# POLITECNICO DI TORINO

Master's Degree in Data Science and Engineering

Master's Degree Thesis

# Multimodal Arithmetic for Zero-Shot Composed Image Retrieval: A Contrastive Post-Pre-Training approach of Vision-Language Models

| Supervisors | Candidate |
|---|---|
| Prof. Giuseppe RIZZO | Marco MAGNANINI |
| Dr. Federico D'ASARO | |
| Dr. Luca CATALANO | |

November 2025

# Summary

Vision–Language Models (VLMs) have emerged as powerful general-purpose models, capable of transferring to a wide range of downstream tasks in a zero-shot manner. These models are typically trained with contrastive objectives on large-scale image-text datasets, aligning images and text into a shared embedding space. Although effective for many applications, tasks such as Composite Image Retrieval (CIR), which consists of retrieving a target image given a reference image and a natural language modification, pose unique challenges. Classical CIR approaches rely on curated triplet datasets (reference, query, target), which are difficult to scale and limited in diversity.

This work introduces Multimodal Arithmetic Loss (MA-Loss), a training objective that learns compositional reasoning directly from readily available image-text pairs, eliminating reliance on costly triplet supervision. Unlike triplets, which require manual curation and annotation, image–text pairs can be collected at scale from the web, making them a practical foundation for large and diverse datasets. MA-Loss models semantic differences as structured transformations in a shared embedding space, aligning textual modifications with corresponding visual changes. This formulation enables CIR in a zero-shot setting while scaling naturally to heterogeneous web-sourced data.

To ground the design of MA-Loss, we conduct a systematic study of multimodal arithmetic using the SIMAT benchmark, analyzing the relationship between embedding space geometry (e.g., modality gap, alignment, uniformity) and compositional reasoning ability. Experiments show that a CLIP model post-pre-trained on MSCOCO using the MA-Loss objective achieves a new state of the art on SIMAT with a 48% score, surpassing the previous best of 42%.

Applying MA-Loss to CIR in a zero-shot setting, we evaluate on FashionIQ and CIRR benchmarks. Although using a relatively small dataset for post-pre-training, our method achieves results comparable to similar state-of-the-art pair-based approaches, while outperforming others on both benchmarks. These findings suggest that modeling semantic differences rather than absolute representations offers a scalable and effective alternative for compositional retrieval tasks.

# Acknowledgements

I would like to express my sincere gratitude to Fondazione LINKS, whose support and research environment have been fundamental to the development of this thesis. The opportunity to work within such a stimulating and collaborative center has greatly enriched my academic and professional growth.

I am deeply thankful to my supervisors, Prof. Giuseppe Rizzo, Dr. Federico D'Asaro, and Dr. Luca Catalano, for their continuous guidance, insightful feedback, and encouragement throughout this work. Their expertise, patience, and dedication have been invaluable at every stage of this thesis.

# Table of Contents

# List of Tables

# List of Figures

# Acronyms

**VLM**

Vision-Language Model

**SoTA**

State of The Art

**MA**

Multimodal Arithmetic

**CIR**

Composite Image Retrieval

**ZS-CIR**

Zero-Shot Composite Image Retrieval

**DNN**

Deep Neural Networks

**ViT**

Vision Transformer

**MSPD**

Mean Squared Pairwise Difference

**XSC-SR**

X-modal Semantic Consistency based on Squared Residuals

# Chapter 1

# Introduction

Vision–Language Models (VLMs) have become a central paradigm in modern multimodal learning. Their ability to align images and natural language into a shared embedding space has enabled strong zero-shot performance on tasks that range from classification to retrieval. This progress has been driven by large-scale image–text datasets and contrastive learning objectives that encourage paired samples to occupy nearby regions of the embedding space. Although this paradigm has produced widely adopted general-purpose models, it also reveals limitations when the target application requires reasoning about transformations rather than absolute descriptions.

Among the tasks that expose these limitations, Composite Image Retrieval (CIR) plays a significant role. CIR requires a model to retrieve a target image given a reference image and a natural language description of how the target differs from the reference. This process cannot rely exclusively on recognizing the content of individual images, but it also requires the ability to interpret relative changes such as color, shape, or context. Classical approaches rely on curated triplets that explicitly define reference, modification, and target. Triplet-based datasets, however, are costly to produce and limited in diversity, which restricts the scalability of CIR systems and their adaptation to open-world settings.

Recent interest in multimodal arithmetic has highlighted an alternative perspective. Multimodal arithmetic investigates the extent to which semantic transformations can be expressed as vector operations in a shared embedding space. This idea is inspired by the regularities observed in word embeddings, where relationships between concepts can be captured by linear transformations. Early work has shown that current VLMs exhibit only partial support for these operations, but also that this capability depends strongly on geometric properties of the embedding space, such as modality alignment, uniformity, and the modality gap. Understanding these properties and the learning objectives that govern them is therefore crucial for improving compositional reasoning.

This thesis investigates whether multimodal arithmetic can serve as a foundation for zero-shot CIR without relying on triplet supervision. The work begins with a systematic study of multimodal arithmetic using the SIMAT benchmark. This study analyzes how geometric properties of the embedding space influence the ability to perform cross-modal compositional reasoning. It also revisits the role of the modality gap, showing that the alignment of modality delta vectors is more important than their average separation.

Building on these insights, the thesis introduces the Multimodal Arithmetic Loss (MA-Loss), a new training objective that learns to model semantic differences using only image–text pairs. MA-Loss aligns transformations across modalities by encouraging the model to represent changes in a consistent way. Since it operates on image–caption pairs, the method eliminates the need for triplet supervision and allows training on large, widely available datasets. The thesis investigates MA-Loss analytically and empirically. It studies how the loss interacts with contrastive learning, how it influences embedding space properties, and how it behaves during post-pre-training.

The proposed methods are first evaluated on multimodal arithmetic, where MA-Loss improves over the classical contrastive objective and achieves a new state of the art on SIMAT. The methods are then applied to zero-shot CIR and evaluated on FashionIQ and CIRR. Despite being trained only on image–caption pairs and using a moderate dataset size, it reaches performance that is comparable to, and in some cases better than, existing zero-shot approaches.

Overall, this thesis shows that modeling semantic differences rather than absolute representations is an effective strategy for compositional retrieval tasks. It also demonstrates that contrastive post-pre-training guided by multimodal arithmetic principles can enhance the generalization ability of VLMs. This opens a path toward multimodal systems that can adapt to new tasks without requiring significant annotation effort. It also contributes to the understanding of how embedding space geometry influences cross-modal reasoning.

## 1.1 Outline of the thesis

This thesis is structured as follows: in Chapter 2, we propose an overview of the key works relevant to this research, starting with a general introduction to VLMs and their embedding space properties, followed by previous approaches related to our chosen tasks. In Chapter 3, we present a comprehensive theoretical framework to analyze the principles of multimodal arithmetic and to ground the design of our learning objective. In Chapter 4, we empirically evaluate our proposed methods and compare them against the current state of the art. Finally, in Chapter 5, we summarize the main conclusions based on our findings and outline potential directions for future research.

# Chapter 2

# Related Works

In this chapter, we will present the main works relevant to this thesis. We will start in Section 2.1 with a brief overview of vision-language models (VLMs), considering the evolution of visual recognition paradigms, the main architectural solutions used in VLMs, the most effective learning objectives, and a more specific discussion about one of the most innovative VLMs actually existing. We will then discuss in Section 2.2 the main geometrical and statistical properties relevant to characterizing an embedding space, spanning from traditional properties like uniformity, variance, and alignment, to the more recent phenomenon of the modality gap; we will then also discuss some proposed approaches to bridge the modality gap. Finally, we introduce the two main downstream tasks relevant for this thesis: multimodal arithmetic (Section 2.3) and composite image retrieval (Section 2.4). For both of them, we will give an overview of the most effective approaches and the main benchmarks used to evaluate performance.

## 2.1 Vision-language models

Vision-Language Models (VLM) have recently established themselves as a versatile and effective paradigm in computer vision, particularly due to their ability to perform zero-shot transfer across a wide range of tasks. These models are typically trained on extensive image–caption datasets using a contrastive learning objective that aligns visual and textual representations into a joint embedding space, enabling general-purpose applications.

### 2.1.1 Evolution of visual recognition paradigms

As well outlined by a recent survey [64], VLMs are the last step in an evolution process of visual recognition paradigms. Early computer vision solutions relied on *end-to-end deep learning* from scratch, which required large amounts of task-specific labeled data and incurred substantial computational costs. Subsequently, the *pre-training, fine-tuning, and prediction* paradigm emerged, where a model is first pre-trained on a large annotated dataset and then fine-tuned for each specific task. While this improved efficiency and performance, it still necessitates task-specific labeled data for effective adaptation.

More recently, the VLM approach of *pre-training and zero-shot prediction* has changed the landscape by utilizing vast amounts of web-scale image-text pairs for model training. This framework allows a single model to learn rich correspondence between images and their natural language descriptions. As a result, the model can directly perform downstream visual recognition tasks without the need of additional fine-tuning. This advancement helps to overcome limitations related to labeled data scarcity and task dependency.

In addition, Yamaguchi et al. [58] have introduced the concept of *post-pre-training*. Different from fine-tuning, this paradigm prescribes a domain-agnostic training phase to be carried out after pre-training in order to refine the properties of the embedding space to improve generalization and zero-shot transfer capabilities.

### 2.1.2 Architectures

VLM architectures are usually made of two primary components: an image encoder and a text encoder. These components work together to project the inputs into a joint embedding space. Image encoders are typically either convolutional neural networks (e.g., ResNet [23]) or Vision Transformers (ViT) [16]. Text encoding is generally realized using transformer-based language models like [52], or variants [15] [43].

From an architectural point of view, there exist two main families of models:

- **Dual-encoder models** (*two-towers* models) which separately encode images and text and use metrics like cosine similarity or dot product for retrieval or classification tasks. They facilitate scalable retrieval by allowing precomputation of embeddings. This family includes models like CLIP [44], and ALIGN [26].

- **Cross-modal transformers** (*one-tower* models) which integrate image patches and text tokens in a fused transformer architecture with cross-attention layers allowing rich interactions between modalities for detailed vision-language understanding. Examples are CLIPPO [51], OneR [24], and UNITER [10].

Additionally, a third hybrid architecture has been proposed. In this case, a dual-encoder architecture is followed by a fusion module to enrich cross-modal reasoning. Such models are exemplified by FLAVA [48], and CoCa [62].

While cross-modal transformers excel in tasks requiring deep reasoning (e.g., VQA, captioning, etc.), dual-encoders excel in zero-shot and retrieval settings due to their scalability and training simplicity.

### 2.1.3 Pretraining objectives and strategies

The foundation of effective VLMs lies in their pre-training methods, which aim to learn aligned vision-language representations that generalize to various downstream tasks. Pre-training leverages large-scale datasets of image-text pairs and employs one or more of the following learning objectives:

- **Contrastive learning**: Models learn to distinguish matching image-text pairs from non-matching pairs within a batch, maximizing similarity of true pairs while minimizing similarity of (negative) mismatched pairs. This objective underpins CLIP [44] and ALIGN [26]. Contrastive training is efficient and scalable, and enables strong zero-shot transfer by tying image embeddings to text descriptions.

- **Masked modeling**: Inspired by masked language modeling in NLP (e.g., BERT [15]), masked image modeling (MIM) and masked language modeling (MLM) objectives have been extended to vision-language pre-training [10] [48]. These models mask parts of the input and learn to reconstruct them conditioned on the remaining content.

- **Image-Text Matching**: they enforce VLMs to align paired images and texts by learning to predict whether the given text describes the given image correctly. They are commonly used as auxiliary losses, e.g., in FLAVA [48], and nCLIP [66].

A significant advancement is also the use of web-scale noisy image-text pairs for pre-training, which dramatically increases data diversity and size beyond curated datasets like COCO or Visual Genome. This scale has been vital for achieving strong generalization and zero-shot transfer capabilities.

### 2.1.4 CLIP

CLIP (Contrastive Language-Image Pre-training) proposed by [44]. It is a pioneering model that leverages an infoNCE contrastive objective [40] to jointly train an image encoder (either ResNet [23] or ViT [16]) and a text transformer [43]. The pre-training is conducted using a proprietary dataset of over 400 million image-text pairs collected from the web. Essentially, CLIP learns to maximize the cosine similarity between the embeddings of matching image-text pairs while minimizing the similarity between non-matching pairs. Refer to Section 3.1.2 for more details on CLIP's training objective.

CLIP differs from traditional image classification pre-training because it learns to ground images in natural language descriptions, instead of curated labels. This process makes it highly flexible and enables zero-shot transfer to a broad spectrum of vision tasks without supervised fine-tuning. For example, CLIP achieves zero-shot ImageNet [14] accuracy comparable to supervised baselines while using none of ImageNet's labels during training.

Architecturally, CLIP has a two-tower structure, and projects both modalities into a shared embedding space where cosine similarity determines alignment. CLIP-based classifiers can be created using text prompts such as "a photo of {class description}" and evaluating the similarities of the prompt's embedding against the image. This allows CLIP to quickly generalize to a set of virtually unlimited classes. CLIP models have also demonstrated robustness to distribution shifts, outperforming comparable supervised ImageNet-trained models on out-of-distribution datasets [44]. Their zero-shot performance also shows effective embedding quality, competitive with few-shot and sometimes supervised linear probe classifiers.

The success of CLIP has inspired numerous follow-up works using contrastive language-image pre-training with varied datasets, architectures, and improved objectives. It stands as a foundational pillar within vision-language research, widely adopted for tasks ranging from retrieval and classification to image generation guidance.

## 2.2 Geometric properties of the embedding space

The geometric landscape of latent embedding spaces is central to the success of modern multimodal and vision–language models. Among the most crucial geometric properties are uniformity, alignment, variance, and the so-called modality gap. These properties govern how information from disparate modalities is represented, compared, and transferred in downstream tasks such as classification, image retrieval, and semantic multimodal arithmetic. A more formal definition of these properties will be given in Section 3.1.3.

### 2.2.1 Uniformity, variance, and alignment

In the context of contrastive learning, the uniformity and alignment of embeddings are recognized as two fundamental objectives. Uniformity describes the extent to which representations are spread over the embedding hypersphere, ideally maximizing mutual information by reducing representational collapse or overcrowding. Alignment refers to the proximity of paired (e.g., image-text) embeddings, seeking to bring corresponding pairs as close as possible in the latent space without collapsing all samples to one point. Wang and Isola's [55] analysis formalizes these intuitions, showing that maximizing alignment (minimizing the distance between positive pairs) and promoting uniformity (spreading out negative pairs) are both desirable and synergistic in well-calibrated unimodal settings.

Other empirical studies [32][17][47][61] have extended these concepts to the multimodal regime; for instance, in CLIP-based models, optimal uniformity and alignment are associated with competitive performance on downstream tasks. However, the introduction of multiple modalities brings additional complexity, as the optimization objectives for uniformity and alignment can become antagonistic, particularly when the two modalities possess heterogeneous information structure or content [47][18][32].

### 2.2.2 The modality gap

A key phenomenon emerging in multimodal embedding spaces is the modality gap: a geometric separation between the submanifolds populated by different modalities, such as images and texts. This gap manifests as non-overlapping hypercones of representations on the embedding hypersphere and has been consistently observed in state-of-the-art models like CLIP and its variants [32][47]. An illustrative example is presented in Figure 2.1.

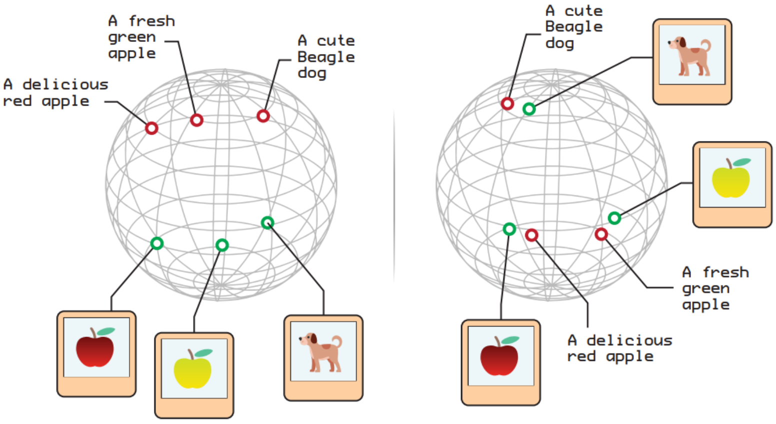
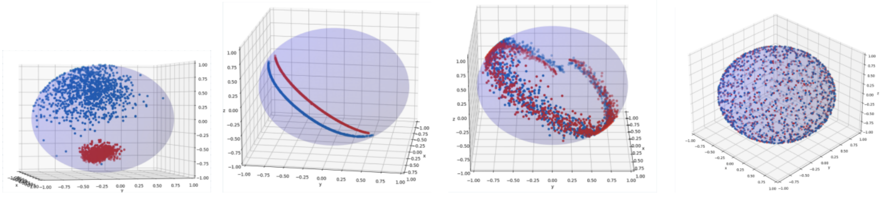


**Figure 2.1:** Illustrative example of the modality gap. On the left, the embeddings of the two modalities form disjoint clusters. On the right, the gap has been mitigated, and the two modalities are spread across the hypersphere. Source: [17].

Several works provide a multifaceted explanation for this phenomenon. From a theoretical perspective, model initialization induces the so-called "cone effect", where deep neural representations shrink to narrow cones; this leads distinctively different random initializations or architectures to produce separable cones for each modality (Figure 2.2)[32][47][18]. During training, the contrastive learning objective then preserves or even amplifies these separations unless explicit penalization is enforced. The local minima of the multimodal contrastive loss favor solutions where modalities stay separated, further stabilized by factors such as the value of the temperature parameter in InfoNCE losses [61][47].



(a) **Epoch 0** $I \rightarrow T$ accuracy: 0.0

(b) **Epoch 37** $I \rightarrow T$ accuracy: 0.0

(c) **Epoch 150** $I \rightarrow T$ accuracy: 0.1

(d) **Epoch 275** $I \rightarrow T$ accuracy: 0.87

**Figure 2.2:** 3D embeddings dynamics during training. At initialization, the two modalities are completely separate due to the cone effect. As the training progresses, the two modalities gradually spread across the hypersphere but stay disjointed. It is only after many epochs that the contrastive objective succeeds in bridging the gap between the modalities. Source: [18].

Although some early research suggested that closing the modality gap might not be necessary [32], or even potentially detrimental depending on the application, more recent evidence [61][18] points to its subtle impact: adjusting the gap can yield quantifiable benefits for downstream classification, retrieval, fairness, and multimodal arithmetic tasks. For instance, Liang et al. [32] show that careful manipulation (instead of closure) of the gap can improve zero-shot accuracy and reduce bias in demographic classification. Other studies confirm that reducing the gap by translation or alignment can enhance SIMAT [12] capabilities, enabling more consistent and meaningful multimodal embedding arithmetic [18]. Finally, a more recent work [56] has shown that the modality gap has a negative impact on out-of-distribution detection.

### 2.2.3 Strategies for improving embedding space properties

The geometric properties discussed above directly translate to downstream performance. Models with better alignment and controlled modality gap consistently outperform baselines on tasks such as classification, image retrieval, and multimodal arithmetic. Multiple strategies have emerged for closing or managing the modality gap:

- **Loss Engineering**: Augmenting the standard contrastive loss with explicit uniformity and alignment penalties (see Section 3.1.4), as proposed by a recent work [18], directly improves the geometric structure of the representation space.
- **Loss tuning**: Adjusting loss-related hyperparameters like temperature has also been shown to mitigate the modality gap [61][47].
- **Parameter-Sharing and Architectural Innovations**: Methods such as those in AlignCLIP [17] show that sharing encoder weights between modalities and introducing intra-modality separation can substantially reduce the gap, also improving alignment, general downstream robustness, and cross-modal retrieval. A similar approach [56] uses cross-modality mappings to enforce image-text consistency and reduce the gap.
- **Latent Space Translation**: Techniques using algebraic (e.g., Procrustes) transformations [36] or anchor-based translation [38] allow for zero-shot stitching of encoder/decoder pairs trained in disjoint spaces. These approaches leverage underlying isomorphisms present in well-trained models, supporting model reuse and composability across architectures and even modalities.

Each of these solutions brings strengths and caveats. For example, latent translation and relative representations unlock compositionality and robustness but may require sufficient anchor correspondences and careful handling of non-isometric factors. Loss-based methods are more broadly applicable but may trade off performance on certain tasks if the fine structure of the gap is not respected. Finally, some recent explanations reframe the modality gap as a contrastive gap, an inherent property of contrastive objectives, suggesting that future progress may require fundamentally new learning formulations [18].

