CROSS THE GAP: INTER-MODAL CLIP REPRESENTATIONS ARE SUPERIOR FOR INTRA-MODAL TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pre-trained multi-modal Vision Language Models like CLIP are widely used off-the-shelf for a variety of applications. Previous work has shown that, due to contrastive pre-training, there is a modality gap between the text and image feature embedding spaces. In this paper, we show that the common practice of individually exploiting the text or image encoders of these powerful multimodal models is highly suboptimal for intra-modal tasks like image-to-image retrieval. We argue that this is inherently due to the inter-modal contrastive loss commonly used to train CLIP models. To demonstrate this, we leverage two optimization-based modality inversion techniques and the inductive bias of the pre-trained encoder of the complementary modality to transform native modality inputs into inter-modal representations. We empirically show in multiple settings (image retrieval, text retrieval, and zero-shot image classification), and at the single-feature level – i.e. each individual feature embedding is mapped to its complementary modality without any need for auxiliary data or additional trained adapters – that approaching these tasks *inter-modally* significantly improves performance with respect to intra-modal baselines on more than fifteen datasets.

1 INTRODUCTION

In recent years the availability of massive, pre-trained Vision-Language Models (VLMs) has enabled a wide variety of applications ranging from zero-shot image segmentation (Zhou et al., 2022a; Lüddecke & Ecker, 2022) to visual question answering (Song et al., 2022; Parelli et al., 2023). These models are typically composed of independent image and text encoders which are simultaneously trained on massive corpora of image-text pairs to align the text and image embeddings of associated inputs. For example, the Contrastive Language-Image Pre-training (CLIP) model is trained on a corpus of 400M image-text pairs to map inputs from both modalities into a shared embedding space (Radford et al., 2021). CLIP is trained with an inter-modal contrastive loss that aims to maximize the similarity of corresponding image-text samples while minimizing the similarity with all the other examples within a batch.

Despite CLIP's shared embedding space, visual and textual features lie in distinct regions. This phenomenon, known as the *modality gap* (Liang et al., 2022), originates from model initialization, and during training the inter-modal contrastive loss preserves and worsens it. Moreover, we note that the CLIP contrastive training strategy focuses on *inter-modal* (i.e. image-text) similarities between paired samples and disregards *intra-modal* (i.e. image-image and text-text) similarities. Consequently, the intra-image and intra-text similarities between CLIP representations might not faithfully correspond to those of the actual images or texts, as depicted in the left section of Fig. 1 and quantified in Sec. 2. We refer to this issue as *intra-modality misalignment*.

Aspects of this misalignment have been accounted for in the limited scope of zero- and few-shot image classification (Udandarao et al., 2023; Yi et al., 2024). However, many recent works overlook this phenomenon and employ CLIP representations for intra-modal similarity comparisons, thus leading to suboptimal similarity measurements. Examples range from KNN classification (Geirhos et al., 2024) to text-to-image generation (Gal et al., 2022; Ruiz et al., 2023) and video synthesis (Esser et al., 2023; Zhang et al., 2024).

In this paper we argue that relying on intra-modal similarities computed using pre-trained CLIP encoders is inherently suboptimal. To support this we conduct an extensive analysis of the behavior

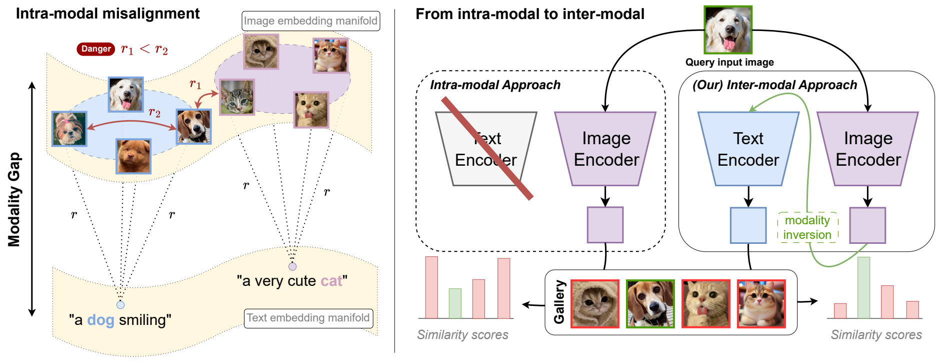


Figure 1: Motivation and overview. **Left:** CLIP is often used off-the-shelf for intra-modal tasks. *Inter-modal* contrastive loss used in pretraining enforces paired images and texts to be at a given distance r but does not encourage *intra-modal* alignment. Consequently, intra-modal similarity scores, might not correspond to those of actual images and texts (*i.e.* $r_1 < r_2$). **Right:** We show that approaching intra-modal tasks (*e.g.* image-to-image retrieval) inter-modally via modality inversion improves the performance.

of intra-modal similarities on the intra-modal tasks of image-to-image and text-to-text retrieval. We contrast this analysis by transforming intra-modal tasks into inter-modal ones to leverage CLIP's inter-modal alignment. Specifically, we map features from their *native* modality (*i.e.* the same as the input) into their *complementary* one. We refer to this process as *modality inversion*. To perform modality inversion we adapt Optimization-based Textual Inversion (OTI) (Baldrati et al., 2023) and introduce Optimization-based Visual Inversion (OVI). OTI and OVI are iterative modality inversion strategies that map image features into text features and vice versa. These techniques operate at the single-feature level, *i.e.* they do not require external data nor the training of a mapping network.

