FPS + `mix`	7.72±4.50	0.04	0.0223
probabilistic FPS + `edge`	8.44±4.18	0.02	0.0244
probabilistic FPS + `saliency`	9.44±4.42	0.02	0.0262
probabilistic FPS + `entropy`	9.46±4.83	0.02	0.0268

In contrast, MCD is markedly more stable across seeds: its seed CV remains below $\approx 10\%$ for all probabilistic configurations. This indicates that, while the seed can affect local patch interactions (and thus overlap), the overall spatial spread of the selected patch set is largely preserved from run to run. Accordingly, dispersion is primarily determined by the selector family and the cue, with the seed acting as a secondary perturbation. Finally, the largest overlap CV values are observed in configurations whose mean overlap is close to zero, whereas the largest absolute overlap variability (in terms of σ_{seed}) is observed for the baseline under `edge`.

Reconstruction indicators. Reconstruction metrics are less sensitive to the seed than geometry. Across probabilistic configurations, the dataset-level mean of the per-sample seed standard deviation lies in $[0.016,\ 0.023]$ for PF and in $[0.008,\ 0.011]$ for SF. The correspond-

Table 16: Top-8 configurations under SF using global within-sample ranks (primary fixed-seed setting, $N = 500$). Rank is reported as mean ± standard deviation. Win denotes the fraction of samples in which a configuration is the best-performing one among all configurations. Loss-to-best denotes the average SF gap from the best configuration on the same sample.

Configuration	Rank (mean±std)	Win rate	Mean loss to best
deterministic FPS + `saliency`	7.34±5.26	0.08	0.0072
deterministic FPS + `entropy`	7.68±5.52	0.18	0.0066
deterministic DPP + `saliency`	8.90±5.76	0.10	0.0089
deterministic FPS + `mix`	9.26±6.20	0.06	0.0086
probabilistic FPS + `mix`	9.40±5.51	0.06	0.0091
deterministic FPS + `edge`	9.46±5.94	0.08	0.0096
probabilistic FPS + `edge`	9.84±5.46	0.06	0.0096
SGC (baseline) + `saliency`	10.26±5.41	0.00	0.0101

ing seed CV ranges are ≈ 2.6–4.1% for PF and ≈ 0.88–1.10% for SF. These values indicate that seed-induced variability in reconstruction quality is limited under the adopted budget and fixed decoder, and that SF is more stable than PF, coherently with its near-saturation regime. Accordingly, in the multi-seed subset PF and SF are appropriately summarized by their seed means, while seed-induced dispersion is retained as a robustness descriptor. This supports the use of fixed-seed PF/SF comparisons in the primary evaluation, with the multi-seed analysis providing evidence that the observed ranking trends are not dominated by random-seed effects.

4.4.7 Decoder inference time

Decoder-side cost is measured through diffusion inference time, defined as the wall-clock time required to produce a single reconstruction. The decoder architecture and inference schedule are kept fixed across all experiments; differences in inference time, therefore, reflect only conditioning effects induced by the transmitted evidence and run-time variability.

Inference times are reported under the same primary fixed-seed setting described in Section 4.4.1. Dataset-level statistics are obtained by aggregating per-sample inference times across samples. Table 18 reports the mean inference time (in seconds) for each selector–cue pair.

Inference time varies within a narrow range across configurations, with differences that are small compared to image-to-image variability. Most configurations cluster around ≈ 25.18 s, indicating that evidence placement has a limited impact on the overall diffusion runtime under

Table 17: Seed robustness of patch-set geometry for probabilistic selector variants on the multi-seed subset ($N = 50$, $seed \in \{42, 43, 44, 45, 46\}$). For each sample–configuration pair, the table reports the mean across seeds and the seed-induced variability, summarized by standard deviation and coefficient of variation over the five seeds. Dataset-level values are obtained by averaging these quantities across samples. Overlap is reported in %, MCD in px, and CV in %.