To conclude, the geometric properties of the embedding space are deeply interconnected with the effectiveness of multimodal representations for downstream tasks. The collective evidence from recent research emphasizes that neither total closure nor total disregard of the modality gap is optimal. Instead, targeted manipulation achieved through architectural design, loss engineering, and geometric alignment enables more generalizable, fair, and high-performing multimodal systems, enhancing the utility and interpretability of large-scale vision language models.

## 2.3 Multimodal arithmetic

Multimodal arithmetic refers to the manipulation and combination of embeddings from multiple modalities, notably vision and language, by applying vector arithmetic operations in a shared embedding space. This concept is inspired by the well-known linguistic analogy property observed in word embeddings, where embeddings exhibit consistent geometric relationships (e.g., $king - man + woman \approx queen$) [37]. Early works in natural language processing established these regularities as fundamental properties of word embeddings, enabling algebraic manipulation of semantic concepts.

Recent studies extend this paradigm to multimodal embeddings that unify images and text representations. After CLIP [44] was introduced in 2021, it demonstrated strong zero-shot classification and retrieval capabilities, but it was initially unclear whether it exhibited multimodal arithmetic properties analogous to language models.

Couairon et al. [12] explicitly explored semantic image transformations conducted by vector arithmetic in joint text-image embedding spaces. They introduced the SIMAT dataset (Section 4.2.2) to quantitatively evaluate text-driven image transformations, where given a source image and a text-based delta vector (e.g., changing "cat" to "dog"), the goal is to retrieve a visually corresponding transformed image. Such a process is depicted in Figure 2.3. Their findings suggested vanilla CLIP embeddings were limited in supporting such arithmetic; however, simple linear adaptation based on datasets like MSCOCO (Section 4.2.1) improved this property considerably. Additionally, incorporating pretrained text encoders such as FastText [7], LASER [2], and LaBSE [20] contributed to better performance in semantic image transformations. This work opens the path to quantifying and improving multimodal arithmetic for visual generation and retrieval tasks. A later work [18] investigated the relationship between SIMAT performance and the modality gap (Section 2.2), revealing that loss functions designed to bridge the modality gap also improved multimodal arithmetic capabilities.

Multimodal arithmetic also underlies various applications in text-driven image editing and retrieval. For instance, methods in image retrieval explore composing queries by combining reference images with modification texts to find target images that satisfy the textual instructions, an approach exemplified in composed image retrieval datasets such as CIRR (Section 4.2.4) and FashionIQ (Section 4.2.3). These datasets and tasks rely on multimodal representations that capture fine-grained semantic relationships to support arithmetic-like transformations between image and text embeddings.

Beyond retrieval, text-driven image manipulation methods employ multimodal arithmetic in latent spaces for controlled editing. Patashnik et al. [42] introduced StyleCLIP, combining CLIP embeddings with generative models (StyleGAN [27]) for semantic editing of images through vector arithmetic on latent vectors. Similar approaches [45] have leveraged arithmetic properties to manipulate image attributes via textual instructions in various generative adversarial and diffusion frameworks.

Complementary studies in multimodal representation learning aim to enhance

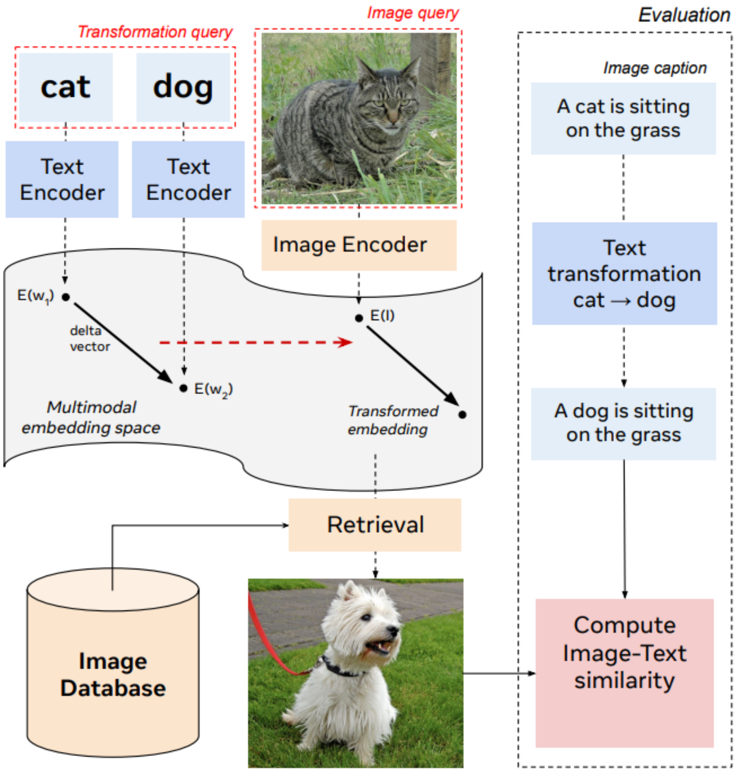


**Figure 2.3:** Example of multimodal arithmetic based on a textual delta vector. The textual embeddings are subtracted to obtain a delta vector; this is then summed to the query image embedding to obtain the transformed image embedding. The result is used to retrieve the most relevant image from a database. Source: [12].

the structure and interpretability of joint embeddings. Works such as Jia et al. [26] have scaled image-text alignment training, demonstrating emergent geometrical regularities and transfer capabilities in massive multimodal datasets. Others [60] have investigated fine-grained alignment of regional visual features and textual tokens to support detailed manipulation and retrieval.

In summary, multimodal arithmetic represents a growing research frontier to harness the geometric properties of joint vision-language embeddings for image transformation, retrieval, and generation. Current progress shows promise through advances in pretraining objectives, multimodal datasets, and fine-tuning strategies that enhance latent space arithmetic for real-world visual and linguistic concept interaction.

## 2.4 Composite image retrieval

Composite Image Retrieval (CIR) is an innovative and challenging task in computer vision and multimodal research. Similar to the traditional image retrieval task, the goal is to retrieve a target image from a database; this task is based on a multimodal query formed by a combination of a reference image and natural language-based modifications. Unlike traditional image retrieval methods that rely on either image or text queries alone, CIR requires models to understand and integrate multimodal cues effectively to retrieve target images that reflect the specified differences in the reference image. This task has significant practical applications, including interactive search in e-commerce, design, and content creation, where users iteratively refine their search results by providing visual examples and textual instructions (see Figure 2.4). A recent comprehensive survey [49] synthesizes the broad landscape of CIR research, encapsulating datasets, fusion techniques, architectures, evaluation protocols, and emerging challenges. CIR remains a dynamic research frontier, progressively driven by larger datasets, more expressive embeddings, and increasingly powerful multimodal fusion strategies.

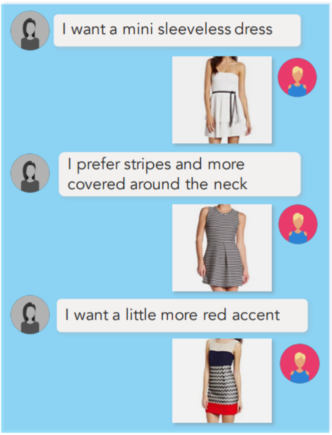

I want a mini sleeveless dress
I prefer stripes and more covered around the neck
I want a little more red accent

**Figure 2.4:** Example of a practical application of CIR. Unlike a traditional retrieval system, CIR prescribes an iterative refinement of the retrieval results. Source: [22]

.

### 2.4.1 Benchmarks evolution

Early efforts in CIR [22][53] primarily focused on simpler domains such as fashion and abstract geometric shapes. Guo et al. [22] introduced the FashionIQ dataset (Section 4.2.3), which consists of fashion images annotated with human-generated relative captions. This dataset supports interactive retrieval scenarios where the system refines search results based on detailed natural language feedback about attributes like color, style, and texture. FashionIQ laid an important foundation for CIR research by highlighting the complexity of real-world user demands and the need for compositional understanding of visual and linguistic modalities. Other earlier datasets for CIR, like CSS (Composed Synthetic Shapes) [53], present a synthetic collection of simple geometric 3D objects with descriptions generated automatically according to visual differences. CSS provides a controlled environment to study compositional image retrieval, focusing on basic shape and color transformations rather than complex real-world scenes.

To broaden the scope of CIR research, another study [34] introduced the CIRR dataset (Section 4.2.4), which includes over 36,000 triplets containing a reference image, a text-based modification, and the target image to be retrieved. CIRR focuses on real-world scenes that are open-domain and feature high levels of semantic and visual complexity. It challenges models to interpret complex textual instructions while distinguishing subtle image differences, demanding sophisticated visiolinguistic reasoning. It also introduces a novel evaluation metric, RecallSubset, to address common pitfalls in CIR benchmarking, like false negatives, and to provide a more reliable assessment framework.

### 2.4.2 Previous approaches

Recent approaches in CIR have explored various strategies, broadly categorized into zero-shot and supervised methods, with further distinctions based on whether they use textual-inversion techniques or vector operations in the embedding space.

#### Zero-shot CIR Approaches

Zero-shot (ZS-CIR) methods aim to perform composed retrieval without requiring annotated triplets for training. They leverage pre-trained VLMs like CLIP [44] and apply innovative fusion or manipulation techniques to enable retrieval in a training-free or minimally supervised manner.

**Textual-inversion-based methods** develop learnable pseudo-word tokens that represent the visual content of the reference image, which are then combined with the modification text tokens to form a text query. The core idea is to invert image embedding features back into the text token space to enable multimodal fusion fully within the text encoder (Figure 2.5). Models like PALAVRA [11] and SEARLE [3] introduce optimization-based textual inversion (OTI) and a mapping network to generate pseudo-word tokens. These tokens capture the visual content of reference images while aligning them semantically with real tokens in the CLIP text embedding space. iSEARLE [1] improves on SEARLE by injecting Gaussian noise and employing similarity-based hard negative sampling for better generalization and reduced modality gap. Pic2word [46] pioneers coarse-grained textual inversion by training a lightweight mapping network from image embeddings to learn a pseudo-token that represents an entire image as a "word". LinCIR [21] extends textual inversion to fine-grained scenarios using a self-masking projection network to flexibly replace keywords in the text with projected latent embeddings, enhancing retrieval performance in zero-shot settings.

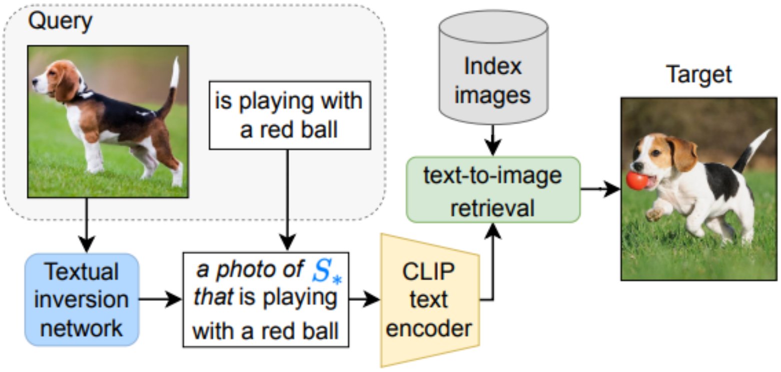

Query
is playing with a red ball
Index images
Target
text-to-image retrieval
Textual inversion network
a photo of $S_*$ that is playing with a red ball
CLIP text encoder

**Figure 2.5:** Inference process of ZS-CIR based on textual inversion. The query image is transformed into a text token that can be directly fused with the relative caption. Source: [3].

**Pseudo-triplet-based approaches** generate synthetic triplets by leveraging large language models (LLMs) or masking strategies to automatically construct training data, thus allowing traditional supervised training paradigms without manual annotations. MTI [8] adopts a masked learning approach where masked images and associated captions simulate the composed query, and the original image plays the role of the target, enabling the model to learn from unlabeled data efficiently.

**Training-free methods** operate without any further training by exploiting pre-trained VLM models and direct embedding space operations. Slerp [25] uses spherical linear interpolation between reference image and text embeddings in the joint embedding space, blending visual and semantic cues for effective zero-shot retrieval. WeiMoCIR [57] employs weighted modality fusion and similarity measures, leveraging multiple captions generated by multimodal large language models (MLLM) to improve the cross-modal matching without model training. CIReVL [28] and SEIZE [59] utilize large language models to transform composed queries into natural language captions suitable for off-the-shelf image retrieval systems, adding local re-ranking mechanisms to improve accuracy.

### Supervised CIR Approaches

Supervised methods rely on annotated triplets of reference image, modification text, and target image for training. They primarily focus on designing sophisticated fusion networks, metric learning losses, and image-text alignment strategies to improve query-target matching.

**Vector operation-based supervised methods** treat the textual modification as an embedding space transformation. CLIP4CIR [4], and its later improvement CLIP4CIR2 [6], stand out as important models leveraging multimodal arithmetic principles. They perform fine-tuning of CLIP encoders, learning a residual vector in the joint embedding space that corresponds to the relative caption, and apply this residual to the reference image embedding (see Figure 2.6), effectively achieving controlled semantic manipulation analogous to vector arithmetic seen in unimodal contexts. This method yields significant performance gains due to integrating the strong prior knowledge of CLIP's multimodal alignment. Combiner [5] similarly combines reference image and modification text features via learned fusion networks but emphasizes adaptive spatial and component-wise fusion to retain more fine-grained features.

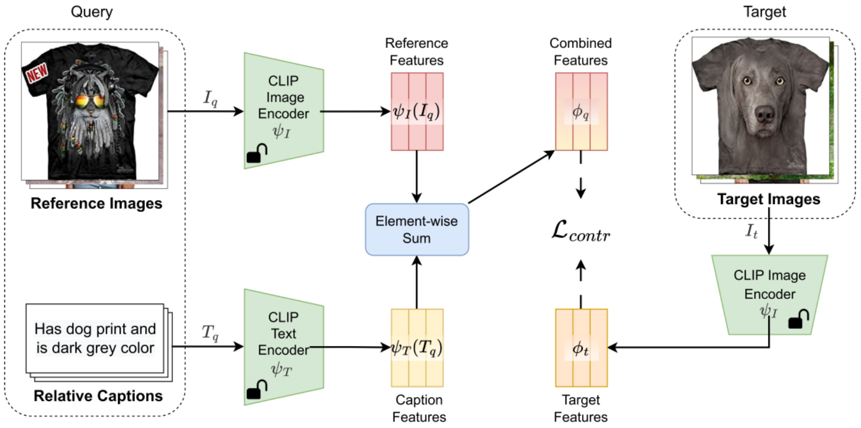


**Figure 2.6:** CIR supervised approach of CLIP4CIR. Relative captions are directly fed to the text encoder. Reference image and relative caption embeddings are combined using vector operations to obtain the combined features, which, in turn, are used to retrieve the target image. CLIP's weights are fine-tuned with a contrastive approach. Source: [6].

Additionally, zero-shot textual-inversion techniques have been integrated into end-to-end supervised frameworks: models like SEARLE [3] and iSEARLE [1] also have supervised versions leveraging annotated triplets for better inversion of visual content and improved alignment with textual modifiers.

Other supervised models utilize neural attention mechanisms or graph representations to enhance the fusion of multimodal inputs, such as SPIRIT [9] (style-guided patch interaction) or JAMMA [63] (joint attribute manipulation with graph attention) for attribute-level reasoning.

Overall, the CIR landscape shows a complex interaction between zero-shot and supervised methods, with textual-inversion techniques pushing the boundaries of zero-shot capabilities by mapping images to pseudo-textual tokens, while vector operation methods exploit embedding arithmetic principles to model semantic modifications as continuous shifts, achieving powerful and interpretable composition in retrieval tasks.

# Chapter 3

# Methods

In this chapter, we will discuss our theoretical approach to the problem of multimodal arithmetic and composite image retrieval. We will start in Section 3.1 with a general background, describing the notation we will use, the structure and properties of the classical contrastive learning objective, the main statistical and geometrical properties of the embedding space, and some auxiliary losses proposed by previous works. In Section 3.2, we will formalize the concept of *cross-modal semantic consistency*, present a theorem that links this property to Multimodal arithmetic performance, and design a metric to globally evaluate the cross-modal consistency of an embedding space. We will also analyze this metric and discuss its global minima and their relationship with the structure of the embedding space. Finally, in Section 3.3, we will design a loss to effectively train a VLM on an image-text pair dataset, applying the insights obtained in the previous section. We will also analyze some particular behavior of this loss, some limitations, and some possible solutions to overcome them.

## 3.1 Background

### 3.1.1 General notation

In this thesis, we decide to adopt a similar notation to a recent work [61]. Consider a set of $N$ paired training samples $\{(x_i^t, x_i^v)\}_{i=1}^N \subseteq \mathbb{R}^{d_t} \times \mathbb{R}^{d_v}$, where each pair $(x_i^t, x_i^v)$ consists of samples from the textual and visual modalities, respectively. We define the encoders of CLIP as $f_\theta : \mathbb{R}^{d_t} \to \mathbb{R}^d$ for textual inputs and $g_\phi : \mathbb{R}^{d_v} \to \mathbb{R}^d$ for visual inputs. By fixing $\theta$ and $\phi$ as the model's weights, we can evaluate the encoders to obtain the textual embeddings $h_i^t = f_\theta(x_i^t)$ and the visual embeddings $h_i^v = g_\phi(x_i^v)$. These embeddings live in a joint vector space with finite dimension $d$. Furthermore, since $f_\theta(\cdot)$ and $g_\phi(\cdot)$ are usually followed by an L2-normalization layer, we can consider $h_i^t$, $h_i^v$ as $\mathbb{R}^d$ vectors with unitary L2-norm. This property allows us to characterize the embedding space as a unit hypersphere, where each encoding corresponds to a specific point on its surface.

### 3.1.2 Contrastive Language–Image objective

Like the majority of VLM pre-training techniques, CLIP[44] is pre-trained using a contrastive infoNCE loss [40]. The general idea for this objective is to pull matching (or positive) image-text pairs close to each other while pushing apart non-matching (or negative) pairs.

Formally, we start by defining a similarity measure for our embedding space. This is usually done by means of the *cosine similarity.* Given two vectors $a$, $b \in \mathbb{R}^d$, the cosine similarity $\cos(a, b)$ between the two has values in $[-1,1]$ and is defined as:

$$\cos(a, b) = \frac{\langle a, b \rangle}{\|a\|\|b\|} \tag{3.1}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, and $\| \cdot \|$ denotes the euclidean L2-norm. In practice, since our embeddings have a unitary L2-norm, we can simplify this definition by omitting the denominator. This function can either be seen as a similarity measure based on the angle between the two vectors, or as a way to compute the logits corresponding to the probability of the two vectors being semantically related (i.e., of the pair $(a, b)$ being a positive pair).

The CLIP optimization objective, hereby denoted by $\mathcal{L}_{\text{CLIP}}$, is composed of two unidirectional losses: one for *image-to-text* classification $\mathcal{L}_{\text{CLIP}}^{V\to T}$, and the other for *text-to-image* classification $\mathcal{L}_{\text{CLIP}}^{T\to V}$. They are both formulated as temperature-scaled cross-entropy losses, structured to maximize the similarity of positive pairs $(k = i)$ against the similarity of negative ones $(k \neq i)$. Analytically, the image-to-text classification loss is defined as:

$$\mathcal{L}_{\text{CLIP}}^{V\to T} = -\frac{1}{N}\sum_{i=1}^{N} \log \left[ \frac{\exp(\langle h_i^v, h_i^t \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle h_i^v, h_k^t \rangle / \tau)} \right] \tag{3.2}$$

and similarly:

$$\mathcal{L}_{\text{CLIP}}^{T\to V} = -\frac{1}{N}\sum_{i=1}^{N} \log \left[ \frac{\exp(\langle h_i^v, h_i^t \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle h_k^v, h_i^t \rangle / \tau)} \right] \tag{3.3}$$

where the scaling factor $\tau$ is an additional parameter known as *temperature.* The overall bidirectional loss function is given by the average between the two unidirectional ones:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}\mathcal{L}_{\text{CLIP}}^{V\to T} + \frac{1}{2}\mathcal{L}_{\text{CLIP}}^{T\to V} \tag{3.4}$$

Note that the above loss function does not directly affect the similarity between samples of the same modality, but acts only on cross-modality pairs.

In practice, the loss is typically applied using minibatches, replacing $N$ with $N_b$, where $N_b$ represents the size of the batch. Since this is a contrastive objective, it generally necessitates a relatively large batch size. For instance, some studies utilize a minibatch size as large as 32,768 [44].