Our experimental results show that tackling intra-modal tasks inter-modally via modality inversion – as illustrated in the right side of Fig. 1 – outperforms intra-modal baselines on more than fifteen datasets. To additionally support our claim that this performance improvement stems from inter-modal alignment and not the modality inversion process itself, we transform inter-modal tasks into intra-modal ones. Specifically, we show that applying modality inversion to the inherently inter-modal zero-shot image classification task yields *worse* performance than the inter-modal baseline. Moreover, we investigate whether the inclusion of an intra-modal loss during image-text contrastive pre-training reduces intra-modal misalignment. For this analysis we use SLIP (Mu et al., 2022), which employs just such an intra-modal loss to improve the alignment within the image embedding space. Results confirm that adding intra-modal loss terms during the pre-training of VLMs significantly mitigates intra-modal misalignment. Finally, we study the relation between the modality gap phenomenon and the intra-modal misalignment. In particular, similar to Liang et al. (2022) we fine-tune CLIP to reduce the modality gap and we observe a decrease in the performance of approaching intra-modal tasks inter-modally. This indicates that a narrower modality gap diminishes the impact of intra-modal misalignment.

The main contributions of this work are:

- we conduct a thorough and comprehensive study of CLIP's intra-modal misalignment. We find that the common practice of relying on intra-modal similarities computed through pre-trained CLIP encoders is inherently suboptimal;
- we propose to transform intra-modal tasks to inter-modal ones via modality inversion to exploit CLIP's inter-modal alignment. To this end we introduce OVI, a single-feature level modality inversion strategy that maps textual features into the image embedding space;
- we conduct extensive experiments that show that approaching intra-modal tasks inter-modally significantly outperforms intra-modal baselines on more than fifteen datasets; and
- we demonstrate that adding intra-modal loss terms during VLM pre-training mitigates the impact of intra-modal misalignment. Moreover, we show that reducing the modality gap also alleviates intra-modal misalignment.

2 QUANTITATIVE INSIGHTS ON INTRA-MODAL MISALIGNMENT

To provide quantitative insights into the intra-modal misalignment issue we conduct a simple experiment using the CLIP ViT-B/32 model and the "Dogs vs Cats" dataset (Elson et al., 2007). This dataset consists of 25K images evenly distributed between two classes: dog and cat. Our goal is to demonstrate that, despite inter-modal alignment, the intra-modal similarity scores are misaligned, *i.e*. they might not reflect those of actual images and texts, as illustrated in the left section of Fig. 1.

We start by filtering out images with incorrect inter-modal alignment to class-specific prompts. Specifically, we remove dog images that exhibit higher similarity to the prompt "a photo of a cat" than to the prompt "a photo of a dog". Then we use the dog-related prompt to query the gallery of all images and filter out the minimal number of images that are incorrectly ranked for this query. We repeat the same procedure for cat images. This filtering ensures perfect inter-modal alignment and text-image retrieval scores. On the resulting filtered dataset, we perform image-to-image retrieval using dog images as queries and the whole set of images as the gallery. If inter-modal alignment guarantees intra-modal alignment, all dog images should rank higher than cat images for any dog query, resulting in perfect retrieval. However, our results contradict this assumption. Specifically, we observe a mean Average Precision (mAP) of 81.4% and an average R-Precision of 71.5%, where R-Precision represents the precision at rank R, with R being the total number of relevant items for a given query. These findings indicate that on average at least 28.5% of dog images are ranked below cat images for a given dog query. Figure 2 qualitatively illustrates this issue, revealing significant overlap between the distributions of pairwise dog-dog and dog-cat image similarities. We observe similar results when employing cat images as queries. Given the evidence of intra-modal misalignment in such a toy dataset, we believe that the issue is likely to be even more pronounced in more complex datasets with more classes.

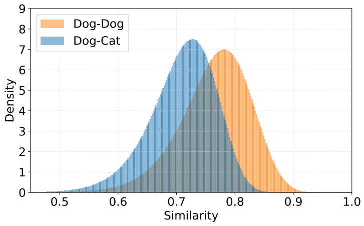


Figure 2: Distribution of pairwise dog-dog and dog-cat image similarities. The significant overlap highlights the intra-modal misalignment issue.

3 RELATED WORK

Contrastively trained Vision-Language Models. VLMs have become increasingly popular for their ability to learn aligned representations across visual and textual modalities (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2022; 2023; Mu et al., 2022; Li et al., 2021). This alignment enables VLMs to be used in a broad variety of downstream tasks, including image-text retrieval and zero-shot image classification, by projecting images and text into a shared feature space. The most prominent example is the Contrastive Language-Image Pretraining (CLIP) model (Radford et al., 2021), which maximizes the similarity between paired images and text captions while minimizing the similarity with the other samples in the batch. SigLIP, on the other hand, employs a sigmoid-based contrastive loss instead of relying on the softmax, thus considering only the single image-text pairs and neglecting the other samples in the same batch (Zhai et al., 2023). More recently, several approaches have extended the CLIP-style contrastive loss by incorporating intra-modal similarities into the training objectives (Mu et al., 2022; Li et al., 2021). For instance, SLIP (Mu et al., 2022) integrates a self-supervised component that maximizes the similarity between different augmentations of the same image, following a strategy akin to SimCLR (Chen et al., 2020).

The modality gap in multi-modal models. Liang et al. (2022) demonstrated a consistent phenomenon affecting VLMs known as the *modality gap*. This refers to the geometric separation between feature embeddings of different modalities (*e.g*. text and images) within their shared representation space (Liang et al., 2022). The modality gap arises due to both model initialization and the contrastive learning objective used during training. At initialization, independent encoders for each modality produce embeddings that are restricted to distinct regions (or cones) within the representation space. During training, the contrastive learning process preserves and worsens this separation. Several works have studied the causes and implications of the modality gap in CLIP (Shi et al., 2023; Schrodi et al., 2024; Zhang et al., 2023). Schrodi et al. (2024) analyzed the embedding space and demonstrated that a minimal number of embedding dimensions – often as few as two – are sufficient to perfectly separate the image and text modalities.