Configuration	Overlap	σ_{seed}	CV	MCD	σ_{seed}	CV
SGC + `entropy`	2.46	1.96	83.7	272.3	25.3	9.5
SGC + `edge`	5.33	2.87	68.4	246.5	21.4	8.9
SGC + `mix`	2.88	2.19	81.7	275.4	27.9	10.2
SGC + `saliency`	2.37	1.95	88.8	276.5	23.3	8.7
probabilistic FPS + `entropy`	1.86	1.91	109.5	267.0	18.9	7.1
probabilistic FPS + `edge`	2.12	2.13	113.8	267.5	21.6	8.2
probabilistic FPS + `mix`	1.99	1.99	117.5	265.9	17.1	6.5
probabilistic FPS + `saliency`	2.00	2.11	115.7	267.9	18.1	6.8
probabilistic DPP + `entropy`	3.66	3.00	98.0	247.1	20.9	8.6
probabilistic DPP + `edge`	3.89	3.13	97.3	251.0	23.7	9.7
probabilistic DPP + `mix`	2.97	2.43	100.2	254.3	20.6	8.3
probabilistic DPP + `saliency`	1.81	1.92	130.1	261.9	15.8	6.1

Table 18: Mean decoder inference time under the primary fixed-seed setting. Rows denote selector variants and columns denote informativeness cues.

Selector variant	`entropy`	`edge`	`saliency`	`mix`
SGC	25.348	25.838	24.972	24.992
DPP (probabilistic)	25.183	25.184	25.178	25.179
DPP (deterministic)	25.184	25.182	25.183	25.184
FPS (probabilistic)	25.171	25.185	25.186	25.190
FPS (deterministic)	25.329	25.203	25.104	25.141

the adopted conditioning interface. No systematic inference-time penalty is associated with the highest-fidelity family (deterministic FPS), suggesting that the PF/SF gains observed in Section 4.4.3 are not achieved by increasing decoder-side compute.

5 Conclusion

5.1 Summary of contributions

This thesis investigated encoder-side allocation policies for selective-fidelity semantic image transfer within a SPIFF-like architecture, under a fixed generative inpainting decoder. The study focused on two coupled decisions under a strict evidence budget: semantic ranking for ROI selection and budgeted patch sampling over the RONI to support decoder-side reconstruction.

A modular experimental framework was developed to enable controlled comparisons across both stages. For semantic ranking, two main methodologies were implemented on a shared set of semantic candidates produced by a GroundingDINO+SAM backend: a task-agnostic methodology based on CLIP semantic self-consistency, and a task-oriented methodology based on face-relevance aggregation derived from occlusion sensitivity in a face-recognition embedding space. In addition, an exploratory LLM-based ranking study was conducted on a reduced subset to provide a complementary perspective on prompt-based multimodal ranking within the same semantic-selection setting.

For RONI evidence sampling, four informativeness measures were implemented, namely `entropy`, `edge_density`, `saliency`, and an adaptive `mix`, together with five selector variants: SGC, deterministic FPS, probabilistic FPS, deterministic DPP, and probabilistic DPP. This yielded a $5 \times 4 = 20$ configuration space evaluated under a shared coverage constraint and a fixed decoder interface.

The evaluation protocol was organized into two complementary tracks. The first assessed ranking agreement against manually curated annotations under a face-driven relevance criterion. The second evaluated end-to-end reconstruction quality for patch selection by jointly reporting perceptual and semantic fidelity, patch-set geometry, seed robustness for probabilistic variants, and runtime indicators. Overall, the study combined two complementary perspectives: a systematic comparison of encoder-side policies and an interpretable analysis of how allocation choices affect end-to-end reconstruction outcomes under fixed-decoder assumptions.

5.2 Key findings

The ranking results should be interpreted with respect to the adopted reference criterion rather than as a direct competition between task-oriented and task-agnostic formulations. On the 780 annotated samples, the task-oriented methodology achieved rank-weighted F_1 values of 0.741 (macro-averaged) and 0.803 (micro-averaged), whereas the task-agnostic methodology reached 0.283 and 0.290, respectively. Since the annotations define relevance in terms of face-recognition utility, the stronger agreement achieved by the task-oriented methodology is best understood as evidence of effective alignment with that objective. By contrast, the lower agreement of the task-agnostic methodology reflects the expected mismatch between a generic semantic-consistency objective and a face-driven notion of relevance. Within the task-oriented results, the main residual error concerned the under-ranking of person-associated but face-distant regions, suggesting that the aggregated face-relevance signal becomes weaker for segments only indirectly related to the facial area.