**Temperature.** The effects of the temperature parameter $\tau$ have been extensively studied in both unimodal and multimodal contrastive learning [54] [65] [61]. Geometrically, this parameter acts as a scaling factor that alters the radius of the hypersphere. To avoid introducing an additional hyperparameter, Radford et al. [44] treat it as a learnable parameter represented by $s$. They incorporate $s$ into the model's set of learnable parameters and scale the logits using the parametrization $\tau = \frac{1}{\exp(s)}$. However, later works [61] found this approach to be counterproductive and have proposed alternatives, such as using a fixed value, implementing a schedule, or adopting a different parametrization for $\tau$.

To better understand the action of $\mathcal{L}_{\text{CLIP}}$, we can refer to the algebraic formulation. Let's define the matrix $H_t \in \mathbb{R}^{N \times d}$ obtained by concatenating the textual embeddings $h_i^t$ along the first axis. Here $N$ denotes the batch size, and $d$ denotes the embedding space dimension. Similarly, we define $H_v \in \mathbb{R}^{N \times d}$ as the matrix created by applying the same operation to the visual embeddings. We then multiply the two embedding matrices to obtain the similarity matrix $S \in \mathbb{R}^{N \times N}$:

$$S := H_v H_t^T \tag{3.5}$$

For construction, this matrix is such that each element $(i, k)$ is the logit quantifying the probability of the image-text pair $(i, k)$ being a positive pair:

$$(S)_{(ik)} = \langle h_i^v, h_k^t \rangle \tag{3.6}$$

A visual representation of this operation is depicted in Figure 3.1. The learning objective can be interpreted as maximizing the diagonal elements against the others. Each row $i$ of the similarity matrix corresponds to a single instance of the cross-entropy in $\mathcal{L}_{\text{CLIP}}^{V \to T}$ for a fixed value of $i$. Similarly, $\mathcal{L}_{\text{CLIP}}^{T \to V}$ acts column-wise, maximizing the ratio (of exponentials) of the diagonal cell over the non-diagonal ones. In practice, this formulation allows for an efficient imwplementation in PyTorch: after applying the scaling factor $\tau$, the matrix can be used as logits in the cross-entropy loss function.

### 3.1.3 Properties of the embedding space

In this subsection, we define some metrics to evaluate the embedding space geometry.

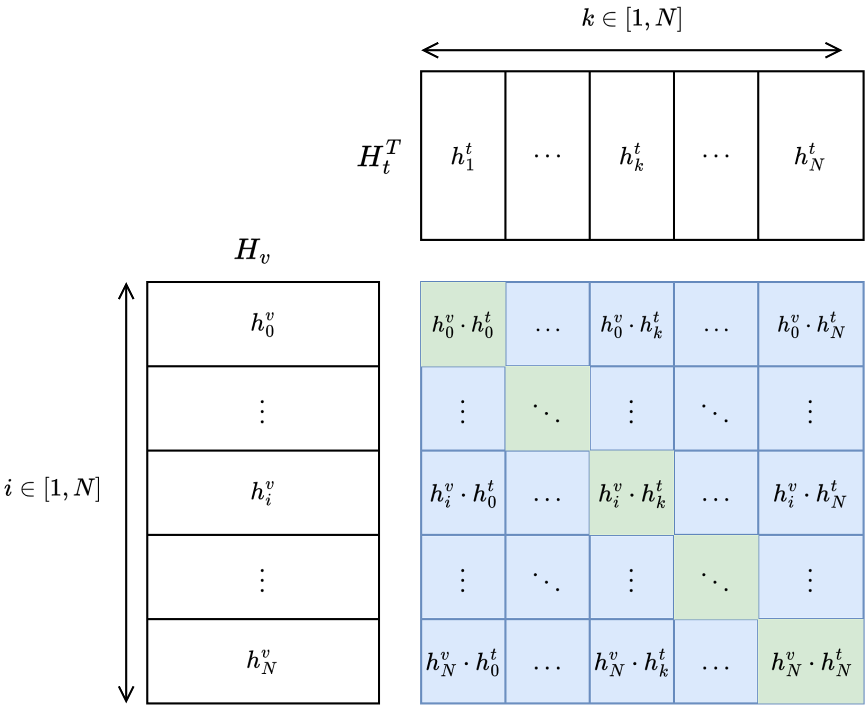


**Figure 3.1:** Visual representation of the CLIP loss: the $H_v$ visual (left) and $H_t$ textual embeddings (top) are multiplied to obtain the similarity matrix (bottom-right). The learning objective can then be interpreted as maximizing the diagonal elements (green) against the others (blue). Each row $i$ of the similarity matrix corresponds to a single instance of the cross-entropy in $\mathcal{L}_{\text{CLIP}}^{V \to T}$ for a fixed value of $i$. Similarly, $\mathcal{L}_{\text{CLIP}}^{T \to V}$ acts column-wise, maximizing the ratio (of exponentials) of the green cell over the blue ones.

### Mean similarity

A simple way to measure the effectiveness of the contrastive objective (Section 3.1.2) is by means of the similarity distribution. To describe this distribution with simple metrics, we refer to the mean similarity between all positive pairs (MPS) and all negative pairs (MNS). Note that in a dataset with size $N$, there are $N$ positive pairs and $N(N-1)$ negative ones. Formally:

$$\text{MPS} := \frac{1}{N} \sum_{i=1}^{N} h_i^t h_i^v \tag{3.7}$$

$$\text{MNS} := \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} h_i^t h_j^v \tag{3.8}$$

Referring to the example in Figure 3.1, MPS is the average value of the green squares, while MNS corresponds to the average value of the blue ones.

### Modality gap

We denote the *modality gap* (see Section 2.2) as $\Delta_{\text{gap}} \in \mathbb{R}^d$, and define it as the centroid difference between the two modalities:

$$\begin{aligned} \Delta_{\text{gap}} :=& \frac{1}{N} \sum_{i=1}^{N} h_i^v - \frac{1}{N} \sum_{i=1}^{N} h_i^t && (3.9) \\ =& \frac{1}{N} \sum_{i=1}^{N} (h_i^v - h_i^t) && (3.10) \end{aligned}$$

Analogously, we define the *modality delta vector* $\Delta_i \in \mathbb{R}^d$ for each pair $(h_i^t, h_i^v)$:

$$\Delta_i := h_i^v - h_i^t \qquad i \in [1, N] \tag{3.11}$$

Using equations (3.9) and (3.11), we can reformulate the modality gap as the mean of the modality delta vectors:

$$\Delta_{\text{gap}} = \frac{1}{N} \sum_{i=1}^{N} \Delta_i \tag{3.12}$$

To assess modality alignment, we often refer to the Euclidean norm $\|\Delta_{\text{gap}}\|$, which represents the centroid distance between visual and text modalities. Unless otherwise specified, when we talk about modality gap, we refer to $\|\Delta_{\text{gap}}\|$.

Additionally, we can also define the *alignment* as the mean squared L2-norm of the modality delta vectors:

$$\mathcal{A} := \frac{1}{N} \sum_{i=1}^{N} \|\Delta_i\|^2 \tag{3.13}$$

We say that the alignment improves as $\mathcal{A}$ goes to 0.

### Variance

From a statistical point of view, the embedding space can also be described with a distribution of the representations over the hypersphere. To have a good semantic expressivity, we seek the embedding space to be as "wide" as possible. This way, we can ensure the model fully exploits the available dimensionality. We measure this "wideness" by means of the *in-modality variance*. We first define the (textual) mean $\mu_t$ and covariance matrix $\Sigma_t$:

$$\mu_t := \frac{1}{N} \sum_{i=1}^{N} h_i^t \tag{3.14}$$

$$\Sigma_t := \frac{1}{N} \sum_{i=1}^{N} (h_i^t - \mu_t)(h_i^t - \mu_t)^T \tag{3.15}$$

The total textual *in-modality variance* is then formulated as:

$$\mathbb{V}ar \left[h^t\right] := Tr(\Sigma_t) = \frac{1}{N} \sum_{i=1}^{N} \|h_i^t - \mu_t\|^2 \tag{3.16}$$

And analogously for the visual modality:

$$\mathbb{V}ar \left[h^v\right] := Tr(\Sigma_v) = \frac{1}{N} \sum_{i=1}^{N} \|h_i^v - \mu_v\|^2 \tag{3.17}$$

For a good semantic expressivity, we seek the in-modality variances to be as large as possible.

**Uniformity**

Another important property for the embedding space is *uniformity*. To assess this property, we use a metric based on the Wasserstein distance between the embedding distribution and an ideal Gaussian distribution $\mathcal{N}(0, \frac{1}{N}\mathbf{I}_N)$ [19]. The formulation is:

$$\mathcal{U} := \sqrt{\|\mu\|^2 + 1 + Tr(\Sigma) - \frac{2}{\sqrt{N}} Tr(\Sigma^{\frac{1}{2}})} \tag{3.18}$$

Where $\Sigma$ is the covariance matrix and $\mu$ is the mean of the embeddings. A small $\mathcal{U}$ indicates a large uniformity of representations.

### 3.1.4 Auxiliary losses

In order to bridge the modality gap and improve cross-modal alignment, Fahim et al. [18] propose some additional losses. In the following paragraphs, we report the formulations for these losses.

### In-modal uniformity

The in-modal uniformity works by pushing apart the samples within each modality. In the classical contrastive loss (3.4), there is no term that actively pushes apart samples from the same modality; thus a specific loss is designed. Formally, we can define the uniformity loss for the visual modality $\mathcal{L}_U^v$ as follows:

$$\mathcal{L}_U^v = \log\left(\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\exp(-2\|h_j^v - h_k^v\|^2)\right) \tag{3.19}$$

Analogously, we can define the in-modal uniformity for the textual modality $\mathcal{L}_U^t$, and compute the average between the two to obtain the total loss:

$$\mathcal{L}_U = \frac{1}{2}\mathcal{L}_U^v + \frac{1}{2}\mathcal{L}_U^t \tag{3.20}$$

Note that this loss pushes apart embeddings from the same modality regardless of whether they are semantically related.

### Cross-modal uniformity

While $\mathcal{L}_U$ acts by pushing apart encodings within each modality, it does not affect cross-modal negative pairs. To enforce a stronger constraint, a cross-modal version has been proposed. The so-called cross-modal uniformity loss $\mathcal{L}_{\text{XU}}$ is defined as follows:

$$\mathcal{L}_{\text{XU}} = \log\left(\frac{1}{N}\sum_{j=1}^{N}\sum_{\substack{k=1\\k\neq j}}^{N}\exp(-2\|h_j^v - h_k^t\|^2)\right) \tag{3.21}$$

### Alignment

Finally, to explicitly improve alignment, they propose the alignment loss $\mathcal{L}_{\text{A}}$:

$$\mathcal{L}_{\text{A}} = \frac{1}{N}\sum_{j=1}^{N}(\|h_j^v - h_j^t\|^2) \tag{3.22}$$

Different from the contrastive objective in Equation (3.4), which maximizes the positive similarity against the negative one, this loss aims to reduce the Euclidean distance between the embeddings of positive pairs.

In practice, these losses are designed to be used in combination with each other and the classical contrastive objective. In particular, we will refer to two specific combinations, proposed by [18]:

$$\mathcal{L}_{\text{CUA}} = \mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{U}} + \mathcal{L}_{\text{A}} \tag{3.23}$$

$$\mathcal{L}_{\text{CUAXU}} = \mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{U}} + \mathcal{L}_{\text{A}} + \mathcal{L}_{\text{XU}} \tag{3.24}$$

## 3.2 Cross-modal semantic consistency

Cross-modal semantic consistency is a multimodal generalization of the famous example « *king-man+woman* ≈ *queen* » [37]. The general principle is to consider the visual-text embedding space as a single unified semantic space, where the encodings are semantically consistent with each other regardless of the modality.

### 3.2.1 Formalization

Let us consider an image-text pair $(h_i^t, h_i^v)$. Ideally, both the textual and visual encodings should represent the same general concept. We denote this general concept (for pair $i$) as $\mathcal{C}_i$. However, since $h_i^t$ and $h_i^v$ belong to different modalities, we could expect their encodings to reflect this difference. In general, this *modality contribution* is modality-dependent and specific to each pair. Without loss of generality, we model the modality contributions as additional components orthogonal to $\mathcal{C}_i$ that we denote as $\psi_i^t$ and $\psi_i^v$ for text and vision, respectively. We can then write the embeddings as:

$$h_i^t = \mathcal{C}_i + \psi_i^t \qquad \psi_i^t \perp \mathcal{C}_i \tag{3.25}$$

$$h_i^v = \mathcal{C}_i + \psi_i^v \qquad \psi_i^v \perp \mathcal{C}_i \tag{3.26}$$

By subtracting the above expressions, we obtain the modality delta vector $\Delta_i$:

$$\Delta_i = h_i^v - h_i^t = \psi_i^v - \psi_i^t \qquad i \in [1, N] \tag{3.27}$$

In practice, to be able to effectively compute semantic arithmetic-based operations, we show that the modality delta vectors should be as aligned as possible. To understand why this is the case, let us examine two different pairs $(h_i^t, h_i^v)$ and $(h_j^t, h_j^v)$. In a unimodal setting, we can transition from sample $i$ to sample $j$ using the semantic delta vector $\mathcal{C}_j - \mathcal{C}_i$ through vector arithmetic [37]. In a multimodal setting, we reformulate this task as the problem of retrieving $h_j^v$ given $h_i^v$, $h_i^t$, and $h_j^t$. This is analogous to the multimodal arithmetic task proposed by [12] (see Section 2.3). We aim to tackle this problem by constructing a query vector $q_{i\to j}^v$ that we define as follows:

$$q_{i\to j}^v := h_i^v + (h_j^t - h_i^t) \qquad (i,j) \in [1, N]^2 \tag{3.28}$$

Theorem 3.2.1 gives us the condition under which we can successfully use the vector $q^v_{i\to j}$ to retrieve $h^v_j$. Note that this approach can be generalized to obtain a textual embedding $h^t_j$ with a query $q^t_{i\to j}$ analogously constructed.

**Theorem 3.2.1.** *We show that, for any $(i, j) \in [1, N]^2$, we have $q^v_{i\to j} = h^v_j$ iff $\Delta_i = \Delta_j$.*

*Proof.* Using the definitions (3.28) and (3.11), we have:

$$\begin{aligned} q^v_{i\to j} =& h^v_i + (h^t_j - h^t_i) \\ =& (h^v_i - h^t_i) + h^t_j \\ =& \Delta_i + h^t_j \end{aligned} \tag{3.29}$$

The difference $q^v_{i\to j} - h^v_j$ is then:

$$\begin{aligned} q^v_{i\to j} - h^v_j =& \Delta_i + h^t_j - h^v_j \\ =& \Delta_i - \Delta_j \end{aligned} \tag{3.30}$$

Thus $q^v_{i\to j} = h^v_j$ iff $\Delta_i = \Delta_j$. □

This theorem states that the most relevant factor for multimodal arithmetic is the alignment of the modality delta vectors. It is important to note that this does not suggest that the modality gap influences multimodal arithmetic performance. Indeed, this result indicates that instead of focusing on $\Delta_{\text{gap}}$ (which represents the mean of the vectors $\Delta_i$), we should consider their variance $\mathbb{V}ar\left[\Delta\right]$ (or a proxy metric, as described in the next section). Empirical results supporting this claim are discussed in Section 4.3.3.

It is also important to note that one way to improve the alignment of modality delta vector is by reducing their magnitude (i.e., setting $\Delta_i \to \vec{0}$); this leads to a decrease in both $\mathbb{V}ar\left[\Delta\right]$ and $\|\Delta_{\text{gap}}\|$. As a result, one might mistakenly interpret the closure of the modality gap as the factor that improves performance, while in reality, it is the alignment of the modality delta vectors that drives the improvement.

In practice, we can relax the equality in Theorem 3.2.1 to obtain the underlying general principle of *cross-modal semantic consistency:*

$$h^v_i + (h^t_j - h^t_i) \approx h^v_j \qquad \forall (i, j) \in [1, N]^2 \tag{3.31}$$

Or alternatively:

$$h^t_j - h^t_i \approx h^v_j - h^v_i \qquad \forall (i, j) \in [1, N]^2 \tag{3.32}$$

In other words, this property is such that making a *step* in one modality is equivalent to making a *step* in the other.

### 3.2.2 Designing a metric

To systematically measure cross-modal semantic consistency, we design a residuals-based metric named XSC-SR (X-modal Semantic Consistency based on mean Squared Residuals). Using the general principle in (3.32), we first define the semantic consistency residual for the $(i, j)$ pair:

$$\begin{aligned} r_{ij} &:= (h^t_j - h^t_i) - (h^v_j - h^v_i) && i, j \in [1, N]^2 && (3.33)\\ &= h^t_j - h^t_i - h^v_j + h^v_i && && (3.34)\\ &= (h^v_i - h^t_i) - (h^v_j - h^t_j) && && (3.35)\\ &= \Delta_i - \Delta_j && && (3.36) \end{aligned}$$

We can notice that:

$$\begin{aligned} & r_{ij} = -r_{ji} && \forall (i, j) \in [1, N]^2 && (3.37)\\ & \|r_{ij}\| = \|r_{ji}\| && \forall (i, j) \in [1, N]^2 && (3.38)\\ & r_{ii} = 0 && \forall i \in [1, N] && (3.39) \end{aligned}$$

To satisfy the general principle, we want $r_{ij}$ to be as close as possible to $\vec{0}$. Thus, we consider its squared L2-norm $\|r_{ij}\|^2$, and define XSC-SR as the mean value across all distinct pairs. Note that, given a dataset of $N$ samples, we can construct $\frac{N(N-1)}{2}$ distinct pairs, therefore:

$$\text{XSC-SR} := \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \|r_{ij}\|^2 \geq 0 \tag{3.40}$$

Given the symmetry property in (3.38), we can reformulate the above, considering also non-distinct pairs:

$$\text{XSC-SR} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \|r_{ij}\|^2 \tag{3.41}$$

Also, given the (3.39), we can include in the sum the improper pairs for which $i = j$:

$$\text{XSC-SR} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \|r_{ij}\|^2 \tag{3.42}$$

And explicitly:

$$\text{XSC-SR} := \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \|h^t_j - h^t_i - h^v_j + h^v_i\|^2 \geq 0 \tag{3.43}$$

By applying a well-known property of the variance (A.1), we can also rewrite it as follows:

$$\text{XSC-SR} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \|\Delta_i - \Delta_j\|^2 \tag{3.44}$$

$$= \frac{2N}{N-1} \mathbb{V}ar\left[\Delta\right] \tag{3.45}$$

Equation (3.43) gives us the definition of XSC-SR, while the (3.45) shows the relationship between this metric and the sample variance of the modality delta vectors, as well as a more practical way to compute it. Ideally, for a good embedding space, we would like the XSC-SR metric to be as small as possible.

Note that, for the construction of this metric, we have that Theorem 3.2.1 is satisfied iff XSC-SR $= 0$, i.e., in each of its global minima.

### 3.2.3 Relationship with other properties

Given Theorem 3.2.1, one might consider using XSC-SR as a loss function to explicitly optimize a VLM for a multimodal arithmetic task. However, we show with the following result why this approach might not be advisable.

**Theorem 3.2.2.** *The following equality holds*

$$\textit{XSC-SR} = \frac{2N}{N-1} \mathbb{V}ar\left[h^v\right] + \frac{2N}{N-1} \mathbb{V}ar\left[h^t\right] + 4(\textit{MNS} - \textit{MPS}) \tag{3.46}$$

*Proof.* We can prove the theorem by decomposing the XSC-SR metric and applying a well-known property of the variance. The complete proof can be found in Appendix A.2. □

Theorem 3.2.2 gives us the relationship between the XSC-SR metric and other properties of the embedding space: in-modality variances, mean positive similarity (MPS), and mean negative similarity (MNS). Recall that XSC-SR follows the principle "the smaller the better", thus Theorem 3.2.2 prescribes to maximize the difference MPS $-$ MNS, while minimizing in-modality variances. This highlights an important trade-off intrinsic to multimodal arithmetic; as a matter of fact, while having a good separation between positive and negative pairs is coherent with the goals of contrastive learning, in-modality variance is generally considered a good property to have, since it allows for a good semantic expressivity of the embedding space. Indeed, we notice that (3.46) has a global minimum corresponding to $\mathbb{V}ar\left[h^t\right] = \mathbb{V}ar\left[h^v\right] = 0$ (see Figure 3.2 left); in such a case, it is easy to see how Theorem 3.2.1 is satisfied for all the samples, even though such embeddings do not carry any meaningful semantic information. From this, we can identify two main takeaways:

1. Multimodal arithmetic, as formulated by [12], is intrinsically limited by the geometry of the embedding space.

2. XSC-SR, while useful to assess cross-modal semantic consistency, should not be used as a standalone learning objective, as it may deteriorate the embedding distribution.