Intra-modal misalignment. Some studies have investigated the problem of misaligned intra-modal embedding distances within the context of zero- and few-shot image classification (Udandarao et al., 2023; Yi et al., 2024). To address this, Udandarao et al. (2023) proposed mitigating the issue by computing similarities in the image-text space, rather than working exclusively with image embeddings, thereby leveraging the inter-modal nature of the feature representations. Similarly, CODER (Yi et al., 2024) introduced an enhanced image representation technique based on measuring distances between images and their K-Nearest Neighboring texts within CLIP's embedding space.

Our contribution with respect to the state-of-the-art. While these prior works have addressed various aspects of intra-modal and inter-modal relationships within VLMs, their scope remains limited, often focusing on specific tasks, datasets, or narrow perspectives on the modality gap and its effects. None of these studies comprehensively investigate the fundamental nature of the intra-modal versus inter-modal representations across diverse tasks and datasets, nor do they fully explore the potential performance improvements achievable by leveraging inter-modal features for intra-modal problems. The motivation behind our work is to shed light on the phenomenon of intra-modal misalignment, its relationship to the modality gap, and to demonstrate the importance of either ensuring intra-modal alignment during pre-training or deriving inter-modal representations better aligned with the semantics relevant to downstream intra-modal tasks like image and text retrieval.

4 CLIP PRELIMINARIES

CLIP (Contrastive Language-Image Pre-training) is a vision-language model trained to align images and textual captions in a shared embedding space (Radford et al., 2021). It consists of an image encoder f_θ and a text encoder g_ϕ. Given an image I, the image encoder extracts its feature representation $f_\theta(I) \in \mathbb{R}^d$, where d is the size of the shared embedding space. Likewise, for a given textual caption Y, first a word embedding layer E_v maps each tokenized word to the token embedding space $\mathcal{V}$. Then, the text encoder g_ϕ generates the textual feature representation $g_\phi(E_v(Y)) \in \mathbb{R}^d$.

When using a Vision Transformer (ViT) (Dosovitskiy et al., 2020) as the visual encoder f_θ, the encoding process begins by splitting the image into U fixed-size non-overlapping patches. Each patch is then transformed into a corresponding patch embedding $\{w_1, w_2, \dots, w_U\}$ through a linear projection by the patch embedding layer E_w, where each w_i resides in the patch embedding space $\mathcal{W}$. A learnable class (CLS) token c is concatenated with the patch embeddings, resulting in the input to the vision transformer being $\bar{I} = \{c, w_1, w_2, \dots, w_U\}$. Finally, the CLS token of the final transformer layer is projected into the shared embedding space via a linear projection to obtain the final representation $f_\theta([c, E_w(I)]) = f_\theta(\bar{I}) \in \mathbb{R}^d$. For brevity, when unnecessary we will omit both the patch embedding layer E_w and the token embedding layer E_v, and use the simplified notations $f_\theta(I)$ instead of $f_\theta([c, E_w(I)])$ and $g_\phi(Y)$ instead of $g_\phi(E_v(Y))$.

Given a batch of image-caption pairs $B = \{(I_n, Y_n)\}_{n=1}^N$, CLIP aims to maximize the cosine similarity for the N correct pairs while minimizing it for the $N^2 - N$ other pairs. This is achieved by optimizing a symmetric, multi-class N-pair contrastive loss (Sohn, 2016). Let $\psi_I^n = f_\theta(I_n)$ and $\psi_T^n = g_\phi(E_v(Y_n))$ denote the image and text embeddings, respectively. The CLIP loss is:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{N} \sum_{n=1}^{N} \left(\log \frac{\exp(c(\psi_I^n, \psi_T^n)/\tau)}{\sum_{m=1}^{N} \exp(c(\psi_I^n, \psi_T^m)/\tau)} + \log \frac{\exp(c(\psi_T^n, \psi_I^n)/\tau)}{\sum_{m=1}^{N} \exp(c(\psi_T^n, \psi_I^m)/\tau)} \right) \tag{1}$$

where $c(\cdot, \cdot)$ denotes the cosine similarity function, and τ is a learnable temperature parameter. As shown by Liang et al. (2022), Eq. (1) leads to a measurable separation between embeddings of the different modalities, creating what is known as the *modality gap*. This gap is significantly affected by the temperature τ, with a larger gap occurring as the temperature decreases.

Note that the CLIP training loss focuses exclusively on inter-modal similarities between paired samples while neglecting intra-modal similarities. For example, consider an image feature anchor ψ_I and two distinct text features ψ_T^1 and ψ_T^2 that are both a distance r from the image feature, such that $c(\psi_I, \psi_T^1) = c(\psi_I, \psi_T^2) = r$. In this case, the text embeddings lie within a hypersphere of radius r centered at ψ_I. The absence of intra-modal constraints means the alignment between ψ_T^1 and ψ_T^2 remains uncalibrated; thus, we have $0 \leq c(\psi_T^1, \psi_T^2) \leq 2r$. This indicates that, while both text features are equidistant from the image feature, their intra-modal similarity is not constrained in any way, leading to potential inconsistencies and intra-modal misalignment. We argue that CLIP's inter-modal contrastive loss directly leads to inferior performance of intra-modal similarity comparisons and thus must either be mitigated via additional intra-modal losses during pre-training or must be compensated for by deriving inter-modal representations for use in intra-modal tasks.