The exploratory LLM-based ranking results should be interpreted in the same way. On the reduced subset of 20 images, the face-oriented LLM output reached rank-weighted F_1 values of 0.559 (macro-averaged) and 0.753 (micro-averaged), whereas the generic variant reached 0.280 and 0.346, respectively. The central issue is therefore not simply whether the face-oriented prompting strategy outperformed the generic one, but how closely each prompting regime matched the adopted relevance criterion. Under this perspective, the face-oriented output showed substantially higher agreement with the reference, whereas the generic variant again reflected objective mismatch. Both LLM-based variants, however, remained below the dedicated task-oriented ranking pipeline, indicating that prompt-based multimodal ranking has not yet reached the same level of reliability in this setting.

For RONI evidence sampling, the dominant performance differences were associated more with the selector family than with the informativeness cue alone. Across the primary fixed-seed setting ($N = 500$), deterministic FPS yielded the most favorable overall fidelity profile, whereas probabilistic FPS generally ranked second, and SGC/DPP-based variants were less competitive on average. Perceptual fidelity also showed a wider dynamic range than semantic fidelity. Across the 20 evaluated configurations, mean PF spanned $[0.592, 0.631]$ around an overall average of about 0.613, whereas mean SF remained in the narrower range

$[0.961, 0.966]$ around an overall average of about 0.963. The best PF mean, equal to 0.631, was attained by deterministic FPS under the edge density and adaptive mixture cues. Under SF, deterministic FPS remained the best-performing family at a mean value of 0.966, but with a much smaller absolute margin, consistent with the near-saturated semantic-fidelity regime.

The analysis of patch-set geometry clarified the mechanisms underlying these trends. Configurations based on deterministic FPS consistently produced low-overlap and highly dispersed evidence sets, and higher spatial dispersion was associated with better reconstruction ranks, especially under perceptual fidelity. These findings support the interpretation that, under strict budgets, the usefulness of sparse evidence depends not only on local informativeness, but also on how effectively the selected patches constrain the image globally through spatial coverage.

Randomness introduced measurable but limited variability. For probabilistic selectors, seed sensitivity was more evident in overlap-based indicators than in reconstruction fidelity, whereas dispersion-related quantities remained comparatively stable. This suggests that randomness affects the exact local arrangement of evidence more than the overall reconstruction behavior.

Finally, decoder inference time remained essentially unchanged across configurations. The strongest-performing selector family did not incur a systematic decoder-side runtime penalty, indicating that the observed fidelity gains were attributable to better encoder-side evidence allocation rather than to increased reconstruction-side compute.

Taken together, these findings indicate that, under strict evidence budgets, reconstruction quality depends less on isolated local informativeness alone than on the ability of the encoder-side policy to distribute sparse evidence in a spatially informative and globally constraining manner.

5.3 Limitations and future work

The conclusions of this study should be understood in relation to the scope and assumptions of the adopted experimental framework. First, end-to-end reconstruction was evaluated under a single coverage budget and a fixed patch specification, so the analysis does not yet characterize operating curves across budgets, patch sizes, or mixed-granularity transmission regimes. Second, the main ranking evaluation was tied to a face-driven relevance definition and should

therefore be interpreted within that specific task setting. Third, the exploratory LLM-based study was conducted on a substantially reduced subset and was included as a complementary extension rather than as a core evaluation track; its findings should therefore be regarded as indicative rather than conclusive. More generally, all results inherit the assumptions of the adopted segmentation backend and fixed generative decoder, both of which constrain the feasible allocation space and may propagate upstream errors to downstream reconstruction.

Several directions emerge for future work. A first extension concerns multi-budget evaluation, to derive perceptual/semantic operating curves and identify progressive transmission regimes under variable resource constraints. This should include adaptive patch sizing, budget-aware selector scheduling, and hybrid strategies that interpolate between sparse evidence and denser explicit transmission.