### 3.2.4 Global minima

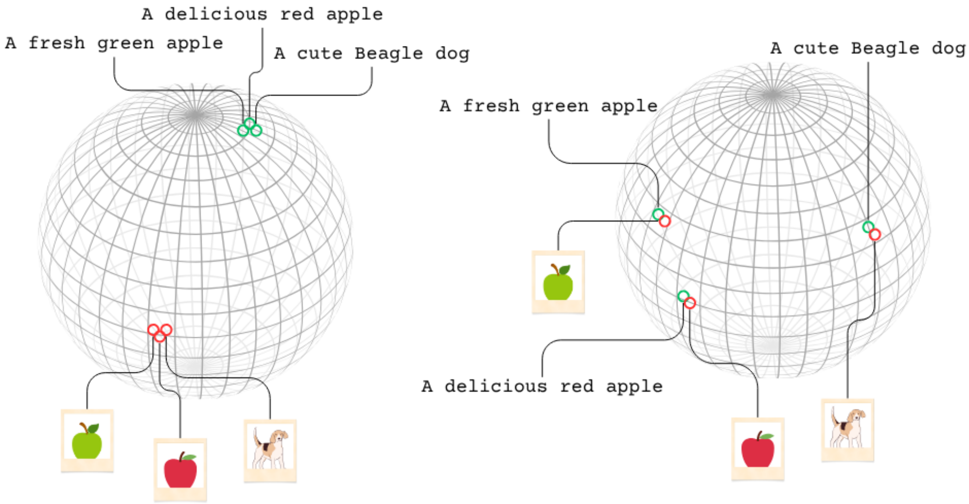


**Figure 3.2:** 3D illustration of the XSC-SR global minima. The zero-variance case (left) is such that each modality is encoded into a single point; in such a case, both the modality gap and the alignment might be large. The perfect alignment case (right) is such that each text-image pair is encoded into a single point; in such a case, we have perfect alignment and zero modality gap.

We have already mentioned how squeezing the in-modality variances gives us one global minimum for XSC-SR. We now prove that proposition.

Having $\mathbb{V}ar\left[h^t\right] = \mathbb{V}ar\left[h^v\right] = 0$ means that all embeddings of each modality are encoded in two single points of the hypersphere (see Figure 3.2 left). Let us denote these points as $h^t_\star$ and $h^v_\star$ such that we have $h^t_i = h^t_\star$ and $h^v_i = h^v_\star$ $\forall i \in [1, N]$. We note that, in such a case:

$$\mathbb{V}ar\left[h^t\right] = \mathbb{V}ar\left[h^t_\star\right] = 0 \tag{3.47}$$

$$\mathbb{V}ar\left[h^v\right] = \mathbb{V}ar\left[h^v_\star\right] = 0 \tag{3.48}$$

$$MPS = \frac{1}{N}\sum_{i=1}^{N} h^v_\star \cdot h^t_\star = h^v_\star \cdot h^t_\star \tag{3.49}$$

$$MNS = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} h^v_\star \cdot h^t_\star = h^v_\star \cdot h^t_\star \tag{3.50}$$

Thus, equation (3.46) gives us XSC-SR $= 0$.

Furthermore, from the definition of the residuals in (3.36), we notice that another minimum is given when the positive pairs are perfectly aligned (i.e., $\Delta_i = \vec{0}$), indeed:

$$\Delta_i = \vec{0} \qquad \forall i \in [1, N] \tag{3.51}$$

$$\implies \Delta_i - \Delta_j = \vec{0} \qquad \forall (i,j) \in [1, N]^2 \tag{3.52}$$

$$\implies r_{ij} = \vec{0} \qquad \forall (i,j) \in [1, N]^2 \tag{3.53}$$

$$\implies \text{XSC-SR} = 0$$

From this latter result, we can deduce that another way to improve cross-modal semantic consistency is by improving the alignment $\mathcal{A}$ as it is defined in (3.13) (see Figure 3.2 right). We summarize these two global minima, along with the corresponding geometrical and statistical properties, in Table 3.1.

| **Description** | **Alignment** | **In-modality variance** | **Modality gap** |
|---|---|---|---|
| Zero-variance | [0,4] | 0 | [0, 2] |
| Perfect alignment | 0 | [0, 1] | 0 |

**Table 3.1:** Summary of the geometrical and statistical properties of the global minima of XSC-SR.

## 3.3 Contrastive query-target objective

We now discuss the problem of designing a loss function $\mathcal{L}_{\text{MA}}$ to explicitly train a model for multimodal arithmetic. As we have seen in Section 3.2.3, using XSC-SR as a loss function might lead the model to the zero-variance global minimum, and deteriorate the embeddings distribution. This global minimum directly follows from the strict condition $q^v_{i\to j} = h^v_j$ in Theorem 3.2.1. However, in a practical application, we do not really need to have a strict equality between our query vector and the target image. Indeed, if we reformulate the task of multimodal arithmetic as a classification problem, where we limit ourselves to match the query with the "best" target image, it becomes sufficient to ensure that the query is positioned closer to the target with respect to any other sample in the database.

A similar argument can be made for the classical contrastive objective discussed in Section 3.1.2: in that context, we did not require the loss to completely align the positive pairs, but we instead constructed it to maximize the similarity of positive pairs *against* the negative ones.

### 3.3.1 Formulation

Applying the above reasoning to the multimodal arithmetic task, we decide to follow a contrastive approach. Recall that our goal is to retrieve a target image $h_j^v$ given a query $q_{i \to j}^v$. From a set of $N$ text-image pairs, we can construct $N^2$ queries; each of these queries should be matched with one among $N$ candidate images. Since, in general, $q_{i \to j}^v$ do not have a unitary L2-norm, we normalize the queries. We then stack all the normalized queries along the first axis to construct a query matrix $Q_v \in \mathbb{R}^{N^2 \times d}$, and do the same operations for the candidate images to obtain a target matrix $T_v \in \mathbb{R}^{N \times d}$. Finally, we compute a similarity matrix $S_{q \to v} \in \mathbb{R}^{N^2 \times N}$ multiplying $Q_v$ and $T_v$:

$$S_{q \to v} = Q_v T_v^T \tag{3.54}$$

For simplicity, we index matrix $S_{q \to v}$ denoting the rows with $(i, j) \in \mathbb{R}^{N^2}$ (corresponding to the query $q_{i \to j}^v$) and columns with $k \in [1, N]$ (corresponding to the candidate image $h_k^v$). For construction, we have:

$$(S_{q \to v})_{[(i,j),k]} = \cos(q_{i \to j}^v, h_k^v) \tag{3.55}$$

Note that the positive elements are given when $j = k$. In practice, we can identify their position with an index matrix $\mathcal{I}$ constructed by vertically stacking $N$ identity matrices $\mathbf{I}_N$:

$$\mathcal{I} = \left[\begin{array}{cccc} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \hline \vdots & \vdots & \vdots & \vdots \\ \hline 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{array}\right] \in \mathbb{R}^{N^2 \times N}$$

The learning objective can then be formulated as the optimization task of maximizing the positive elements against the negative ones. A visual depiction of this objective is proposed in Figure 3.3.

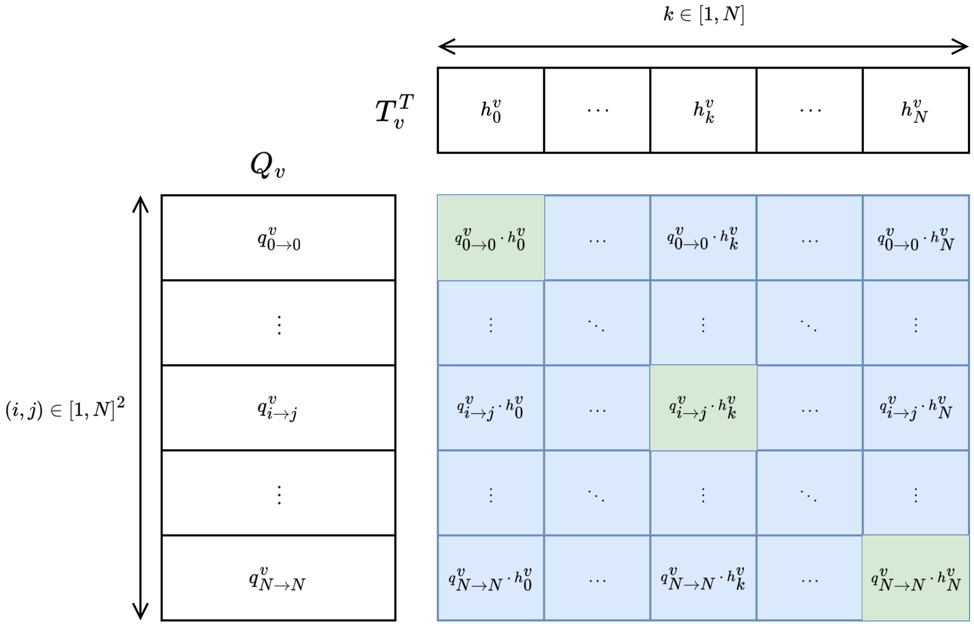

$k \in [1, N]$
$T_v^T$
$h_0^v$
$\cdots$
$h_k^v$
$\cdots$
$h_N^v$
$Q_v$
$(i, j) \in [1, N]^2$
$q_{0\to 0}^v$
$\vdots$
$q_{i\to j}^v$
$\vdots$
$q_{N\to N}^v$
$q_{0\to 0}^v \cdot h_0^v$
$q_{0\to 0}^v \cdot h_k^v$
$q_{0\to 0}^v \cdot h_N^v$
$\ddots$
$q_{i\to j}^v \cdot h_0^v$
$q_{i\to j}^v \cdot h_k^v$
$q_{i\to j}^v \cdot h_N^v$
$q_{N\to N}^v \cdot h_0^v$
$q_{N\to N}^v \cdot h_k^v$
$q_{N\to N}^v \cdot h_N^v$

**Figure 3.3:** Visual depiction of the query-target contrastive objective. The query matrix $Q_v$ (left) is multiplied with the target matrix $T_v$ (top) to obtain the similarity matrix $S_{q\to v}$ (bottom-right). The positive elements ($j = k$) are depicted in green, while the negative ones ($j \neq k$) are colored in blue. Due to spatial limitations, we can only show a small subset of the similarity matrix rows. In reality, this matrix is rectangular, with $N^2$ rows and $N$ columns.

Note that until now, we have only considered the query-to-image case, but the same approach can be followed to design a query-to-text similarity matrix $S_{q\to t}$.

To enforce the optimization objective, we write an infoNCE loss function that, for the query-to-vision case, we can write as follows:

$$\mathcal{L}_{\mathrm{MA}}^{q\to v} = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\log\left[\frac{\exp(\cos(q_{i\to j}^{v}, h_j^v)/\tau)}{\sum_{k=1}^{N}\exp(\cos(q_{i\to j}^{v}, h_k^v)/\tau)}\right] \tag{3.56}$$

and analogously for the query-to-text case:

$$\mathcal{L}_{\mathrm{MA}}^{q\to t} = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\log\left[\frac{\exp(\cos(q_{i\to j}^{t}, h_j^t)/\tau)}{\sum_{k=1}^{N}\exp(\cos(q_{i\to j}^{t}, h_k^t)/\tau)}\right] \tag{3.57}$$

Finally, we define the *bidirectional* loss as:

$$\mathcal{L}_{\mathrm{MA}} = \frac{1}{2}\mathcal{L}_{\mathrm{MA}}^{q\to v} + \frac{1}{2}\mathcal{L}_{\mathrm{MA}}^{q\to t} \tag{3.58}$$

From now on, we will refer to $\mathcal{L}_{\mathrm{MA}}$ as the *unweighted bidirectional* loss, and to $\mathcal{L}_{\mathrm{MA}}^{q\to v}$ as its *monodirectional* version.

**Applications for CIR.** In the scope of CIR, $\mathcal{L}_{\mathrm{MA}}$ can be interpreted as a supervised contrastive loss similar to the one used in CLIP4CIR [4]. The difference is that we do not require an annotated relative caption to transform $h_i^v$ into $h_j^v$; instead, we approximate the relative caption with the difference vector $h_j^t - h_i^t$, and use this vector to construct the query. This approach is based on the assumption that any image $i$ can be meaningfully transformed into an image $j$. We will discuss a possible strategy to relax this assumption in Section 3.3.3.

### 3.3.2 Edge cases behaviour

We now present a special behavior that arises from $\mathcal{L}_{\mathrm{MA}}$ when we consider the particular case $i = j$. For simplicity, let us consider the monodirectional vision-to-text loss. Under the condition $i = j$, we have:

$$q_{i\to j}^v = h_i^v + (h_j^t - h_i^t) \tag{3.59}$$

$$q_{i\to i}^v = h_i^v \qquad (i = j) \tag{3.60}$$

We can also notice that in the loss written in (3.56) there appear $N$ different terms for which $i = j$. We can isolate from (3.56) the partial sum that depends on those terms, and write it as follows:

$$-\frac{1}{N^2}\sum_{i=1}^{N}\log\left[\frac{\exp(\cos((q^v_{i\to i}, h^v_i)/\tau)}{\sum_{k=1}^{N}\exp(\cos(q^v_{i\to i}, h^v_k)/\tau)}\right] \tag{3.61}$$

$$= -\frac{1}{N^2}\sum_{i=1}^{N}\log\left[\frac{\exp(\cos(h^v_i, h^v_i)/\tau)}{\sum_{k=1}^{N}\exp(\cos(h^v_i, h^v_k)/\tau)}\right] \tag{3.62}$$

$$= -\frac{1}{N^2}\sum_{i=1}^{N}\log\left[\frac{\exp(1/\tau)}{\sum_{k=1}^{N}\exp(\langle h^v_i, h^v_k\rangle/\tau)}\right] \tag{3.63}$$

$$= -\frac{1}{N^2}\sum_{i=1}^{N}\log\left[\frac{1}{\sum_{k=1}^{N}\exp(\langle h^v_i, h^v_k\rangle/\tau)}\right] - \frac{1}{\tau N} \tag{3.64}$$

We notice that, as written in (3.64), this partial sum acts to minimize the similarity between each pair of embeddings inside the target modality.

To summarize, we can conclude that inside each monodirectional loss there appear $N$ terms for which $i = j$, and that these terms behave like an in-modal uniformity loss that pushes apart the embeddings of the target modality.

### 3.3.3 Tackling incompatible samples

Until now, we have explored a loss function based on the principles of multimodal arithmetic. $\mathcal{L}_{\text{MA}}$ is based on the assumption that we can write a meaningful transformation query between any couple of pairs $i, j$. However, this is hardly the case in practice. Indeed, there are many couples of images that are too different to allow for a transformation that makes sense. Let's consider, for example, the images in Figure 3.4: for a human annotator, it would be hard to write a relative caption to transform one image into the other. Yet, the $\mathcal{L}_{\text{MA}}$ would consider the difference between their textual embeddings as a relative caption, even if it hardly reflects any meaningful visual transformation.

To tackle the problem of incompatible samples, we propose a weights-based approach. We aim to define a weighting function $w : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$, such that $w(h^t_i, h^t_j)$ expresses how "compatible" the captions $i$ and $j$ are. We then use this function to re-weight the cross-entropy loss. The unidirectional query-to-vision loss thus becomes:

$$\mathcal{L}^{q\to v}_{\text{MA}^\star} = -\frac{1}{\sum_{l=1}^{N}\sum_{m=1}^{N} w(h^t_l, h^t_m)}\sum_{i=1}^{N}\sum_{j=1}^{N} w(h^t_i, h^t_j)\log\left[\frac{\exp(\cos(q^v_{i\to j}, h^v_j)/\tau)}{\sum_{k=1}^{N}\exp(\cos(q^v_{i\to j}, h^v_k)/\tau)}\right] \tag{3.65}$$

and similarly:

$$\mathcal{L}^{q\to t}_{\text{MA}^\star} = -\frac{1}{\sum_{l=1}^{N}\sum_{m=1}^{N} w(h^t_l, h^t_m)}\sum_{i=1}^{N}\sum_{j=1}^{N} w(h^t_i, h^t_j)\log\left[\frac{\exp(\cos(q^t_{i\to j}, h^t_j)/\tau)}{\sum_{k=1}^{N}\exp(\cos(q^t_{i\to j}, h^t_k)/\tau)}\right] \tag{3.66}$$

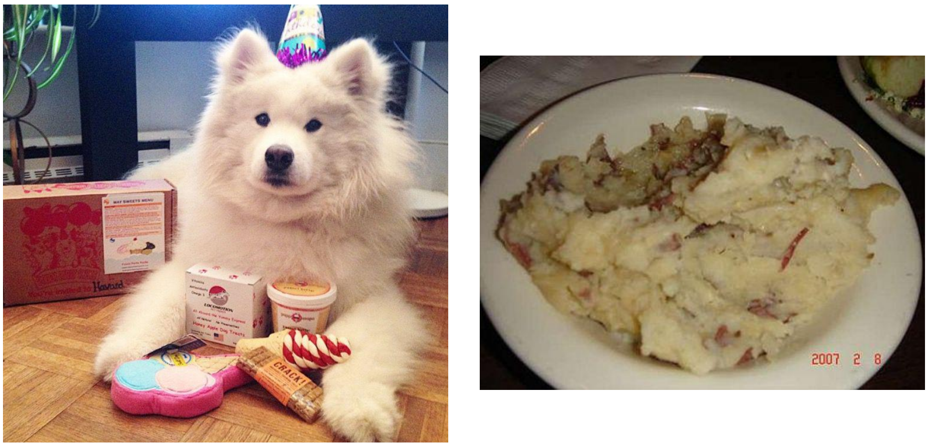


**Figure 3.4:** Example of two incompatible images. A human annotator would likely not be able to write a meaningful relative caption to transform a dog (left) into a serving of mashed potatoes (right). These images are taken from CIRR's imageset.

As before, the bidirectional version is the average of the two unidirectional ones:

$$\mathcal{L}_{\mathrm{MA}^\star} = \frac{1}{2}\mathcal{L}_{\mathrm{MA}^\star}^{q\to v} + \frac{1}{2}\mathcal{L}_{\mathrm{MA}^\star}^{q\to t} \tag{3.67}$$

For the rest of this work, we will use the weighted versions $\mathcal{L}_{\mathrm{MA}^\star}$ and $\mathcal{L}_{\mathrm{MA}^\star}^{q\to v}$ of our loss function, referring to them as the Multimodal Arithmetic loss (MA-loss bi) and its monodirectional variant (MA-loss mono).

### Weighting function

We now want to define the function $w(\cdot,\cdot)$. We want this function to:

1. be symmetrical, since if the couple $(i, j)$ is incompatible, so is couple $(j, i)$
2. have values in [0,1], so that we can use it to re-weight the samples

We decide to use an approach based on the similarity; the rationale is that if two samples are similar enough, they are also compatible. Thus, we formulate $w(\cdot,\cdot)$ as:

$$w(h_i, h_j) := \mathbb{1}_{\langle h_i, h_j\rangle > 0} \left\langle h_i, h_j\right\rangle^2 \tag{3.68}$$

where $\mathbb{1}$ denotes the indicator function. With this formulation, we find that $w(h_i, h_j)$ is symmetric, has values in [0,1], and equals 0 if the similarity $\langle h_i, h_j\rangle$ is negative. The latter property enables us to disregard samples that are too dissimilar to represent a meaningful transformation. A plot for this function is presented in Figure 3.5.

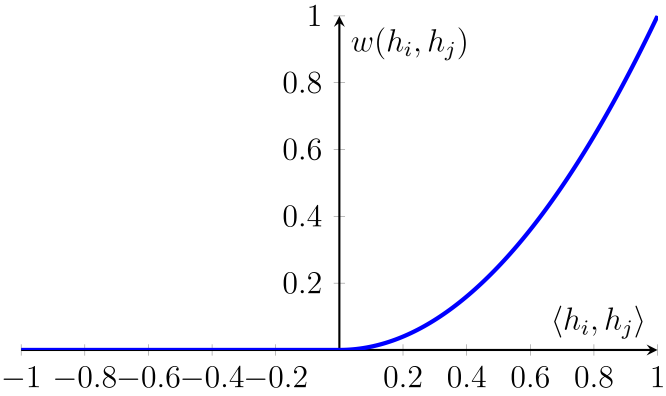


**Figure 3.5:** Behavior of the weighting function $w(h_i, h_j)$. The value only depends on the similarity $\langle h_i, h_j \rangle$.

In Appendix B.1, we report a simple PyTorch code to efficiently compute the weighted version of the MA-loss using this weighting function.

### Alternative weighting strategies

Until now, we have discussed a weighting strategy based on the textual embedding similarity, but we could as easily adopt a strategy based on the visual embeddings. All we have to do is compute the weighting function as $w(h_i^v, h_j^v)$ using the visual embeddings in place of the textual ones.