5 INTER-MODAL REPRESENTATIONS VIA MODALITY INVERSION

Due to the modality gap, images and text features lie in distinct regions in the shared embedding space. Previous work introduced modality inversion techniques to map features from the native modality to the complementary one (Ramesh et al., 2022; Patel et al., 2024; Li et al., 2023).

Our goal is to demonstrate that tackling intra-modal tasks in an inter-modal way outperforms CLIP's intra-modal representations. To this end, we propose to employ a modality inversion strategy to derive representations that exploit both native and complementary modality encoders. Most existing modality inversion techniques rely on external data or the training of a mapping network, making the inversion process dependent on factors beyond CLIP. Therefore, we leverage two modality inversion strategies that operate at a single-feature level, *i.e.* that maps each individual feature to its complementary modality without the need for any external resources.

Specifically, we adapt Optimization-based Textual Inversion (OTI) (Baldrati et al., 2023; Agnolucci et al., 2024) and we introduce Optimization-based Visual Inversion (OVI) to map an image to the text embedding space and vice versa. Both are iterative, optimization-based approaches. The core concept behind OTI and OVI is to learn vectors of trainable parameters that are passed through the encoder of the *complementary* modality to yield features aligned with the representations of the *native* modality encoder. In the following we define OTI and OVI for CLIP, but they can be applied to any VLM that maps images and texts into a shared embedding space.

5.1 OPTIMIZATION-BASED TEXTUAL INVERSION (OTI)

Starting from an image I, OTI involves iteratively optimizing a set of R pseudo-word tokens $v^* = \{v_1^*, v_2^*, \dots, v_R^*\}$, with $v_i^* \in \mathcal{V}$ for $i \in \{1, \dots, R\}$, for a given number of optimization steps S. We refer to v^* as pseudo-word tokens since it belongs to the word-embedding space $\mathcal{V}$ but it is not associated with an existing word. Algorithm 1 in Appendix A shows the pseudo-code of OTI.

The pseudo-word tokens v^* are randomly initialized and concatenated with the template sentence "a photo of" to form $\overline{Y}_{v^*} = [E_v(\text{"a photo of"}), v^*]$ input into the CLIP text encoder g_ϕ to obtain $\psi_T = g_\phi(\overline{Y}_{v^*})$. Then we extract the features of the image I with the CLIP image encoder f_θ, resulting in $\psi_I = f_\theta(I)$. Since we aim to obtain a textual feature representation ψ_T that captures the informative content of I, we minimize the gap between image and text features via a cosine loss:

$$\mathcal{L}_{\text{cos}} = 1 - \cos\left(\psi_I, \psi_T\right) \tag{2}$$

While OTI is adapted from Baldrati et al. (2023) our goal is significantly different. Their work focuses on deriving a single pseudo-word token that captures the informative content of the image I and can interact with existing words to form meaningful sentences (e.g., "a photo of v^* that is running …"). In contrast, we use OTI purely as a mapping technique from visual to textual features. We do not focus on the pseudo-word tokens themselves but aim to obtain an accurate final feature representation that effectively captures the content of the image I. Additionally, the original OTI technique employs a regularization loss that exploits an auxiliary vocabulary that constrains the pseudo-word token to remain in the CLIP token embedding space. However we are not interested in using the learned v^* in different contexts – and more importantly, we aim to avoid influencing the inversion process with external data. For this reason we do not use a regularization loss.

5.2 OPTIMIZATION-BASED VISUAL INVERSION (OVI)

We propose the OVI approach to map text features from the CLIP text embedding space to the visual embedding space. Since OVI learns vectors of trainable parameters in the patch embedding space $\mathcal{W}$, it can be applied only to ViT-based image encoders. Given a sentence Y, we first extract its text features $\psi_T = g_\phi(E_v(Y))$. OVI then optimizes a set of P randomly initialized pseudo-patches in CLIP's patch embedding space $\mathcal{W}$, denoted as $w^* = \{w_1^*, w_2^*, \dots, w_P^*\}$, where each $w_i^* \in \mathcal{W}$.

This optimization is performed for a fixed number of optimization steps S. Similarly to the terminology introduced in Sec. 5.1, we refer to w^* as pseudo-patches since they belong to the patch embedding space $\mathcal{W}$ but are not associated with any existing image. Algorithm 2 in Appendix A illustrates the pseudo-code of the OVI method.

Since the ViT employs learned positional embeddings, the number of input patches U to the image encoder is fixed. Consequently, when $P < U$ directly using w^* as input is impossible. In cases, we repeat the pseudo-patches to match the U by applying nearest-neighbor interpolation to w^*. Specifically, given the pre-trained CLS token c, the input to the ViT is given by:

$$\bar{I}_{w^*} = \{c, \underbrace{w_1^*, w_1^*, \dots, w_1^*}_{H_1 \text{ times}}, \underbrace{w_2^*, w_2^*, \dots, w_2^*}_{H_2 \text{ times}}, \dots, \underbrace{w_P^*, w_P^*, \dots, w_P^*}_{H_P \text{ times}}\}, \tag{3}$$

where $H_1, H_2, \dots, H_P$ represent the number of times each pseudo-patch is repeated, and $H_1 + H_2 + \cdots + H_P = U$. The specific values are determined by the nearest-neighbor interpolation.

Finally, the input $\bar{I}_{w^*}$ is passed through CLIP's image encoder to obtain the features $\psi_I = f_\theta(\bar{I}_{w^*})$. To obtain a visual feature representation ψ_I that captures the informative content of Y, we minimize the gap between the image and text features using the same cosine-based loss in Eq. (2).