A second direction concerns the generalization of semantic ranking beyond the face-recognition case study. Task-oriented and task-agnostic ranking should be evaluated against multiple task-matched reference criteria, so that their behavior can be assessed under different notions of relevance and their agreement with the corresponding annotation protocols can be characterized more systematically. In the same spirit, the exploratory LLM-based methodology could be extended through more systematic prompt design, larger evaluation subsets, and clearer integration into the semantic-candidate ranking framework, while preserving its role as a complementary multimodal perspective.

A third direction concerns a possible extension toward communication-level constraints. In the present study, transmission cost is modeled through a coverage-budget abstraction, which is convenient for controlled algorithmic comparison. A more complete networking-oriented evaluation could relate encoder-side evidence allocation to concrete communication quantities such as bitrate, latency, and protocol overhead.

Overall, the results support a clear conclusion within the adopted fixed-decoder setting: encoder-side allocation is not a secondary implementation detail, but a primary design dimension in selective-fidelity semantic image transfer. Semantic ranking determines which visual content is granted high-fidelity preservation, while RONI evidence placement determines how effectively the remaining scene can be constrained for decoder-side reconstruction. Under strict transmission constraints, reconstruction quality depends not only on the informativeness of the transmitted evidence, but also on how coherently that evidence is selected

and spatially distributed. In this sense, the study shows that meaningful reconstruction gains can be achieved through encoder-side policy design alone, even when the generative decoder is kept unchanged.

References

[1] Marco Palena, Jose A. Ayala-Romero, Andres Garcia-Saavedra, and Carla Fabiana Chiasserini. Spiff: Selective preservation of image fidelity for bandwidth-constrained heterogeneous networks. In *Proceedings of the IEEE INFOCOM 2026 – IEEE Conference on Computer Communications*, 2026.

[2] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[4] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations (ICLR)*, 2018.

[5] Yu Ren, Zhenming Ni, Xincen Ruan, and Ben Liu. Adaptive image semantic communications: A mask-based dual-mode transmission approach for segmentation tasks. *Physical Communication*, 2025.

[6] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] Eirina Bourtsoulatze, David Burth Kurka, and Deniz Gündüz. Deep joint source-channel coding for wireless image transmission. In *ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[8] Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y. Zeevi. The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing*, 6(9):1305–1315, 1997.

[9] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012.

[10] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[13] Thrupthi Ann John, Vineeth N Balasubramanian, and C V Jawahar. Canonical saliency maps: Decoding deep face models. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(4):561–572, 2021.

[14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019.

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[16] Gemini Team et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[17] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-nms – improving object detection with one line of code. *arXiv preprint arXiv:1704.04503*, 2017.

[18] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3–4):379–423, 623–656, 1948.

[19] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[20] Robert Bridson. Fast poisson disk sampling in arbitrary dimensions. In *ACM SIGGRAPH 2007 Sketches*, 2007.

[21] Mohamed S. Ebeida, Anjul Patney, Scott A. Mitchell, Andrew A. Davidson, Patrick M. Knupp, and John D. Owens. Efficient maximal poisson-disk sampling. *ACM Transactions on Graphics*, 30(4), 2011.

[22] Florin-Alexandru Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

[23] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.

[24] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7514–7528, 2021.

[25] 3gpp tr 38.821: Solutions for nr to support non-terrestrial networks (ntn). Technical Report TR 38.821, 3rd Generation Partnership Project (3GPP), 2023. Release 16; v16.2.0.

[26] X. Luo, H.-C. Chen, and O. Guo. Semantic communications: Overview, open issues, and future research directions. *IEEE Wireless Communications*, 2022.

[27] Bo Guo, Z. Xiong, T. Q. S. Wang, T. Q. S. Quek, and Z. Han. Semantic communication-aware end-to-end routing in large-scale leo satellite networks. In *2024 IEEE Interna-*

tional Conference on Metaverse Computing, Networking and Applications (MetaCom), pages 137–142, 2024.

[28] Jia Guo and Jiankang Deng. Insightface: 2d and 3d face analysis project, 2025.

[29] Google. Structured outputs | gemini api | google ai for developers, 2026.

[30] Google. Models | gemini api | google ai for developers, 2026.

[31] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1209–1218, 2018.

[32] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.