One could argue that using the same model to compute embeddings for both the query-target learning objective and the weighting function might confuse the model. Specifically, allowing a single model to optimize both the learning objective and the weighting function could result in the model assigning a low compatibility score to samples that are challenging but still "compatible." This situation could ultimately lead to a model with reduced generalization capabilities. To address this issue, we propose a weighting strategy called *frozen weights*. This strategy involves using a separate, frozen model instance solely for computing compatibility scores based on embedding similarities. Furthermore, since this auxiliary model does not partake in the training process, we can compute the compatibility scores once and cache the results to speed up training.

To summarize, defining a weighting strategy requires us to choose both the modality for calculating the weighting function and whether to use a frozen auxiliary model to compute the embeddings. This results in a total of four different strategies, which are outlined in Figure 3.6.

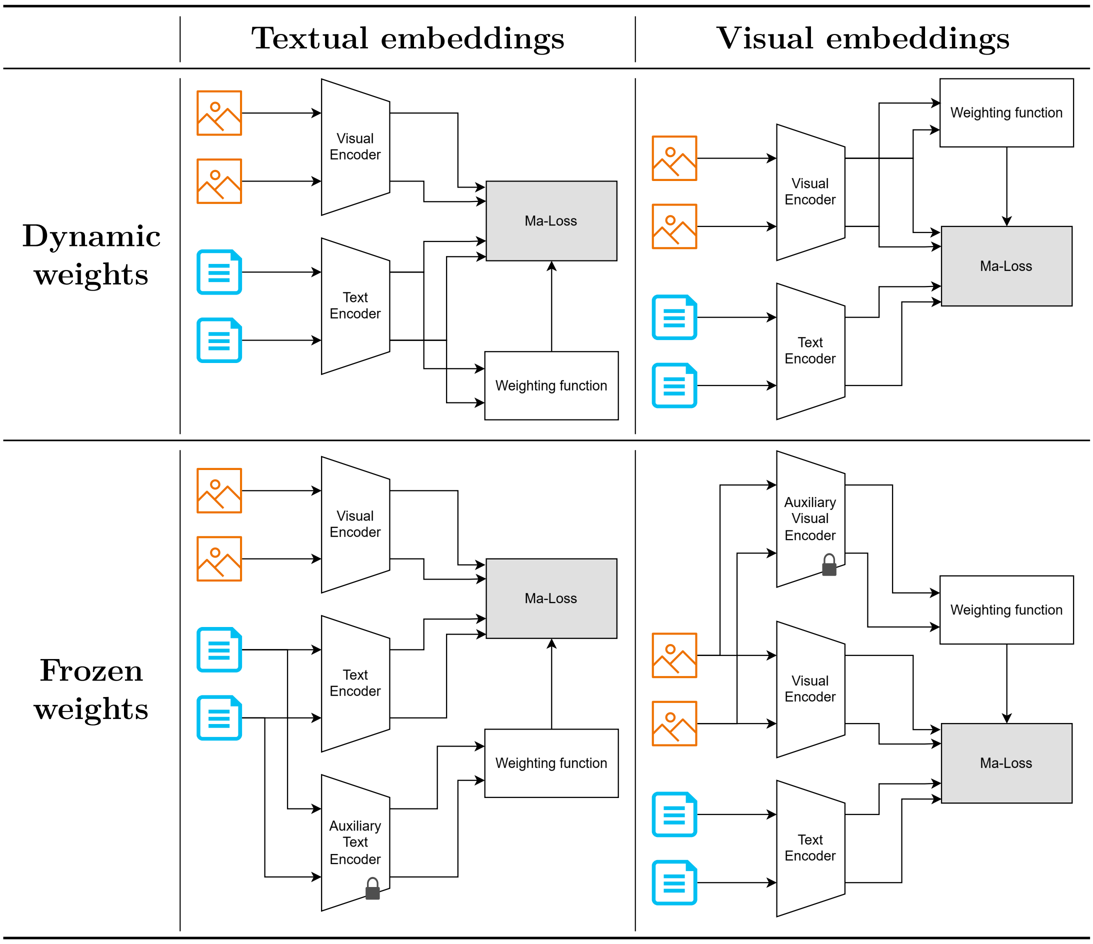


**Figure 3.6:** Visual representation of the weighting strategies. Each strategy is defined by the modality used to compute the weighting function: Textual (left) or Visual (right); and whether the weighting function is computed using the same model (top), or an auxiliary frozen model (bottom).

### 3.3.4 Adapting the MA-Loss to supervised CIR

In Section 3.3.1, we have discussed how MA-loss can be used to train a model for ZS-CIR. We did so based on two assumptions:

- Our linear fusion strategy (the query construction in (3.28)) is complex enough to apply a meaningful multimodal transformation
- When post-pre-training on an image-text pair dataset for ZS-CIR, we can approximate a text-based transformation using the difference of the captions' embeddings

In order to evaluate the first assumption, we aim to develop a method in order to evaluate the effectiveness of the fusion strategy, excluding any possible effects due to the zero-shot transfer. In other words, we want to establish a supervised upper bound for CIR.

Unlike MA-loss, which is formulated for an image-text dataset like MSCOCO [33], in supervised CIR, we have a relative caption $x_i^c$ that describes the difference between source $x_i^v$ and target $z_i^v$ images. We then proceed by encoding the images with the visual encoder to obtain the embeddings $h_i^x = g_\phi(x_i^v) \in \mathbb{R}^d$ and $h_i^z = g_\phi(z_i^v) \in \mathbb{R}^d$, for source and target, respectively. Similarly, we encode the relative caption with the textual encoder and obtain the encoding $h_i^c = f_\theta(x_i^c) \in \mathbb{R}^d$.

Recall that for the MA-loss, we approximated the relative caption as the difference between the text embeddings; in supervised CIR, we can directly use the relative caption embedding $h_i^c$. The CIR query to retrieve the target image can then be written as a simple sum between the source image and the relative caption embeddings:

$$q_i^{\text{CIR}} := h_i^x + h_i^c \tag{3.69}$$

We then use the above query inside the unidirectional unweighted loss in (3.56). We call the resulting loss function MA-CIR (Multimodal Arithmetic for CIR), and we formulate it as follows:

$$\mathcal{L}_{\text{MA-CIR}} = -\frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{\exp(\cos(q_i^{\text{CIR}}, h_j^z)/\tau)}{\sum_{k=1}^{N}\exp(\cos(q_i^{\text{CIR}}, h_k^z)/\tau)}\right] \tag{3.70}$$

Since CIR is fundamentally a modality-asymmetrical problem, i.e., we always retrieve an element from the visual modality, we do not need a bidirectional version of the MA-CIR loss.

This supervised CIR loss is equivalent to the fusion strategy in [4], without the combiner.

# Chapter 4

# Experiments

In this chapter we will discuss the experimental setup, datasets and results of our experiments. In Section 4.1, we start with a brief overview of our setup and main hyperparameters used across most of our experiments. In Section 4.2, we then describe the datasets we used, their structure, origin, and usage within the scope of this thesis. Finally, we present our experimental results. These are divided into two categories: first, in Section 4.3 we discuss the multimodal arithmetic performance of our methods, along with a comparison with some replicated methods from the SoTA, an ablation study of the various weighting strategies, an analysis of the embedding space geometry in relation with SIMAT, and a study of the SIMAT score dynamics during the post-pre-trainging process; secondly, in Section 4.4 we present a set of experiments centered on CIR, including a performance comparison across the weighting strategies, and a comparison with SoTA performance on both FashionIQ and CIRR.s

## 4.1 Experimental setup

All the experiments are run on a single NVIDIA GeForce RTX 2080 Ti GPU with 11GB of memory. The implementation is developed in Python, leveraging the PyTorch [41] deep learning framework for model training, evaluation, and experimentation.

The main architecture chosen for the experiments is CLIP, introduced by Radford et al. [44]. We use the ViT-B/32 version, with a vision transformer for the vision tower and a slightly improved transformer [43] for the text tower. Model's weights are initialized with the checkpoints released by OpenAI. A standard size of 512 is chosen for the embedding space. The practical implementation was adapted from OpenAI's official repository on GitHub [1]. Unless otherwise specified, the post-pre-training process was applied exclusively to the final projection layers of CLIP, while keeping the backbone encoders frozen. The optimization is carried out using AdamW [35], which combines the fast convergence of Adam with decoupled weight decay, ensuring both fast convergence and improved generalization during training. Unlike standard Adam [29], its treatment of weight decay as a separate regularization term avoids the tendency to overfit and makes it particularly effective for training large-scale VLMs. The main hyperparameters chosen for post-pre-training are summarized in Table 4.1.

| **Hyperparameter** | **Value** |
|---|---|
| Image encoder | Vit-B/32 |
| Embedding size | 512 |
| Batch size | 128 |
| Epochs | 20 |
| Learning rate | 1e-6 |
| Weight decay | 0.1 |
| Epsilon | 1e-8 |
| Betas | (0.9, 0.99) |
| Scheduler | Cosine |

**Table 4.1:** Hyperparameters used for post-pre-trainging CLIP. Unless otherwise specified, these hyperparameters are used to post-pre-train all our models.

## 4.2 Datasets

The datasets used in our experiments can be divided into three categories:

- **Pair-based**. Mainly used for post-pre-training. Includes **MSCOCO**, described in Section 4.2.1
- **Text-driven image transformation**. Used to test for multimodal arithmetic. Includes **SIMAT**, described in Section 4.2.2
- **Triplet-based**. Used both for supervised CIR fine-tuning and CIR/ZS-CIR testing. Includes **FashionIQ** and **CIRR**, described in Section 4.2.3 and Section 4.2.4, respectively.

The main characteristics and usage of these datasets are summarized in Table 4.2.

[1] https://github.com/openai/CLIP

| Dataset | N. Images | Usage | Structure |
|---|---|---|---|
| MSCOCO | 123287 | post-pre-train | Text-image pairs |
| SIMAT | 5860 | test | Text-driven image transformations |
| FashionIQ | 75254 | fine-tune/test | Image-text-image triplets |
| CIRR | 16939 | fine-tune/test | Image-text-image triplets |

**Table 4.2:** Summary of the datasets' characteristics. The number of images refers to the total number of images, considering all the splits. The usage is specifically referred to our experiments.

### 4.2.1 MSCOCO

MS COCO (Microsoft Common Objects in Context) is a widely used general-domain dataset for computer vision tasks, particularly suitable for object detection, segmentation, and image captioning. The dataset is known for its rich contextual information, containing complex everyday scenes with multiple objects per image, which provides a challenging environment for models to learn object recognition in a natural setting. It was first introduced in 2014 by Lin et al. [33], while a refined version was later released in 2017. This latter version is the most prominent and the one used in our training process. It is divided into a train/validation/test split featuring around 118,000 and 5000 images for training and validation, respectively. Each image is associated with five human-generated captions that describe the scene in natural language, making it ideal for an image-text contrastive objective. A practical example of the dataset image-text pairs structure is offered in Table 4.3.

### 4.2.2 SIMAT

The SIMAT dataset (Semantic IMage Transformation) was first introduced in 2022 by Couairon et al. [12] to evaluate models on the task of text-driven image transformation. It consists of approximately 6,000 images and 18,000 "transformation queries", where each query aims to either replace scene elements or modify their pairwise relationships in the image. The goal is to retrieve an image consistent with the given source image and transformation query.

The dataset is built on top of the Visual Genome dataset [30] using its annotations, which consist of subject-relation-object triplets. These triplets are then filtered and expanded to create feasible transformation queries. The transformation queries are, in turn, used to produce human-written transformed captions. The so-built samples are thus composed of a source image, a source caption, a transformation, and a transformed (target) caption. Some practical examples are

| Image | Captions |
|---|---|
| | 1. A table with pies being made and a person standing near a wall with pots and pans hanging on the wall.<br>2. A man is in a kitchen making pizzas.<br>3. Man in apron standing on front of oven with pans and bakeware<br>4. A baker is working in the kitchen rolling dough.<br>5. A person standing by a stove in a kitchen. |
| | 1. City dwellers walk by as a homeless man begs for cash.<br>2. A homeless man holding a cup and standing next to a shopping cart on a street<br>3. People walking past a homeless man begging on a city street<br>4. A person with a shopping cart on a city street<br>5. People are walking on the street by a homeless person. |

**Table 4.3:** Example structure of MSCOCO: each image is paired with five natural language descriptions.

reported in Table 4.4. For evaluation, an external image/text matching oracle (OSCAR[31]) is used to assess the probability of the retrieved image corresponding to the target caption. The final score is then given by a weighted sum of these oracle-computed probabilities. To ease the evaluation pipeline, the dataset also features pre-computed OSCAR scores for every combination of images and target captions. In this sense, unlike triplet-based datasets, SIMAT does not offer a hard ground truth in terms of compositional retrieval. Nevertheless, this dataset offers the ideal tool for quantifying the cross-modal semantic structure of the embedding space.

### 4.2.3 FashionIQ

The FashionIQ dataset is a comprehensive resource designed for interactive fashion image retrieval using natural language feedback. It was first introduced by Guo et al.[22] in 2019. The dataset includes about 77.6K images and 30.1K triplets,

| Source image | Transformation | Source caption<br>↓<br>Target caption |
|---|---|---|
| | Dog → Girl | A dog laying on a pillow.<br>↓<br>A girl laying on a pillow. |
| | Sitting on → Touching | A woman sitting on a horse.<br>↓<br>A woman touching a horse. |

**Table 4.4:** Example structure of SIMAT: each sample is composed of a source image, a source caption, a transformation, and a target caption. Unlike CIR-specific datasets, only a textual description of the target image is provided.

divided into training, validation, and test sets, with the test set being a challenge dataset that is not publicly accessible. FashionIQ covers three main categories of fashion items: dresses, shirts, and tops/tees. It provides two evaluation protocols: the VAL-Split and the Original-Split, which differ in the candidate image sets used for testing; in our experiments, we elected to use the VAL-Split. The dataset was constructed by collecting images and rich side information, including real-world product descriptions and detailed attribute labels extracted and refined from product websites. A key feature of FashionIQ is the inclusion of high-quality, human-written relative captions, which describe subtle visual differences between pairs of similar garment images to facilitate fine-grained retrieval. These relative captions were collected via crowd-sourcing, with annotators focusing on differences in color, texture, style, and other attributes of fashion.

FashionIQ supports several tasks, notably dialog-based interactive image retrieval, where a system iteratively refines retrieval results based on textual feedback provided by users comparing a reference image to the target image. It also supports single-turn retrieval and relative captioning tasks. Its structure is composed of typical CIR triplets, where each sample is composed of a query image, a target image, and two alternative relative captions describing the difference between the two images. An illustrative example of the dataset structure is reported in Table 4.5.

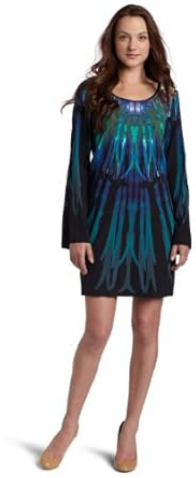
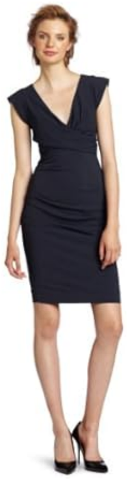
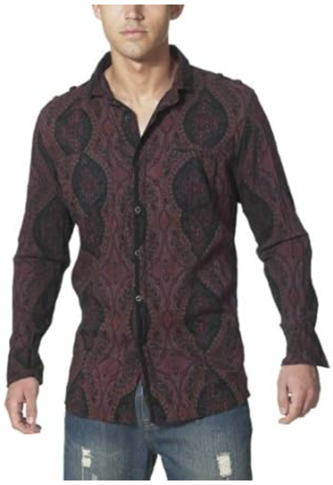
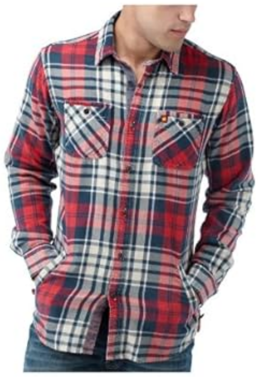
| Query image | Relative captions | Target image |
|---|---|---|
|  | 1. is solid black with no sleeves<br>2. is black with straps |  |
|  | 1. is plaid and is more colorful<br>2. is red and black plaid |  |

**Table 4.5:** Example structure of FashionIQ: each sample is composed of a query image, a target image, and two alternative relative captions that describe the difference between the two images.

### 4.2.4 CIRR

The CIRR dataset (Composed Image Retrieval on Real-life Images) is a large-scale benchmark designed for the task of composed image retrieval (CIR), where queries are composed of a reference image paired with a natural language text describing desired modifications to retrieve a target image. First proposed by Liu et al. [34], CIRR consists of over 36,000 annotated triplets (reference image, modification text, and target image) based on the images of NLVR2 [50], a dataset that provides a diverse collection of real-world images with rich context and semantic complexity. CIRR specifically emphasizes visually similar images to challenge models to perform fine-grained visiolinguistic reasoning, requiring understanding what in the image should be preserved or altered based on the text query. It includes carefully crafted subsets of semantically and visually related images, and the annotation process involves crowd-sourcing modification sentences that clearly discriminate the target image from other candidates within the subset, reducing false-negative issues common in prior datasets. A structured example is presented in Table 4.6.

The main purpose of CIRR is to push forward research on understanding and efficiently combining multimodal inputs (image and text) for retrieval in open-domain settings beyond narrow domains like fashion. It serves as a challenging benchmark to evaluate models on their ability to fuse visual and linguistic information to achieve accurate and subtle image retrieval based on text-guided modifications. CIRR additionally supports detailed analysis through metrics like RecallSubset, which considers ranking within subsets to evaluate fine-grained retrieval capability without bias from false negatives. Overall, CIRR is structured to enable and stimulate studies in joint image-text representation learning, multimodal reasoning, and interactive image retrieval in realistic environments.

## 4.3 Multimodal arithmetic

We start our set of experiments with a comparison focused on the Multimodal Arithmetic task. For this first experiment, we post-pre-train CLIP on MSCOCO using a variety of techniques. The models are evaluated on SIMAT using the test split. The results are presented in Table 4.7.

We compare the performance of our proposed method against two baselines. We also consider a number of techniques proposed to bridge the modality gap and/or improve the overall geometrical structure of the embedding space.

**W/O post-pre-training.** To establish a first baseline, we use the checkpoints released by OpenAI without any additional training process.

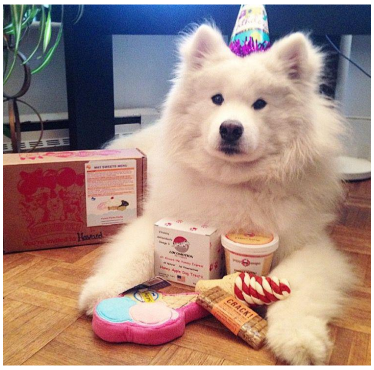
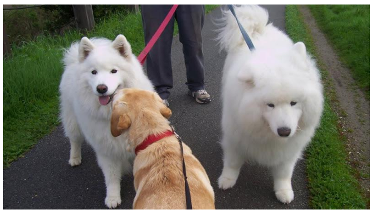
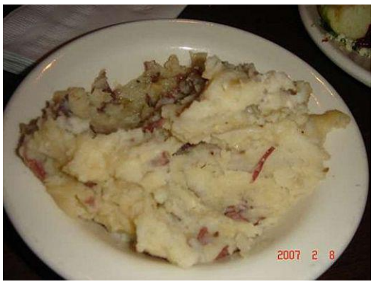
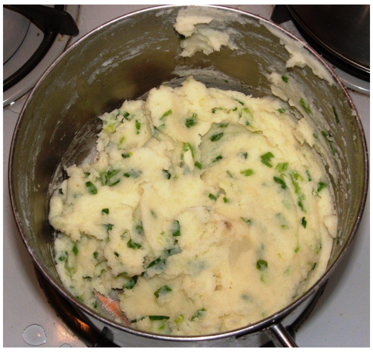
| Query image | Relative caption | Target image |
|---|---|---|
|  | Have two dogs of the same breed |  |
|  | Leave the mashed potatoes in the pot |  |

**Table 4.6:** Example structure of CIRR: each sample is composed of a query image, a target image, and a relative caption that describes the difference between the two.

**CLIP loss.** The second baseline is established using the standard CLIP loss, as described in [44]. The post-pre-training process is performed with the hyperparameters described in Table 4.1. As OpenAI's checkpoints were obtained using a different dataset, this baseline enables a fair comparison of the effect of our proposed methods against the traditional contrastive objective, excluding any impact due to potential domain differences.