5.3 CROSSING THE MODALITY GAP WITH OTI AND OVI

The goal of OTI and OVI is to map features from the native modality into the complementary one. We observe that in cases where the loss $\mathcal{L}_{\text{cos}}$ approaches zero, the complementary features converge to the native ones, thus drifting onto the native modality embedding manifold. This undermines the goal of leveraging the image-text alignment inherent in the CLIP training objective.

For OTI, in our experiments the loss never approaches zero – within a reasonable number of optimization steps – when considering a single pseudo-word token (*i.e.* $R = 1$). We argue that this stems from the strong inductive biases of the frozen encoders and the modality gap, making it challenging for a single pseudo-word token to bridge the distance between image and text representations. Nevertheless, the OTI-inverted features retain the informative content of the corresponding image. As a result, the potential drift related to $\mathcal{L}_{\text{cos}}$ does not pose a significant issue, and inter-modal alignment is preserved. In all experiments we use $R = 1$ unless stated otherwise.

Also for OVI we observe that the loss only approaches zero when the number of pseudo patches P is relatively large. Unlike OTI, we find that for some experiments a single pseudo-patch (*i.e.* $P = 1$) is insufficient for embedding the informative content of the corresponding text. We believe that this discrepancy stems from the inherent differences between images and texts. Specifically, in textual inputs a single word (or pseudo-word token) can significantly alter the meaning of a sentence. For instance, the sentences "a photo of a building" and "a photo of a dog" convey completely different meanings, despite differing by only one word. In contrast, a single (pseudo-)patch has less influence on the overall semantic content of an image. Therefore, while a single pseudo-word token is enough for an effective modality inversion with OTI, more pseudo-patches may be necessary when applying OVI. Consequently, in our experiments, we employ a number of pseudo-patches P ranging from 1 to 4, based on the considered model (see Appendix C for more details). For such values, inter-modal alignment is maintained and the drift does not constitute a significant problem.

6 EXPERIMENTAL RESULTS

Here we report on a broad range of experiments supporting our claims. We show that transforming intra-modal tasks into inter-modal ones consistently improves performance by better aligning with the original CLIP training objective. We first evaluate two intra-modal tasks: image-to-image and text-to-text retrieval, using OTI and OVI to transform them into inter-modal tasks. Next, we evaluate zero-shot image classification, an inter-modal task, and show that modality inversion *reduces* performance by making the task intra-modal. Finally, we investigate OTI and analyze the modality gap phenomenon. In the following, we denote as *inter-modal approaches* those involving inter-modal similarity comparisons, *i.e.* similarity comparisons between features of two different modalities (such as image-text, OTI-image, and OVI-text). Conversely, *intra-modal approaches* refer to methods that employ intra-modal similarity comparisons (such as image-image, text-text, OTI-OTI, and OVI-OVI).

Table 1: Performance (mAP) evaluation on the image-to-image retrieval task. ✓ and ✗ denote inter-modal and intra-modal approaches, respectively. Blue rows indicate the usage of OTI-inverted features, while white rows refer to the intra-modal baseline.

	Backbone	Inter modal	CUB	SOP	$\mathcal{R}$Oxford	$\mathcal{R}$Paris	Cars	Pets	Flowers	Aircraft	DTD	EuroSAT	Food101	SUN397	Caltech	UCF101	ImageNet	*Average*
CLIP	B/32	✗	22.9	34.4	42.6	67.9	24.6	30.5	62.0	**14.5**	28.1	**47.9**	32.3	34.3	77.1	47.1	21.4	39.2
		✓	**24.6**	**35.1**	**43.0**	**70.3**	**28.0**	**37.5**	**62.6**	14.4	**31.9**	47.2	**34.7**	**36.3**	**79.9**	**48.6**	**23.8**	**41.2**
	L/14	✗	43.0	40.8	57.5	76.9	43.3	47.3	84.0	25.8	34.1	**59.0**	53.0	39.1	83.2	60.0	33.1	52.0
		✓	**47.1**	**41.2**	**62.4**	**77.1**	**50.5**	**56.0**	**86.0**	**27.1**	**37.7**	56.3	**55.9**	**43.5**	**87.3**	**62.8**	**38.2**	**55.3**
OPEN	B/32	✗	32.1	43.0	50.8	74.7	46.7	44.1	77.0	19.6	36.9	**56.4**	39.6	36.2	82.3	45.7	24.7	47.3
		✓	**34.3**	**44.5**	**54.4**	**75.8**	**50.5**	**50.5**	**78.0**	**20.1**	**40.9**	54.5	**42.9**	**37.8**	**83.3**	**48.2**	**27.3**	**49.5**
	L/14	✗	56.4	50.7	69.0	83.9	65.4	61.4	91.6	32.5	40.4	**63.8**	61.1	42.2	86.9	62.6	38.8	60.4
		✓	**58.9**	**51.9**	**73.2**	**87.7**	**72.6**	**67.3**	**92.7**	**34.3**	**44.3**	63.1	**65.2**	**45.8**	**89.7**	**64.7**	**42.6**	**63.6**
Sig LIP	B/16	✗	39.4	49.9	50.6	73.9	65.7	56.5	87.0	**37.9**	39.9	52.4	56.3	42.8	87.3	**56.7**	35.9	55.5
		✓	**41.8**	**53.0**	**55.2**	**79.1**	**71.8**	**64.2**	**89.7**	37.6	**43.3**	**52.9**	**59.0**	**43.6**	**88.9**	54.9	**38.8**	**58.3**

To ensure a comprehensive analysis, we experiment using multiple CLIP models with different backbones and pre-training datasets. We also consider SigLIP to demonstrate that our observations are not specific to the CLIP loss but generalize to other inter-modal contrastive losses. Specifically, we use OpenAI CLIP with ViT-B/32 and ViT-L/14 backbones, OpenCLIP pre-trained on the DataComp dataset (Gadre et al., 2024) with the same backbones, and SigLIP-B/16. Due to space limitations, implementation details and description of all datasets used are given in Appendices A and E.

6.1 IMAGE-TO-IMAGE RETRIEVAL

Pre-trained CLIP image encoders are often used to extract features for image-to-image similarity comparisons. Here we perform image-to-image retrieval experiments in order to compare intra-modal features with inter-modal ones derived using our OTI approach described in Sec. 5.1.

Experiment design. The objective is to retrieve images from a gallery that are visually similar to a given query image. We consider a total of 15 datasets commonly employed for image-to-image retrieval and image classification. We begin by extracting the features of the gallery images using a pre-trained CLIP image encoder and then consider two evaluation settings. In the first, which we call *intra-modal*, we directly extract the features of the query image using the same pre-trained CLIP image encoder and retrieve gallery images according to cosine similarity. In the second, we transform the intra-modal image-to-image retrieval task into an inter-modal task by applying OTI to the query image to obtain the corresponding inter-modal features. Then we again retrieve the images from the gallery most similar to the query features using cosine similarity.

Results. In Tab. 1 we report results on image-to-image retrieval on 15 datasets and for five different pre-trained models. Using OTI features outperforms the native image features, highlighting that intra-modal features lead to suboptimal results. Moreover, the performance improvement for OpenCLIP and SigLIP shows that the misaligned representation phenomenon is independent of the pre-training dataset and pre-training contrastive loss, respectively.

6.2 TEXT-TO-TEXT RETRIEVAL

Although text features from pre-trained CLIP models are not commonly used for text-to-text tasks, we believe that it is important to show that our findings also apply to the textual embedding space.

Experiment design. Applying the CLIP text encoder to text-only tasks presents several challenges. Specifically, the CLIP text encoder is trained on short, descriptive texts. As a result, using it for tasks such as sentiment analysis or text classification, which involve longer texts and abstract concepts, results in a mismatch with the pre-training data. Moreover, VLMs like CLIP have a limited input token capacity (*e.g.* 77 tokens for CLIP), which makes using longer texts impractical.

To avoid these problems, we formulate a text-to-text retrieval task using image captioning datasets. Specifically, we select datasets in which each sample consists of an image and multiple associated captions (*e.g.* Flickr30K (Plummer et al., 2015)). These captions are comparable to those used in VLM training and are short enough to avoid token limit issues. We ignore the images and use the first caption associated with each image as the query text. The goal is to retrieve the other captions

Table 2: **Left**: Performance (mAP) evaluation on the text-to-text retrieval task. ✓ and X denote inter-modal and intra-modal approaches, respectively. Purple rows indicate the usage of OVI-inverted features, while white rows refer to the intra-modal baseline. **Right**: Performance (accuracy) evaluation on the zero-shot image classification task. Blue rows indicate the usage of OTI-inverted features, while white rows refer to the inter-modal baseline.

	Backbone	Inter modal	Flickr30k	COCO	nocaps	*Average*
CLIP	B/32	X	51.7	26.2	35.1	37.7
		✓	**54.8**	**28.3**	**38.8**	**40.6**
	L/14	X	52.3	26.7	35.7	38.2
		✓	**54.9**	**29.4**	**39.5**	**41.3**
OPEN	B/32	X	58.0	30.0	40.3	42.8
		✓	**60.2**	**32.0**	**43.6**	**45.3**
	L/14	X	61.0	31.8	42.5	45.1
		✓	**63.6**	**33.0**	**44.5**	**47.0**
Sig LIP	B/16	X	56.7	27.0	38.6	40.8
		✓	**60.1**	**29.6**	**43.4**	**44.4**

	Backbone	Inter modal	Cars	Pets	Flowers	Aircraft	DTD	EuroSAT	Food101	SUN397	Caltech	UCF101	ImageNet	*Average*
CLIP	B/32	✓	**60.4**	**87.5**	**67.0**	**19.1**	**43.6**	**45.2**	**80.5**	**62.0**	**91.2**	**62.0**	**62.1**	**61.9**
		X	54.5	80.9	61.2	17.3	41.8	39.4	75.3	54.6	83.7	58.5	53.6	**56.4**
	L/14	✓	**76.8**	**93.6**	**79.3**	**32.5**	**53.0**	**58.1**	**91.0**	**67.6**	**94.9**	**74.2**	**73.5**	**72.2**
		X	72.1	89.8	73.1	29.4	52.3	56.4	87.6	62.4	90.2	71.3	68.0	**68.4**
OPEN	B/32	✓	**88.4**	**90.3**	**73.5**	**24.4**	**53.9**	**56.5**	**83.0**	**67.0**	**96.2**	**61.6**	**68.6**	**69.4**
		X	86.0	87.6	70.9	23.1	52.8	47.5	80.3	61.5	93.6	59.8	63.9	**66.1**
	L/14	✓	**93.7**	**95.0**	**82.5**	**47.6**	**62.7**	**68.0**	**92.3**	**74.2**	**97.6**	**75.0**	**78.9**	**78.9**
		X	93.0	94.0	82.0	44.9	61.2	66.6	91.8	71.7	91.6	73.1	77.0	**77.0**
Sig LIP	B/16	✓	**90.7**	**94.1**	**85.8**	**43.9**	**62.0**	**42.3**	**89.2**	**69.6**	**97.4**	**74.9**	**75.7**	**75.1**
		X	86.3	90.4	69.5	35.1	58.6	32.5	84.6	55.9	89.5	64.8	62.1	**66.3**

related to the same image from a gallery of all captions in the dataset. As in the image-to-image retrieval experiments above, we consider two evaluation settings. In the first, we use the textual features of the query to retrieve from the gallery of captions. In the second, we apply OVI to each query, transforming the task from intra-modal to an inter-modal retrieval task.