**CUAXU loss.** Originally proposed by [18], this loss aims at improving both the pairwise alignment and the cross-modal uniformity (see Section 3.1.4). The proposing paper showed that this loss improves both modality gap and SIMAT performance. Note that our experimental setup is different from the one in [18], thus we obtain slightly different SIMAT scores.

**CUA loss.** Similar to CUAXU but without the cross-modal uniformity term (see Section 3.1.4). This loss was also proposed by [18] to bridge the modality gap.

**Hard Swapping Between Modalities (HS).** This method was proposed by [61] to mitigate the modality gap. It involves randomly selecting images and their corresponding text descriptions, and exchanging the projected embeddings of these images and paired texts within the joint embedding space. Following the original paper, we apply the swap with a probability of $1e-3$.

**Soft Swapping Between Modalities (SS).** Similar to HS, this method consists of merging the selected features using a weighted sum. In our implementation, we elected to use a weight of $\lambda = 0.5$ for the modality fusion. This corresponds to computing the arithmetic average between textual and visual features. Following [61], we apply the swap with a probability of $5e-2$.

**Fixed temperature (FT).** Proposed by [61], this strategy involves freezing the temperature parameter $\tau$. In this experiment, we report the results for $\tau = 0.1$.

**Bidirectional MA-loss.** We perform a post-pre-training process using our proposed (bidirectional) MA-loss (see Section 3.3.1). In Table 4.7 we report the results using a weighting strategy based on the textual embeddings similarities (see Section 3.3.3) without freezing the weights. In Table 4.8 we report the performance using different weighting strategies.

**Monodirectional MA-loss.** Similar to the above, we perform a post-pre-training process using our proposed MA-loss in a monodirectional configuration (see Section 3.3.1). The result reported in Table 4.7 is obtained using a weighting strategy based on textual embedding similarities and frozen weights. In Table 4.8 we report the performance using different weighting strategies.

| **Method** | **SIMAT ↑** | **Reported SIMAT ↑** |
|---|---|---|
| W/O post-pre-training | 16.30 | - |
| CLIP loss | 33.85 | - |
| CUAXU | 41.45 | 42.47 |
| CUA | 44.69 | 42.18 |
| SS | 43.80 | - |
| HF | 28.59 | - |
| FT | 42.54 | - |
| MA-loss (bi) | **48.13** | - |
| MA-loss (mono) | 43.84 | - |

**Table 4.7:** SIMAT score of our methods (bottom rows) compared to the baselines (top rows) and other SoTA methods (center rows). The reported SIMAT column refers to the score reported by the proposing paper [18].

The comparative results reported in Table 4.7 reveal a clear progression in SIMAT performance across the tested methods, with our proposed approaches achieving the strongest overall results. The baseline without post-pre-training yields the lowest score (16.30), as expected, since OpenAI's CLIP checkpoints were optimized on a more general dataset. This baseline thus establishes the lower bound for performance on this evaluation. The standard CLIP loss substantially improves results (33.85), confirming a considerable domain difference between MSCOCO and CLIP's proprietary pre-training dataset. Among the existing techniques designed to reduce the modality gap, both the CUAXU and CUA losses further enhance compositional structure, achieving 41.45 and 44.69, respectively. The improvement of CUA over CUAXU might indicate that excessive regularization on cross-modal uniformity may occasionally limit fine-grained optimization. Similarly, the swapping-based strategies (SS and HS) and the fixed-temperature configuration (FT) provide moderate gains, with SS (43.80) and FT (42.54) performing comparably to CUA, while HS (28.59) underperforms due to the instability introduced by hard embedding exchanges. In this sense, further hyperparameter tuning might yield some minor improvements; however, such efforts fall outside the scope of this work.

Our proposed multimodal arithmetic losses (MA-loss) deliver the highest SIMAT scores among all compared methods, confirming their effectiveness in reinforcing the geometric consistency of the joint embedding space. The bidirectional variant reaches 48.13, outperforming all baselines and previous SoTA techniques by a considerable margin of +3.4 points over the strongest competitor (CUA). This demonstrates that jointly considering both directions of the arithmetic relations contributes to more robust multimodal representations. The monodirectional configuration achieves 43.84, aligning closely with the best SoTA methods but still below the bidirectional approach. This contrast highlights the importance of symmetric cross-modal constraints in preserving semantic coherence across visual and textual domains. Overall, these results validate the hypothesis that the proposed MA-loss effectively improves multimodal compositional reasoning by enhancing the structure of the embedding space, ultimately leading to superior performance on SIMAT.

### 4.3.1 Similarity distributions

We propose an analysis of the similarity distributions between the constructed queries and the targets. For this experiment, we identify the SIMAT's targets as the images that maximize the oracle score. For comparison, we use images picked at random from SIMAT's image database, excluding those with an oracle's score greater than 0.5. In Figure 4.1, we plot the similarity distributions of the constructed queries with the targets and with random images.

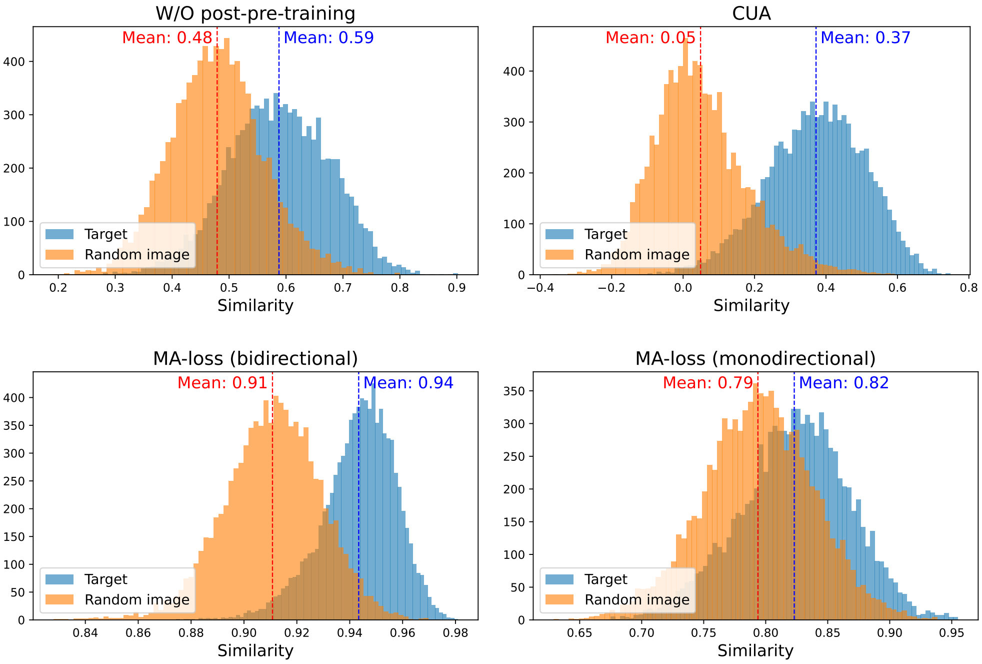


**Figure 4.1:** Similarity distribution considering targets (orange) and random images (blue) on SIMAT. Ideally, a good model should be able to separate the two distributions. The plots use different scales for better visibility.

This analysis gives some valuable information regarding the way different models behave on SIMAT. Ideally, a good model should be able to distinguish between a target and a random image; this would show as two separate curves, with a significant separation. The baseline without post-pre-training (Figure 4.1 top left) shows a significant overlap between the target and random image distributions; this overlap might produce confusion and reduce the model's ability to identify the correct target. In the same way, better-performing models like CUA and bidirectional MA-loss are able to better recognize the true targets, yielding a significant improvement on SIMAT and more separate similarity distributions.

An interesting result is obtained for the monodirectional MA-loss (Figure 4.1 bottom right); this model produces two very similar distributions with a difference in means of just 0.3 and a significant overlap, yet this configuration greatly improves the performance over the baseline, and performs only slightly worse than CUA. At this time, we cannot give a satisfactory explanation of why the monodirectional loss achieves good SIMAT results despite this behavior.

Another interesting observation is how different losses influence the means and standard deviations of the similarity distributions. The baseline model produces exclusively similarities on the positive range, spanning from around 0.2 to 0.8. CUA obtains similar results for the target distribution, but pushes the random distribution towards the left, producing queries that are more likely to be orthogonal to the non-target embeddings. The MA losses, instead, follow an alternative strategy: they increase the similarity of both the targets and the random images, but compensate by shrinking the standard deviations to achieve better separation. This suggests that CUA is able to produce queries that are more widely spread in the hyperspace, and potentially more informative. However, since the MA-loss outperforms CUA on SIMAT, it is not clear whether the "wideness" of the query space influences retrieval performance.

### 4.3.2 Weighting strategies ablation study

As discussed in Section 3.3.3, choosing a good weighting strategy to tackle incompatible samples is vital to improve the robustness of our loss. In this section, we study the effect that the weighting strategy has on SIMAT. In Table 4.8 we report the results considering the symmetrical bidirectional and asymmetrical monodirectional versions of MA-loss.

Starting with the overall trend, the bidirectional formulation clearly outperforms the monodirectional one across nearly all configurations. The best bidirectional score (48.13) surpasses the best monodirectional score (43.84) by more than four points, confirming that encouraging cross-modal consistency in both directions enhances the semantic alignment between textual and visual representations. This improvement suggests that the bidirectional constraint encourages a more globally coherent embedding space, in which both modalities contribute symmetrically to the underlying structure.

| **Loss direction** | **Weighting strategy** | **Frozen weights** | **SIMAT ↑** |
|---|---|---|---|
| Bi | None | - | 48.10 |
| Bi | Images | ✓ | 41.29 |
| Bi | Images | ✗ | 40.96 |
| Bi | Texts | ✓ | 48.00 |
| Bi | Texts | ✗ | **48.13** |
| Mono | None | - | 40.47 |
| Mono | Images | ✓ | 41.28 |
| Mono | Images | ✗ | 40.93 |
| Mono | Texts | ✓ | **43.84** |
| Mono | Texts | ✗ | 43.70 |

**Table 4.8:** SIMAT score comparison of MA-loss in different directional configurations and with different weighting strategies.

When considering the effect of weighting, the “None” configurations provide useful baselines for each loss direction. In the bidirectional case, the absence of weighting already yields a strong result (48.10), nearly matching the best-performing weighted variant (48.13). This indicates that the bidirectional loss itself captures much of the necessary relational information, and that weighting primarily serves a minor role rather than a decisive one. In contrast, the monodirectional loss without weighting performs substantially worse (40.47), suggesting that weighting plays a more crucial role when semantic relationships are enforced in only one direction. Here, the absence of weighting may leave the optimization biased toward one modality, reducing overall coherence in the joint embedding space.

Comparing the different weighting strategies reveals another consistent pattern: text-based weighting outperforms image-based weighting in both loss directions. For the bidirectional loss, text-based weighting (48.13 with unfrozen weights) slightly improves upon the unweighted baseline, while image-based weighting significantly underperforms (around 41 points). This discrepancy likely arises from the richer semantic structure of text embeddings, which provide more reliable cues for balancing the contribution of pairs. In the monodirectional case, a similar behavior emerges: text-based weighting reaches 43.84 (frozen) and 43.70 (unfrozen), while image-based variants hover near 41, confirming that text-derived weighting better guides the optimization in asymmetrical setups.

The impact of weight freezing is relatively minor but nonetheless informative. For image-based strategies, frozen weights consistently perform marginally better than unfrozen ones (41.29 vs. 40.96 for the bidirectional case, and 41.28 vs. 40.93 for the monodirectional), indicating that an auxiliary model during training offers a small but consistent gain. For text-based strategies, however, the difference between frozen and unfrozen configurations is negligible, suggesting that the more stable textual features do not benefit from independent reweighting.

In summary, these results collectively demonstrate that

- bidirectionality is the dominant factor driving performance gains
- text-based weighting is the most effective and stable strategy across both loss directions
- independent, frozen weights provide slight but consistent improvements
- in the absence of weighting, the bidirectional formulation still performs remarkably well, whereas the monodirectional loss requires weighting to approach competitive results

This interplay indicates that while weighting strategies can refine performance, the structural advantages of the bidirectional loss are fundamentally responsible for the superior multimodal arithmetic alignment observed.

### 4.3.3 Relationship with embedding space properties

Many previous works have tried to establish a connection between embedding space properties and downstream performance [32] [18] [61]. In Section 3.2.2, we expand those works proposing the XSC-SR metric to evaluate the semantic composability of our models. In this section, we evaluate the ability of XSC-SR to predict the downstream performance on the SIMAT benchmark.

Using the checkpoints reported in Table 4.7 and Table 4.8, we propose a simple correlation analysis between the SIMAT score and the main embedding space properties. In particular, we consider modality gap (using the centroid distance), alignment, uniformity, and XSC-SR. Note that modality gap, alignment, and XSC-SR follow the principle of "the smaller the better", while uniformity yields higher values for more uniform spaces. Figure 4.2 shows the scatter plots of these metrics against SIMAT.

None of the metrics considered achieves a definite strong correlation on multimodal arithmetic. The modality gap shows a behavior almost independent of SIMAT, with some techniques that are quite good at closing the modality gap performing similarly, or worse, than the ones with a pronounced gap. A similar observation can be made for the alignment, with checkpoints that produce embedding spaces with a wide range of alignment scores spanning between 0.20 and 1.80, but no significant relationship with SIMAT. Uniformity shows a significant negative correlation, suggesting that *less* uniform spaces yield better SIMAT results. XSC-SR, on the other hand, yields a moderate negative correlation with the downstream task..

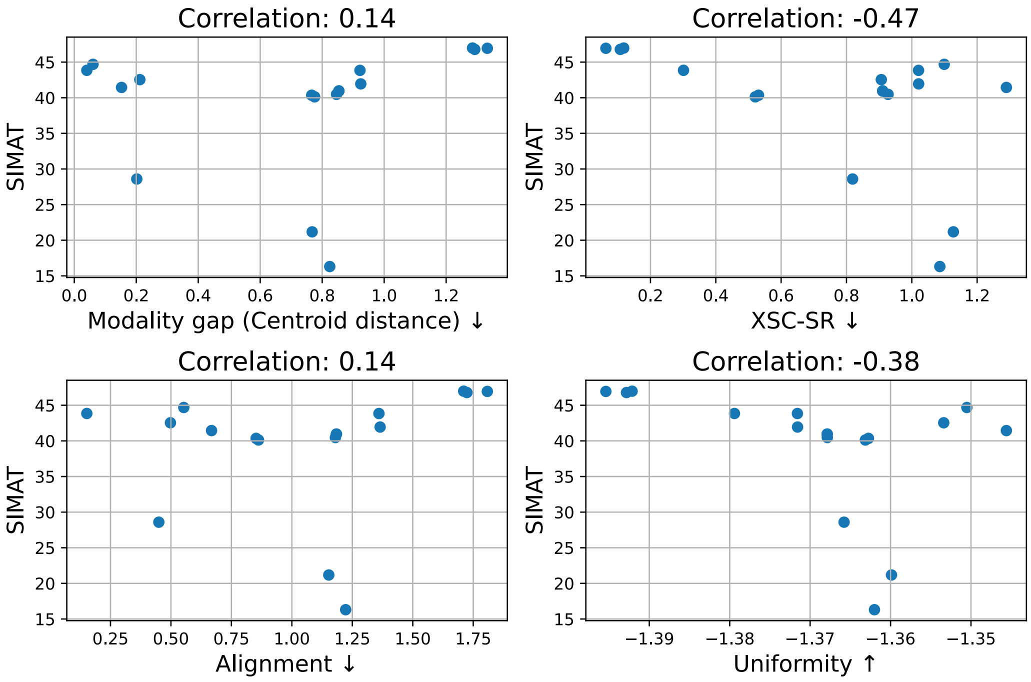


**Figure 4.2:** Correlations between SIMAT and other embedding space properties. Alignment (bottom left) and modality gap (top left) show almost no relationship with SIMAT. XSC-SR (top right) shows a mild but significant negative correlation with SIMAT.

These results support the theoretical insights in Section 3.2.2, demonstrating that improving alignment and reducing the modality gap can occasionally enhance semantic composability, though they are not essential to SIMAT. On the other hand, XSC-SR better correlates with multimodal arithmetic performance, even if it cannot explain the observed results alone. This is because a simple scalar metric like XSC-SR might not be complex enough to capture the nuanced structure of the embedding space, and thus is unable to provide definitive results.

### 4.3.4 SIMAT dynamics during post-pre-training

We finish this first set of experiments, presenting an interesting yet mysterious finding observed by analyzing the dynamics of the SIMAT score during post-pre-training with the MA-loss.

In Figure 4.3, we plot the SIMAT score as post-pre-training progresses. For this experiment, we evaluate the MA-loss in both unidirectional and bidirectional versions. We observe that the SIMAT score reaches a notable peak after only a

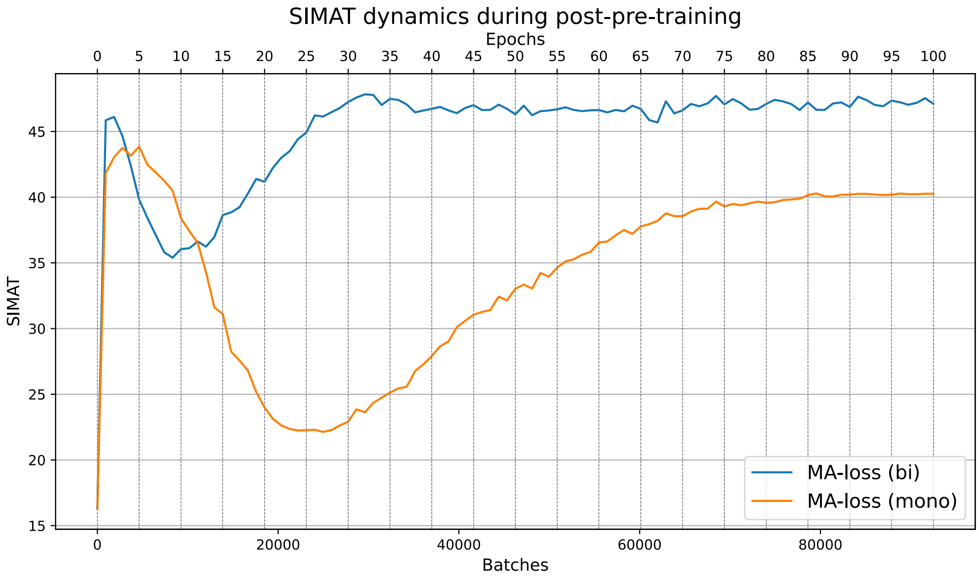


**Figure 4.3:** Epoch-wise SIMAT dynamics during post-pre-training. For the monodirectional loss, the drop is more pronounced, and the subsequent rise stabilizes at a lower value. The bidirectional version has a smaller drop and is then able to recover, achieving a second peak before stabilizing.

few epochs, increasing from around 16 at initialization to approximately 43 for the unidirectional loss and 46 for the bidirectional loss. Following this peak, there is a significant drop in downstream performance before it begins to rise again after some additional epochs.

Notably, the behavior of the unidirectional MA loss is more pronounced; it experiences a substantial drop in performance of about 20 points before stabilizing at a value of 40, which is lower than the initial peak. In contrast, the bidirectional loss has a smaller decline of 10 points before stabilizing at a new high of 47.

The fluctuations observed in the SIMAT scores during post-pre-training may be interpreted through the lens of the double descent phenomenon, which describes non-monotonic behaviors in model performance as the effective complexity of the system increases [39]. In our case, training length, or even introducing weighting strategies and changing loss directionality, can be seen as altering the implicit capacity of the model, which might explain why some configurations initially underperform and then recover or even surpass the initial peak. Although this perspective provides an appealing conceptual framework, double descent has not been systematically studied in zero-shot transfer settings, and the mechanisms governing transfer performance may differ from those typically examined in supervised learning [13]. Therefore, while double descent provides a plausible interpretation of the observed dynamics, further investigation is necessary before drawing any definitive conclusions.

On a more practical note, we can use this finding to refine our post-pre-training process. Since epoch-wise plots, like the one in Figure 4.3, are expensive and potentially difficult to replicate on datasets without a clear validation split, we can decide to implement an a priori early-stopping strategy. This approach would allow us to stop training during the first peak, but it may limit performance if a second peak occurs. However, since the presence of the second peak could be influenced by many unknown factors, an a priori early-stopping strategy would provide more consistent results across different datasets and configurations, other than a simpler and quicker post-pre-training process.