Results. Tab. 2 (left) gives the results on text-to-text retrieval. Similar to image-to-image retrieval, intra-modal representations yield suboptimal performance. By applying OVI to the query text features, we obtain inter-modal features that exploit the inter-modal alignment of the pre-trained CLIP model. With OVI we achieve better performance for all VLMs and backbones.

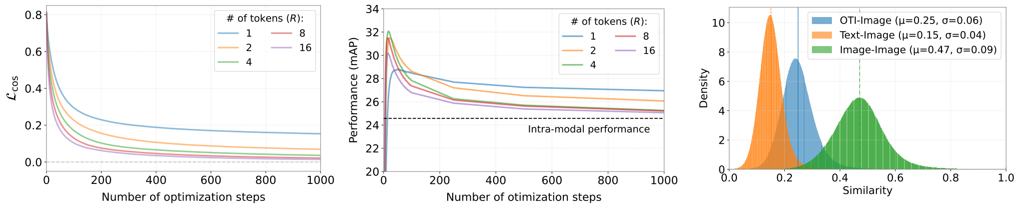
6.3 ZERO-SHOT IMAGE CLASSIFICATION

Experiment design. We evaluate the performance of modality inversion on zero-shot image classification. CLIP-like models can perform this task by predicting the output class based on the similarity between the input image and a set of textual prompts in the form of "*a photo of a [CLASS]*", where CLASS represents each class name, such as "cat" or "dog". Since an image is compared with texts, this task is inherently inter-modal, and we expect that converting it to an intra-modal task by applying a modality inversion technique should *hinder* performance due to intra-modal misalignment. Following Zhou et al. (2022b), we take into account 11 different datasets (see Appendix E for dataset details). We consider three evaluation settings. The first is the standard one, where we use the input image features and the textual features of the set of prompts. In the second, we apply OTI to the input image. In the third, we apply OVI to each textual prompt.

Results. In Tab. 2 (right) we show the results of the first two evaluation settings described above. Results for the third setting are given in Appendix F. As expected, using modality inversion results in performance degradation as we are transforming an inter-modal task into an intra-modal one. Note that the datasets used in zero-shot image classification are the same as those employed for image-to-image retrieval in Sec. 6.1. This allows us to reuse the *same* OTI-inverted features for both tasks. Interestingly, the results are the opposite: performance improves in image-to-image retrieval but decreases in zero-shot image classification. The reason for this is that in the former we are transforming an intra-modal task into an inter-modal one, while in the latter we are doing the opposite. This experiment demonstrates that modality inversion does not inherently improve performance, as the same OTI-inverted features can either enhance or hinder results depending on the nature of the task. Performance improvement is observed only when an intra-modal task is converted into an inter-modal one. To further confirm this claim, we also performed experiments on image-text retrieval and report the results in Appendix F.

6.4 ANALYZING MODALITY INVERSION

In this section we study how and why transforming native modality features into complementary ones via modality inversion leads to performance improvement on intra-modal tasks. For brevity, we consider only OTI, but we find that the same considerations apply to OVI.



(a) (b) (c)

Figure 3: **(a, b)** Loss values and retrieval performance over OTI optimization steps for different numbers of pseudo-word tokens R. **(c)** Distribution of pairwise image-image, text-image, and OTI-image cosine similarities. The similarities related to the OTI-inverted features are closer to those related to texts than images.

In Figs. 3(a) and 3(b) we investigate how the image-to-image retrieval performance and the values of the loss $\mathcal{L}_{\text{cos}}$ vary based on the number of optimization steps and the number of pseudo-word tokens R. We consider the Cars dataset (Krause et al., 2013) and the CLIP ViT-B/32 model.

We notice first of all that with a single pseudo-word token (*i.e.* $R = 1$) the loss does not approach zero within a reasonable number of optimization steps. On the contrary, as R increases (*i.e.* the number of trainable parameters grows) the loss decreases more rapidly and approaches zero. As discussed in Sec. 5.3, as the loss decreases the OTI-inverted features shift away from the text manifold towards the image embedding manifold as they approach the original, native image features. This phenomenon is reflected in the image retrieval performance shown in Fig. 3(b) since for enough optimization steps and pseudo-word tokens the performance approaches those obtained by the native, intra-modal image features. Moreover, we observe that, regardless of the value of R, the best performance corresponds to a relatively low number of optimization steps.

We argue that in proximity to the performance peak observed during OTI optimization, the OTI-inverted features capture the informative content of the corresponding image while still remaining within the text embedding manifold. To support this claim, we compute the pairwise image-image, text-image, and OTI-image cosine similarities for features extracted from the COCO validation set (Lin et al., 2014). For the OTI-inverted features, we consider those obtained using $R = 1$ after 150 optimization steps. In Fig. 3(c) we plot the distribution of these inter- and intra-modal similarities. We observe that the similarities related to the OTI-inverted features are closer to those related to texts than images. This suggests that the OTI-inverted features still lie in the text manifold for $R = 1$ and only 150 optimization steps, confirming our hypothesis that the performance improvement obtained by OTI stems from leveraging CLIP's inter-modal alignment.