## 4.4 Composite image retrieval

We now generalize the approach we adopted for SIMAT in the ZS-CIR task. Similarly to what we did before, we post-pre-train a CLIP model on MSCOCO and then evaluate the checkpoints on two different CIR datasets, without any additional transfer procedure. Unlike what we did for SIMAT, for this set of experiments, we follow the approach of [58], training the models for just two epochs. Due to the reduced length of training, we do not employ any learning scheduler.

**Baselines.** We compare our methods against two basic baselines. The first baseline (**W/O post-pre-training**) simply evaluates the performance of OpenAI's pre-trained checkpoints used at initialization. This baseline aims at establishing a null hypothesis score to evaluate the effectiveness of our proposed methods. The second baseline (**CLIP loss**) is obtained by performing a post-pre-training process using the standard CLIP loss [44]. Since OpenaAi's checkpoints were trained on a general-purpose proprietary dataset, the second baseline is important to exclude any effect due to possible domain differences between the datasets.

**Supervised upper bound.** As discussed in Section 3.3.4, our proposed MA-loss is applicable to CIR under the assumption that our linear query construction is complex enough to express a meaningful relative transformation. Our proposed MA-loss is applicable to CIR under two main assumptions:

- The linear fusion strategy is powerful enough to effectively combine the source image with the relative caption
- When using an image-text pair dataset, we can effectively approximate a textual transformation using the difference of the captions' embeddings

The first assumption has partially been tested by some recent works that adopted a similar fusion strategy [4] [6], but still requires additional experiments for a fair comparison with our experimental setup. The second assumption is paramount to the zero-shot transfer and is measurable by comparing the supervised performance with our zero-shot methods.

To empirically evaluate these assumptions, we designed a supervised version of our MA-loss, namely, **MA-CIR**. This loss is used to fine-tune CLIP directly on the train split of our CIR datasets and serves as an important upper bound comparison for our zero-shot methods.

**Zero-shot methods.** Following the methods discussed in Section 3.3.1, we evaluate our proposed **MA-loss** in two different directional configurations using the best performing weighting strategies. A more comprehensive exploration of the weighting strategies is presented in Table 4.11.

| **Method** | **R@10** | **R@50** | **Avg.** |
|---|---|---|---|
| W/O post-pre-training | 8.59 | 19.88 | 14.24 |
| CLIP loss | 11.53 | 24.87 | 18.20 |
| MA-CIR (supervised) | 25.64 | 46.92 | 36.28 |
| MA-loss (bi) | 19.47 | 35.45 | 27.46 |
| MA-loss (mono) | 19.47 | 35.62 | 27.55 |

**Table 4.9:** Performance on FashionIQ of our zero-shot methods (bottom rows) compared to the zero-shot baselines (top rows) and the supervised upper bound (center row).

| **Method** | **R@k** | | | | **$\mathbf{R_{subset}}$@k** | | |
|---|---|---|---|---|---|---|---|
| | k=1 | k=5 | k=10 | k=50 | k=1 | k=2 | k=3 |
| W/O post-pre-training | 10.84 | 32.27 | 46.71 | 75.47 | 30.15 | 53.78 | 73.78 |
| CLIP loss | 12.91 | 37.16 | 51.95 | 80.00 | 34.05 | 58.53 | 78.00 |
| MA-CIR (supervised) | 23.50 | 54.07 | 67.78 | 89.16 | 51.88 | 74.65 | 87.71 |
| MA-loss (bi) | 19.01 | 46.434 | 61.831 | 86.145 | 44.63 | 68.27 | 83.54 |
| MA-loss (mono) | 18.94 | 46.19 | 61.86 | 86.36 | 44.55 | 68.19 | 83.64 |

**Table 4.10:** Performance on CIRR of our zero-shot methods (bottom rows) compared to the zero-shot baselines (top rows) and the supervised upper bound (center row).

Table 4.9 and Table 4.10 report the performance of our zero-shot methods compared to the baselines and the supervised upper bound, on FashionIQ and CIRR, respectively. Across both FashionIQ and CIRR, the supervised MA-CIR model establishes a clear upper bound that is essential for interpreting the zero-shot results. Since the supervised model directly learns the transformation from annotated CIR data, the gap between this upper bound and our zero-shot variants reflects how well the MA-loss can approximate CIR-specific relational reasoning without any task-level supervision. On FashionIQ, the supervised MA-CIR score (36.28 average) is notably higher than all zero-shot methods, showing a gap of roughly nine points over the best zero-shot variant. This difference is substantial and indicates that, while our MA-loss effectively improves over the unsupervised baselines, zero-shot transfer still captures only part of the transformation needed for FashionIQ-style attribute modifications.

The CIRR results in Table 4.10 follow a similar structure. The unsupervised baselines remain the weakest performers, and the supervised MA-CIR model again provides a strong upper bound. Here, the bidirectional and monodirectional variants perform almost identically on global retrieval metrics, with the bidirectional version showing a slight advantage at lower k.

Across both datasets, the zero-shot models improve meaningfully over unadapted CLIP and over CLIP-loss post-pre-training, but they do not close the distance to the supervised upper bound. This distance is informative: it quantifies how much of CIR reasoning can be recovered through multimodal arithmetic alone. The smaller the gap, the stronger the evidence that the underlying transformation is expressible through text–image relations present in MSCOCO. The larger the gap, the more the transformation appears to depend on dataset-specific structure that must be learned explicitly. These results show that our MA-loss achieves non-trivial zero-shot transfer while still leaving room for improvement before matching supervised performance

### 4.4.1 Weighting strategies comparison

As we did for multimodal arithmetic, we compare the different weighting strategies described in Section 3.3.3, evaluating the different combinations on FashionIQ. This comparison is then used to select the best configuration for CIR used for the other results reported in this section.

These results reveal several interesting dynamics regarding the behavior of the proposed MA-loss on CIR. In contrast with the SIMAT results, where bidirectional training offered a clear advantage, CIR retrieval performance displays a more nuanced pattern. The monodirectional formulation reaches the overall highest score in the unweighted configuration, with R@10 of 19.466 and R@50 of 35.615, slightly outperforming the corresponding bidirectional baseline. This suggests that,

| **Loss direction** | **Weighting strategy** | **Frozen weights** | **R@10** | **R@50** | Avg. |
|---|---|---|---|---|---|
| Bi | None | - | 15.841 | 30.517 | 23.179 |
| Bi | Images | ✓ | 19.324 | 35.597 | <u>27.461</u> |
| Bi | Images | ✗ | 19.466 | 35.447 | **27.457** |
| Bi | Texts | ✓ | 15.514 | 30.340 | 22.927 |
| Bi | Texts | ✗ | 15.585 | 30.261 | 22.923 |
| Mono | None | - | 19.466 | 35.615 | **27.541** |
| Mono | Images | ✓ | 19.316 | 35.606 | <u>27.461</u> |
| Mono | Images | ✗ | 19.474 | 35.438 | 27.456 |
| Mono | Texts | ✓ | 19.033 | 35.430 | 27.232 |
| Mono | Texts | ✗ | 19.219 | 35.350 | 27.285 |

**Table 4.11:** Comparison of recall scores on FashionIQ with the MA-loss in different directional configurations and with different weighting strategies. Due to the small numerical differences, we report the scores up to the third decimal digit.

for CIR, enforcing the compositional constraint in a single direction may already provide sufficient structure for the retrieval objective, and that adding the reverse constraint does not necessarily yield additional benefits. This is consistent with previous findings, which argued that CIR is inherently asymmetrical and thus benefits from a monodirectional transformation [4]. Still, the differences remain small, indicating that both directionalities behave comparably once the model is adapted to the CIR setting.

The impact of weighting strategies differs markedly from what was observed on SIMAT. In the bidirectional case, image-based weighting leads to large improvements over the unweighted baseline, raising R@10 from 15.8 to more than 19.3 and R@50 from 30.5 to around 35.5. This shows that, unlike SIMAT, where image-based weighting was consistently suboptimal, CIR retrieval benefits strongly from emphasizing visual similarities during the loss computation. Text-based weighting, however, has the opposite effect: both frozen and unfrozen variants decrease performance relative to the unweighted bidirectional baseline. The asymmetry between image-based and text-based weighting in this setting hints at a retrieval behavior driven primarily by visual attributes, which is consistent with the nature of CIR.

In the monodirectional setting, weighting strategies induce far milder effects. Neither image nor text-based weighting meaningfully improves over the unweighted baseline, and all variants remain extremely close to the best monodirectional score. This suggests that, once the loss is constrained in a single direction, the model becomes less sensitive to reweighting of the multimodal pairs, likely because the directional constraint already biases the optimization toward the visual path relevant for CIR.

The effect of freezing the weights also varies across strategies. For image-based weighting, frozen and unfrozen configurations behave almost identically in both directionalities, mirroring the observations made for SIMAT. For text-based weighting, freezing the weights has only minor effects, consistently positive in the bidirectional case and neutral or slightly negative in the monodirectional one. Overall, weight freezing appears to influence CIR performance far less than on SIMAT, and its impact does not seem tied to a clear trend across settings.

Comparing these findings to the SIMAT results highlights the task-dependent nature of the MA-loss behavior. On SIMAT, bidirectionality and text-based weighting provided the strongest gains, whereas on FashionIQ, the best results arise either from monodirectional training without weighting or from bidirectional training with image-based weighting. This divergence suggests that CIR relies more heavily on visual similarity structure, while multimodal arithmetic benefits from richer semantic balancing introduced by text-driven weighting and bidirectionality. Together, the two sets of results illustrate how the optimal MA-loss configuration depends strongly on the characteristics of the downstream task and the modality that carries the most discriminative information.

### 4.4.2 Comparison with the state of the art

We conclude our experiments by comparing our methods with the leading works in the state-of-the-art (SoTA). The values reported in Tables Table 4.12 and Table 4.13 represent the performance of various models on FashionIQ and CIRR, as cited in [49]. All models considered were trained using CLIP-B. To ensure a fair comparison, we also include an additional version of our model, denoted as **U**, which is obtained by post-pre-training CLIP on all parameters.

| **Method** | **R@10** | **R@50** | **Avg.** |
|---|---|---|---|
| MA-loss (bi) | 19.47 | 35.45 | 27.46 |
| MA-loss (mono) | 19.47 | 35.62 | 27.55 |
| MA-loss **U** (mono) | 18.07 | 34.16 | 26.12 |
| SEARLE | 22.89 | 42.53 | 32.71 |
| MagicLens | 26.30 | 47.40 | 36.90 |
| MTI | 31.31 | 53.24 | 42.28 |

**Table 4.12:** Zero-shot performance on FashionIQ of our methods compared to the SoTA.

| **Method** | **R@k** | | | | **$R_{subset}$@k** | | |
|---|---|---|---|---|---|---|---|
| | k=1 | k=5 | k=10 | k=50 | k=1 | k=2 | k=3 |
| MA-loss (bi) | 19.01 | 46.434 | 61.831 | 86.145 | 44.63 | 68.27 | 83.54 |
| MA-loss (mono) | 18.94 | 46.19 | 61.86 | 86.36 | 44.55 | 68.19 | 83.64 |
| MA-loss **U** (mono) | 21.08 | 46.96 | 60.89 | 84.82 | 54.77 | 75.57 | 87.04 |
| SEARLE | 24.00 | 53.42 | 66.82 | 89.78 | 54.89 | 76.60 | 88.19 |
| MTI | 18.80 | 46.07 | 60.75 | 86.41 | 44.29 | 68.10 | 83.42 |
| MagicLens | 27.00 | 58.00 | 76.90 | 91.10 | 66.70 | 83.90 | 92.40 |

**Table 4.13:** Zero-shot performance on CIRR of our methods compared to the SoTA.

On FashionIQ, the unlocked variant performs slightly worse than both the bidirectional and monodirectional MA-loss configurations. While the standard MA-loss models reach an average score of about 27.5, MA-loss U drops to 26.12. This performance drop suggests that full post-pre-training on MSCOCO does not transfer well to FashionIQ. This is because FashionIQ requires modeling very fine-grained visual differences across similar items, often tied to subtle attributes such as fit, texture, or local color changes. Updating all CLIP parameters on MSCOCO, a dataset not designed for such delicate distinctions, may distort parts of the embedding space that FashionIQ relies on. In contrast, the post-pre-training limited to the projection layers of MA-loss modifies the representation more conservatively and thus retains the structure needed for fine-grained, attribute-driven retrieval.

When compared to zero-shot SoTA systems, our MA-loss variants remain below SEARLE, MagicLens, and MTI, all of which were also trained on CLIP-B and evaluated in zero-shot mode. The gap reflects the relative difficulty of transferring FashionIQ-style fine-grained relational information purely from MSCOCO. MTI, for instance, reaches an average of 42.28, significantly higher than our best result. Still, our methods improve substantially over non-arithmetic baselines presented earlier, showing that MA-loss introduces meaningful transferable structure even without CIR-specific training.

On CIRR (Table 4.13), the behavior changes. The locked bidirectional and monodirectional MA-loss variants again achieve similar results, but the unlocked model performs noticeably better on the subset metrics, with an Rsubset@1 of 54.77 compared to roughly 44.6 for the other variants. CIRR contains broader domain diversity than FashionIQ, with less emphasis on tiny attribute differences and more on general relational cues. Full post-training on MSCOCO can therefore be advantageous: adapting the entire encoder helps align the representation to a distribution more compatible with CIRR's open-domain content. This explains why

MA-loss U outperforms the other MA-loss variants on the subset-based rankings and performs competitively on global metrics.

Compared to the zero-shot SoTA, our methods still fall below MagicLens, which remains the strongest performer across most metrics. However, the unlocked MA-loss model narrows the gap more effectively on CIRR than on FashionIQ. Its subset-level performance, in particular, approaches SEARLE and surpasses MTI. This difference with FashionIQ underlines that the usefulness of full post-pre-training depends heavily on the dataset.

Overall, these results reinforce a pattern that is consistent with the detailed analyses carried out earlier in this chapter. In the SIMAT experiments, we observed that the bidirectional MA-loss and text-based weighting strategies were particularly effective, suggesting that SIMAT benefits from strong semantic consistency and symmetry between modalities. In other words, SIMAT relies heavily on the structure of the multimodal space and on preserving fine semantic relations during post-pre-training.

In contrast, the FashionIQ and CIRR results show that CIR tasks do not respond uniformly to these same principles. FashionIQ, which depends on very subtle attribute modifications, is harmed by full post-pre-training and benefits from keeping CLIP's fine-grained structure largely intact. CIRR, on the other hand, involves more general relational reasoning over a broader visual domain, and therefore benefits from unlocking all parameters during training. By comparing these behaviors with those observed on SIMAT, we see more clearly that the effectiveness of the MA-loss, and especially the choice between partial or full parameter updates, is highly dependent on the type of compositionality and granularity required by each downstream task.

# Chapter 5

# Conclusions

This thesis explored whether compositional reasoning can emerge from vision language models trained exclusively on image-text pairs. The goal was to understand how semantic transformations can be encoded in a shared embedding space and how this capability can be leveraged for multimodal arithmetic and zero-shot composite image retrieval. The work followed a progression from theoretical analysis to empirical evaluation, and the key findings are summarized in this chapter.

The first part of the thesis focused on understanding how embedding space geometry influences multimodal arithmetic. Through an extensive analysis, it was shown that the classical interpretation of the modality gap is not sufficient to explain performance. In particular, the results indicated that the magnitude of the gap, commonly measured by the centroid distance between modalities, correlates only weakly with SIMAT scores. Instead, the study found that the variance of the modality delta vectors, and more generally their alignment across samples, are better predictors of compositional performance. This observation motivated the formalization of cross-modal semantic consistency, which captures the idea that semantic differences should be represented similarly across modalities.

To quantify this concept, the thesis introduced a metric based on squared residuals. The analysis of this metric revealed two relevant insights. First, improvements in cross-modal semantics coincide with reductions in the variance of the modality delta vectors, even when the modality gap does not change significantly. Second, extremely small variance can be obtained by collapsing modality embeddings towards a single point, but this leads to poor semantic expressivity. Therefore, the metric identifies useful regions of the embedding space that balance consistency with meaningful variation. This understanding guided the development of a new learning objective.

The second part of the thesis introduced the Multimodal Arithmetic Loss. This loss aims to align transformations of visual and textual embeddings by directly supervising differences rather than absolute points. The loss was characterized both analytically and empirically. Theoretical considerations showed that MA Loss preserves the invariances of the contrastive objective but introduces a supervision signal that encourages similarity between cross-modal semantic deltas. Experimental observations confirmed this behavior. During post-pre-training, the SIMAT score exhibited a non-monotonic trend, where an initial degradation was followed by recovery and stabilization. This phenomenon appeared both in the mono-directional and bi-directional versions of MA Loss, although the latter recovered more effectively. The study also showed that MA Loss reshapes similarity distributions by reducing the overlap between positive and negative samples, which is consistent with its intended effect on semantic differences.

A key part of the experimental study involved understanding the effect of different weighting strategies. The results showed that the best strategy depends on the task. On SIMAT, strategies based on textual similarity, especially those using the frozen model, consistently outperformed visual-based variants. These strategies provided a more stable signal for selecting compatible caption pairs and improved the representation of the semantic transformations. In contrast, the CIR experiments showed a different pattern. Visual-based weighting performed competitively or better in several settings, and the advantage of frozen textual weighting was less consistent. This difference suggests that CIR, which relies on natural image variation and relative captions, benefits from visual similarity cues in ways that SIMAT does not. Overall, the results show that weighting strategies influence the model in task-dependent ways rather than producing a single optimal choice.

After establishing the behavior of MA Loss, the thesis evaluated its performance on downstream tasks. On SIMAT, MA Loss achieved a new state of the art with a score of 48 percent, improving over the previously reported 42 percent. This confirmed that supervising semantic differences strengthens the model's ability to perform structured transformations. The analysis of the embedding space during training also showed that improvements on SIMAT correlate with improvements in cross-modal consistency rather than with changes in classical metrics.

The evaluation of composite image retrieval produced more nuanced results. The zero-shot models were tested on FashionIQ and CIRR without training on any triplet-based supervision. On FashionIQ, all MA Loss variants remained below the zero-shot state of the art. The best MA Loss configuration achieved an average score of roughly 27.5, which is significantly lower than MTI, MagicLens, and SEARLE, all of which surpassed 30 points and, in some cases, exceeded 40. These results show that the fine-grained attribute differences required by FashionIQ are not fully recoverable from MSCOCO-level supervision.

On CIRR, the picture was different. While our methods did not reach the state of the art and remained below MagicLens and SEARLE on the main retrieval metrics, some configurations performed competitively with other zero-shot baselines.

In particular, the unlocked variant improved subset-level metrics and surpassed MTI on most scores, reflecting the fact that CIRR benefits more from broad domain alignment than from fine-grained attribute structure. This behavior highlights that full post-pre-training on MSCOCO can be useful for datasets that rely on general relational cues rather than subtle attribute modifications. Still, the results also showed a persistent gap between zero-shot MA-Loss and the supervised upper bound, confirming that some aspects of CIRR-style transformations require dataset-specific supervision.

The thesis makes three main contributions. First, it provides a detailed analysis of how embedding space geometry affects multimodal arithmetic, highlighting the limitations of classical metrics and the importance of semantic delta alignment. Second, it introduces the Multimodal Arithmetic Loss, which operationalizes these insights using only image-text pairs. Third, it demonstrates that MA Loss improves both multimodal arithmetic and zero-shot composite image retrieval, establishing new state-of-the-art results on SIMAT and achieving strong performance on FashionIQ and CIRR.

### Future directions

While the results are encouraging, several research directions could further expand this work. First, MA Loss currently relies on a linear fusion of embeddings. More expressive fusion strategies, such as non-linear transformations inspired by CLIP4CIR [4] or goniometric interpolations [25], could capture richer patterns in semantic relationships. These techniques may also enhance robustness when the relationship between image and text embeddings is not well approximated by linear differences. Second, the post-training procedure uses raw captions from MSCOCO to simulate relative descriptions. A dedicated data augmentation pipeline could help the model learn a wider variety of transformations. Such augmentation could incorporate paraphrasing, attribute manipulation, or structured templates to emulate the nuances of real relative captions. Third, MA Loss exhibits non-monotonic training behavior, which suggests that the optimization landscape is complex. Future research could further explore this phenomenon and improve the optimization process. This might include things like regularization techniques, adaptive scheduling, or multi-phase optimization.

# Appendix A

# Proofs

## A.1 Relationship between MSPD and total variance

Here we present a well-known property of the total variance of a set of vectors. This property shows the relationship between MSPD (Mean Squared Pairwise Difference) and the total variance.

**Theorem A.1.1.** *Let's consider a set of vectors $X = \{x_i\}_{i=1}^N$ in $\mathbb{R}^d$. Then:*

$$\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \|x_i - x_j\|^2 = \frac{2N}{N-1} \mathbb{V}ar\left[X\right] \tag{A.1}$$

*Proof.* We start by decomposing the squared norm inside the definition of MSPD. For simplicity we define the constant $K = \frac{1}{N(N-1)}$.

$$K \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \|x_i - x_j\|^2 = K \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left(\|x_i\|^2 + \|x_j\|^2 - 2x_i \cdot x_j\right) \tag{A.2}$$

$$= K \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \|x_i\|^2 + K \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \|x_j\|^2 - 2K \sum_{i=1}^{N} x_i \sum_{\substack{j=1 \\ j \neq i}}^{N} x_j \tag{A.3}$$

In the first double sum, we notice that each element $x_i$ appears $(N-1)$ times; analogously, each element $x_j$ also appears in the second double sum $(N-1)$ times. We also note that $\sum_{j=1}^{N} x_j = N\mu$, where $\mu$ is the mean vector of X; removing the $j = i$ index, we have that $\sum_{\substack{j=1 \\ j \neq i}}^{N} x_j = N\mu - x_i$. Using these results, we can rewrite the above expression as follows:

$$\text{MSPD} = 2K(N-1)\sum_{i=1}^{N} \|x_i\|^2 - 2K\sum_{i=1}^{N} x_i(N\mu - x_i) \tag{A.4}$$

$$= 2K(N-1)\sum_{i=1}^{N} \|x_i\|^2 - 2KN\mu\sum_{i=1}^{N} x_i + 2K\sum_{i=1}^{N} \|x_i\|^2 \tag{A.5}$$

Again, we note that $\sum_{i=1}^{N} x_i = N\mu$, so:

$$\text{MSPD} = 2K(N-1)\sum_{i=1}^{N} \|x_i\|^2 - 2KN^2\|\mu\|^2 + 2K\sum_{i=1}^{N} \|x_i\|^2 \tag{A.6}$$

$$= 2KN\sum_{i=1}^{N} \|x_i\|^2 - 2KN^2\|\mu\|^2 \tag{A.7}$$

$$= 2KN\left(\sum_{i=1}^{N} \|x_i\|^2 - N\|\mu\|^2\right) \tag{A.8}$$

Using the definition of the constant $K = \frac{1}{N(N-1)}$:

$$\text{MSPD} = \frac{2}{N-1}\left(\sum_{i=1}^{N} \|x_i\|^2 - N\|\mu\|^2\right) \tag{A.9}$$

Finally, we multiply and divide by N to arrive at the definition of total variance:

$$\text{MSPD} = \frac{2N}{N-1}\left(\frac{1}{N}\sum_{i=1}^{N} \|x_i\|^2 - \|\mu\|^2\right) \tag{A.10}$$

$$= \frac{2N}{N-1}\mathbb{V}ar\left[X\right] \tag{A.11}$$

□

## A.2 Proof of Theorem 3.2.2

**Theorem.** We prove the following equality:

$$\text{XSC-SR} = \frac{2N}{N-1}\mathbb{V}ar\left[h^v\right] + \frac{2N}{N-1}\mathbb{V}ar\left[h^t\right] + 4(\text{MNS} - \text{MPS}) \tag{A.12}$$

Where:

$$\text{MPS} = \frac{1}{N}\sum_{i=1}^{N} h_i^v \cdot h_i^t \tag{A.13}$$

$$\text{MNS} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} h_i^v \cdot h_j^t \tag{A.14}$$

*Proof.* Recall the definition of XSC-SR:

$$\text{XSC-SR} := \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \| h_j^t - h_i^t - h_j^v + h_i^v \|^2 \tag{A.15}$$

We define two vectors $\delta_{ij}^v = h_i^v - h_j^v \in \mathbb{R}^d$, $\delta_{ij}^t = h_i^t - h_j^t \in \mathbb{R}^d$ and rewrite the above equation as follows:

$$\text{XSC-SR} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \| \delta_{ij}^v - \delta_{ij}^t \|^2 \tag{A.16}$$

We note that, when $i = j$, we have $\delta_{ij}^v = \delta_{ij}^t = 0$. So we can rewrite the double sum:

$$\text{XSC-SR} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \| \delta_{ij}^v - \delta_{ij}^t \|^2 \tag{A.17}$$

and splitting the squared norm we get:

$$\begin{aligned} \text{XSC-SR} &= \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left( \|\delta_{ij}^v\|^2 + \|\delta_{ij}^t\|^2 - 2\delta_{ij}^v \cdot \delta_{ij}^t \right) &\text{(A.18)} \\ &= \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left( \|\delta_{ij}^v\|^2 + \|\delta_{ij}^t\|^2 \right) - \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \delta_{ij}^v \cdot \delta_{ij}^t &\text{(A.19)} \end{aligned}$$

We then proceed by noticing that:

$$\begin{aligned} \delta_{ij}^v \cdot \delta_{ij}^t &= (h_i^v - h_j^v) \cdot (h_i^t - h_j^t) &\text{(A.20)} \\ &= h_i^v \cdot h_i^t - h_i^v \cdot h_j^t - h_j^v \cdot h_i^t + h_j^v \cdot h_j^t &\text{(A.21)} \end{aligned}$$

thus:

$$\begin{aligned} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \delta_{ij}^v \cdot \delta_{ij}^t &= \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} (h_i^v \cdot h_i^t + h_j^v \cdot h_j^t) - \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} (h_i^v \cdot h_j^t + h_j^v \cdot h_i^t) &\text{(A.22)} \\ &= 2(N-1) \sum_{i=1}^{N} h_i^v \cdot h_i^t - 2 \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} h_i^v \cdot h_j^t &\text{(A.23)} \end{aligned}$$

At the same time, using a well-known property of the variance (A.1), we have:

$$\begin{aligned} \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \|\delta_{ij}^v\|^2 &= \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \|h_i^v - h_j^v\|^2 &\text{(A.24)} \\ &= \frac{2N}{(N-1)} \mathbb{V}ar\left[h^v\right] &\text{(A.25)} \end{aligned}$$

and analogously:

$$\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\|\delta_{ij}^{t}\|^2 = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\|h_i^t - h_j^t\|^2 \tag{A.26}$$

$$= \frac{2N}{(N-1)}\mathbb{V}ar\left[h^t\right] \tag{A.27}$$

Finally, we substitue equations (A.22), (A.25), and (A.27) into (A.19):

$$\text{XSC-SR} = \frac{2N}{N-1}\mathbb{V}ar\left[h^v\right] + \frac{2N}{N-1}\mathbb{V}ar\left[h^t\right] - \frac{4}{N}\sum_{i=1}^{N} h_i^v h_i^t + \frac{4}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N} h_i^v h_j^t \tag{A.28}$$

And using the definitions for MPS and MNS, we conclude our proof:

$$\text{XSC-SR} = \frac{2N}{N-1}\mathbb{V}ar\left[h^v\right] + \frac{2N}{N-1}\mathbb{V}ar\left[h^t\right] + 4(\text{MNS} - \text{MPS}) \tag{A.29}$$

□

# Appendix B

# Implementations

## B.1 Pytorch code for the MA-loss

Below, we report the code for a simple PyTorch function to efficiently compute the weighted version of the unidirectional MA-loss.

**Listing B.1:** PyTorch implementation of the unidirectional MA-loss.

```python
def compute_MAq2i_sw_loss(image_features, text_features, temperature=1, lambd=1):
    N = image_features.shape[0]
    device = image_features.device

    # Expand for broadcasting: (N, N, D)
    img_i = image_features.unsqueeze(1)    # (N, 1, D)
    txt_j = text_features.unsqueeze(0)     # (1, N, D)
    txt_i = text_features.unsqueeze(1)     # (N, 1, D)

    # Compute y for all i, j:  y[i, j, :] = image_features[i] + lambd * (text_features[j] - text_features[i])
    y = img_i + lambd * (txt_j - txt_i)   # (N, N, D)
    y = y / y.norm(dim=-1, keepdim=True)  # Normalize along last dim

    # Reshape to (N*N, D)
    query = y.reshape(-1, y.shape[-1])

    # Labels: for each i, repeat torch.arange(N) N times
    labels = torch.arange(N, device=device).repeat(N)

    # Normalize text features
    text_features = text_features / text_features.norm(dim=-1, keepdim=True)
    self_similarities = text_features @ text_features.T  # (N, N)
    # Compute weights (N*N,)
    weights = self_similarities.relu().pow(2).flatten()

    logits = temperature * query @ image_features.t() #(N*N, N)

    unweighted_loss = torch.nn.functional.cross_entropy(logits, labels, reduction='none')
    return (unweighted_loss * weights).sum() / weights.sum()
```

# Bibliography

[1] L. Agnolucci, A. Baldrati, A. Del Bimbo, and M. Bertini, «Isearle: Improving textual inversion for zero-shot composed image retrieval», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025 (cit. on pp. 14, 16).

[2] M. Artetxe and H. Schwenk, «Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond», *Transactions of the association for computational linguistics*, vol. 7, pp. 597–610, 2019 (cit. on p. 11).

[3] A. Baldrati, L. Agnolucci, M. Bertini, and A. Del Bimbo, «Zero-shot composed image retrieval with textual inversion», in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 338–15 347 (cit. on pp. 14–16).

[4] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo, «Conditioned and composed image retrieval combining and partially fine-tuning clip-based features», in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4959–4968 (cit. on pp. 16, 33, 38, 55, 58, 64).

[5] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo, «Effective conditioned and composed image retrieval combining clip-based features», in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 21 466–21 474 (cit. on p. 16).

[6] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo, «Composed image retrieval using contrastive learning and task-oriented clip-based features», *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 3, pp. 1–24, 2023 (cit. on pp. 16, 55).

[7] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, «Enriching word vectors with subword information», *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017 (cit. on p. 11).

[8] J. Chen and H. Lai, «Pretrain like your inference: Masked tuning improves zero-shot composed image retrieval», *arXiv preprint arXiv:2311.07622*, 2023 (cit. on p. 15).

[9] Y. Chen, J. Zhou, and Y. Peng, «Spirit: Style-guided patch interaction for fashion image retrieval with text feedback», *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 6, pp. 1–17, 2024 (cit. on p. 17).

[10] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, «Uniter: Universal image-text representation learning», in *European conference on computer vision*, Springer, 2020, pp. 104–120 (cit. on pp. 5, 6).

[11] N. Cohen, R. Gal, E. A. Meirom, G. Chechik, and Y. Atzmon, «"this is my unicorn, fluffy": Personalizing frozen vision-language representations», in *European conference on computer vision*, Springer, 2022, pp. 558–577 (cit. on p. 14).

[12] G. Couairon, M. Douze, M. Cord, and H. Schwenk, «Embedding arithmetic of multimodal queries for image retrieval», in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4950–4958 (cit. on pp. 9, 11, 12, 25, 29, 41).

[13] Y. Dar, L. Luzi, and R. G. Baraniuk, «Frozen overparameterization: A double descent perspective on transfer learning of deep neural networks», *arXiv preprint arXiv:2211.11074*, 2022 (cit. on p. 54).

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, «Imagenet: A large-scale hierarchical image database», in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255 (cit. on p. 7).

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, «Bert: Pre-training of deep bidirectional transformers for language understanding», in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186 (cit. on pp. 5, 6).

[16] A. Dosovitskiy *et al.*, «An image is worth 16x16 words: Transformers for image recognition at scale», *arXiv preprint arXiv:2010.11929*, 2020 (cit. on pp. 5, 6).

[17] S. Eslami and G. de Melo, «Mitigate the gap: Investigating approaches for improving cross-modal alignment in clip», *arXiv preprint arXiv:2406.17639*, 2024 (cit. on pp. 8, 10).

[18] A. Fahim, A. Murphy, and A. Fyshe, «It's not a modality gap: Characterizing and addressing the contrastive gap», *arXiv preprint arXiv:2405.18570*, 2024 (cit. on pp. 8–11, 23, 24, 46–48, 52).

[19] X. Fang, J. Li, Q. Sun, and B. Wang, «Rethinking the uniformity metric in self-supervised learning», *arXiv preprint arXiv:2403.00642*, 2024 (cit. on p. 23).

[20] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, «Language-agnostic bert sentence embedding», *arXiv preprint arXiv:2007.01852*, 2020 (cit. on p. 11).

[21] G. Gu, S. Chun, W. Kim, Y. Kang, and S. Yun, «Language-only training of zero-shot composed image retrieval», in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 225–13 234 (cit. on p. 15).

[22] X. Guo, H. Wu, Y. Gao, S. Rennie, and R. Feris, «The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback», *arXiv preprint arXiv:1905.12794*, vol. 1, no. 2, p. 7, 2019 (cit. on pp. 13, 42).

[23] K. He, X. Zhang, S. Ren, and J. Sun, «Deep residual learning for image recognition», in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: `10.1109/CVPR.2016.90` (cit. on pp. 5, 6).

[24] J. Jang, C. Kong, D. Jeon, S. Kim, and N. Kwak, «Unifying vision-language representation space with single-tower transformer», in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, 2023, pp. 980–988 (cit. on p. 5).

[25] Y. K. Jang, D. Huynh, A. Shah, W.-K. Chen, and S.-N. Lim, «Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval», in *European Conference on Computer Vision*, Springer, 2024, pp. 239–254 (cit. on pp. 15, 64).

[26] C. Jia *et al.*, «Scaling up visual and vision-language representation learning with noisy text supervision», in *International conference on machine learning*, PMLR, 2021, pp. 4904–4916 (cit. on pp. 5, 6, 12).

[27] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, «Analyzing and improving the image quality of stylegan», in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119 (cit. on p. 11).

[28] S. Karthik, K. Roth, M. Mancini, and Z. Akata, «Vision-by-language for training-free compositional image retrieval», *arXiv preprint arXiv:2310.09291*, 2023 (cit. on p. 15).

[29] D. Kingma and J. Ba, «Adam: A method for stochastic optimization», *International Conference on Learning Representations*, Dec. 2014 (cit. on p. 40).

[30] R. Krishna *et al.*, «Visual genome: Connecting language and vision using crowdsourced dense image annotations», *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017 (cit. on p. 41).

[31] X. Li *et al.*, «Oscar: Object-semantics aligned pre-training for vision-language tasks», in *European conference on computer vision*, Springer, 2020, pp. 121–137 (cit. on p. 42).

[32] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou, «Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning», *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 612–17 625, 2022 (cit. on pp. 8, 9, 52).

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, «Microsoft coco: Common objects in context», in *European conference on computer vision*, Springer, 2014, pp. 740–755 (cit. on pp. 38, 41).

[34] Z. Liu, C. Rodriguez-Opazo, D. Teney, and S. Gould, «Image retrieval on real-life images with pre-trained vision-and-language models», in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2125–2134 (cit. on pp. 14, 45).

[35] I. Loshchilov and F. Hutter, «Decoupled weight decay regularization», *arXiv preprint arXiv:1711.05101*, 2017 (cit. on p. 40).

[36] V. Maiorca, L. Moschella, A. Norelli, M. Fumero, F. Locatello, and E. Rodolà, «Latent space translation via semantic alignment», *Advances in Neural Information Processing Systems*, vol. 36, pp. 55 394–55 414, 2023 (cit. on p. 10).

[37] T. Mikolov, K. Chen, G. Corrado, and J. Dean, «Efficient estimation of word representations in vector space», *arXiv preprint arXiv:1301.3781*, 2013 (cit. on pp. 11, 25).

[38] L. Moschella, V. Maiorca, M. Fumero, A. Norelli, F. Locatello, and E. Rodolà, «Relative representations enable zero-shot latent space communication (2023)», *arXiv preprint arXiv:2209.15430*, 2023 (cit. on p. 10).

[39] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, «Deep double descent: Where bigger models and more data hurt», *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, no. 12, p. 124 003, 2021 (cit. on p. 54).

[40] A. v. d. Oord, Y. Li, and O. Vinyals, «Representation learning with contrastive predictive coding», *arXiv preprint arXiv:1807.03748*, 2018 (cit. on pp. 6, 19).

[41] A. Paszke *et al.*, «Pytorch: An imperative style, high-performance deep learning library», in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf` (cit. on p. 39).

[42] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, «Styleclip: Text-driven manipulation of stylegan imagery», in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2085–2094 (cit. on p. 11).

[43] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, «Language models are unsupervised multitask learners», *OpenAI blog*, vol. 1, no. 8, p. 9, 2019 (cit. on pp. 5, 6, 39).

[44] A. Radford *et al.*, «Learning transferable visual models from natural language supervision», in *International conference on machine learning*, PmLR, 2021, pp. 8748–8763 (cit. on pp. 5–7, 11, 14, 19, 20, 39, 46, 55).

[45] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, «Hierarchical text-conditional image generation with clip latents», *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022 (cit. on p. 11).

[46] K. Saito, K. Sohn, X. Zhang, C.-L. Li, C.-Y. Lee, K. Saenko, and T. Pfister, «Pic2word: Mapping pictures to words for zero-shot composed image retrieval», in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 305–19 314 (cit. on p. 15).

[47] P. Shi, M. Welle, M. Bjørkman, and D. Kragic, «Understanding the modality gap in clip», *ICLR, Stockholm, Sweden*, 2023 (cit. on pp. 8–10).

[48] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela, «Flava: A foundational language and vision alignment model», in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 15 638–15 650 (cit. on pp. 5, 6).

[49] X. Song, H. Lin, H. Wen, B. Hou, M. Xu, and L. Nie, «A comprehensive survey on composed image retrieval», *ACM Transactions on Information Systems*, 2025 (cit. on pp. 13, 59).

[50] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, «A corpus for reasoning about natural language grounded in photographs», *arXiv preprint arXiv:1811.00491*, 2018 (cit. on p. 45).

[51] M. Tschannen, B. Mustafa, and N. Houlsby, «Clippo: Image-and-language understanding from pixels only», in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 006–11 017 (cit. on p. 5).

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, «Attention is all you need», *Advances in neural information processing systems*, vol. 30, 2017 (cit. on p. 5).

[53] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays, «Composing text and image for image retrieval-an empirical odyssey», in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6439–6448 (cit. on pp. 13, 14).

[54] F. Wang and H. Liu, «Understanding the behaviour of contrastive loss», in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2495–2504 (cit. on p. 20).

[55] T. Wang and P. Isola, «Understanding contrastive representation learning through alignment and uniformity on the hypersphere», in *International conference on machine learning*, PMLR, 2020, pp. 9929–9939 (cit. on p. 8).

[56] Y. Wang, E. Riddell, A. Chow, S. Sedwards, and K. Czarnecki, «Mitigating the modality gap: Few-shot out-of-distribution detection with multi-modal prototypes and image bias estimation», *arXiv preprint arXiv:2502.00662*, 2025 (cit. on pp. 9, 10).

[57] R.-D. Wu, Y.-Y. Lin, and H.-F. Yang, «Training-free zero-shot composed image retrieval via weighted modality fusion and similarity», in *International Conference on Technologies and Applications of Artificial Intelligence*, Springer, 2024, pp. 77–90 (cit. on p. 15).

[58] S. Yamaguchi, D. Feng, S. Kanai, K. Adachi, and D. Chijiwa, «Post-pre-training for modality alignment in vision-language foundation models», in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 4256–4266 (cit. on pp. 5, 55).

[59] Z. Yang, S. Qian, D. Xue, J. Wu, F. Yang, W. Dong, and C. Xu, «Semantic editing increment benefits zero-shot composed image retrieval», in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1245–1254 (cit. on p. 15).

[60] L. Yao *et al.*, «Filip: Fine-grained interactive language-image pre-training», *arXiv preprint arXiv:2111.07783*, 2021 (cit. on p. 12).

[61] C. Yaras, S. Chen, P. Wang, and Q. Qu, «Explaining and mitigating the modality gap in contrastive multimodal learning», *arXiv preprint arXiv:2412.07909*, 2024 (cit. on pp. 8–10, 18, 20, 47, 52).

[62] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, «Coca: Contrastive captioners are image-text foundation models», *arXiv preprint arXiv:2205.01917*, 2022 (cit. on p. 5).

[63] F. Zhang, M. Xu, Q. Mao, and C. Xu, «Joint attribute manipulation and modality alignment learning for composing text and image to image retrieval», in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 3367–3376 (cit. on p. 17).

[64] J. Zhang, J. Huang, S. Jin, and S. Lu, «Vision-language models for vision tasks: A survey», *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 8, pp. 5625–5644, 2024 (cit. on p. 4).

[65] O. Zhang, M. Wu, J. Bayrooti, and N. Goodman, «Temperature as uncertainty in contrastive learning», *arXiv preprint arXiv:2110.04403*, 2021 (cit. on p. 20).

[66] J. Zhou, L. Dong, Z. Gan, L. Wang, and F. Wei, «Non-contrastive learning meets language-image pre-training», in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 11 028–11 038 (cit. on p. 6).