Finally, we notice that $R = 1$ is not the optimal choice to achieve the best performance with OTI. Still, we use $R = 1$ in the experiments as the associated OTI-inverted features are less prone to drift towards the native image features, thus being more robust to the number of optimization steps. Moreover, the main goal of this work is *not* to achieve the best results on the downstream tasks but rather show that using VLMs intra-modally is suboptimal. The number of learnable tokens and optimization steps are hyperparameters that could be cross-validated to further improve performance.

6.5 THE ROLE OF INTRA-MODAL CONSTRAINTS

We study whether incorporating an intra-modal loss term during image-text contrastive pre-training effectively mitigates the issue of intra-modality misalignment. To this end, we consider SLIP (Mu et al., 2022), which, in addition to the standard CLIP inter-modal contrastive loss (Eq. (1)), adds a self-supervised intra-modal loss based on SimCLR (Chen et al., 2020) (see Appendix B for more details). This loss encourages the model to yield similar representations for two augmentations of the same image, aiming to improve the intra-modality alignment within the image embedding space.

To analyze this, we performed an image-to-image retrieval experiment following the evaluation protocol from Sec. 6.1. We report these results in Tab. 3. Notably, the complementary representations obtained via OTI achieve comparable performance to native image features. This contrasts with results from VLMs trained solely with an inter-modal contrastive loss (see Tab. 1) in which OTI led to a substantial performance boost. This experiment demonstrates that the intra-modal loss used in

Table 3: Performance (mAP) evaluation on the image-to-image retrieval task using SLIP. ✓ and ✗ denote inter-modal and intra-modal approaches, respectively. Blue rows indicate the usage of OTI-inverted features, while white rows refer to the intra-modal baseline.

	Backbone	Inter modal	CUB	SOP	$\mathcal{R}$Oxford	$\mathcal{R}$Paris	Cars	Pets	Flowers	Aircraft	DTD	EuroSAT	Food101	SUN397	Caltech	UCF101	ImageNet	*Average*
SLIP	B/16	✗	**16.6**	**49.3**	36.2	78.9	4.9	17.8	**65.2**	**9.1**	29.8	**53.0**	19.3	26.2	65.5	40.3	14.5	35.1
		✓	16.2	48.8	**36.4**	**79.3**	**5.0**	**19.3**	65.1	9.0	**30.5**	50.6	**20.0**	**26.4**	**67.6**	**40.6**	**14.8**	**35.3**
	L/16	✗	19.5	**46.4**	36.3	**75.3**	5.3	21.7	69.2	9.7	28.8	**56.5**	24.25	27.4	71.0	41.2	17.4	36.7
		✓	**19.6**	45.8	**38.0**	75.1	**5.5**	**23.3**	**70.2**	**9.8**	**29.7**	53.7	**25.2**	**27.8**	**72.3**	**41.4**	**18.2**	**37.1**

SLIP effectively reduces intra-modal misalignment and suggests the importance of including such a loss when pre-training VLMs if intra-modal similarity comparisons are important downstream.

6.6 THE ROLE OF THE MODALITY GAP

The temperature parameter τ in Eq. (1) critically affects the resulting modality gap: higher temperatures considerably reduce or close it (Liang et al., 2022). To examine whether reducing the modality gap alleviates intra-modal misalignment we fine-tune a CLIP B/32 model on the COCO dataset (Lin et al., 2014) using a temperature $\tau = 1.0$, which closes the modality gap. To provide a reference, we repeat the experiment with $\tau = 0.01$, *i.e.* the value employed during CLIP pre-training. See Tab. A2 for more details about the specific magnitudes of the modality gap for the different models.

Table 4: Impact of the modality gap on the performance (mAP) for the image-to-image retrieval task on image retrieval datasets. ✓ and ✗ denote inter-modal and intra-modal approaches, respectively. Blue rows indicate the usage of OTI-inverted features, while white rows refer to the intra-modal baseline.

Temperature	Inter modal	CUB	SOP	$\mathcal{R}$Oxford	$\mathcal{R}$Paris	Cars	*Average*
$\tau = 1$ *(no gap)*	✗	**15.9**	**23.7**	**29.3**	**46.6**	**19.3**	**27.0**
	✓	14.0	20.4	26.7	43.1	17.4	**24.2**
$\tau = 0.01$	✗	24.0	35.0	43.1	68.6	25.7	39.3
	✓	**24.1**	**35.2**	**44.0**	**70.2**	**27.6**	**40.2**

We reproduce our image-to-image retrieval experiments using these fine-tuned models and report results in Tab. 4. Clearly, in the absence of the modality gap tackling intra-modal tasks inter-modally does not improve performance. The results of the reference model demonstrate that this outcome does not stem from the fine-tuning strategy.

Therefore, this experiment shows that closing the modality gap mitigates the intra-modality misalignment. However, as also observed by Liang et al. (2022), we note that using higher temperature values during training leads to an overall performance decrease in downstream tasks, despite reducing the modality gap. For this reason, we argue that – in practice – simply increasing the temperature value in Eq. (1) does not represent a viable strategy to address intra-modal misalignment.

7 CONCLUSIONS

In this work we show that relying solely on intra-modal similarity comparisons of features extracted using contrastively-trained VLMs hinders performance on intra-modal tasks like image-to-image and text-to-text retrieval. The inter-modal contrastive loss frequently employed for pre-training these models leads to a gap between the image and text modalities and misaligned intra-modal representations. We demonstrate that transforming native modality inputs to the complementary modality through modality inversion improves performance since these representations exploit inter-modal alignment. In addition, we show that employing an intra-modal loss or reducing the modality gap alleviates the intra-modal misalignment induced by the CLIP contrastive loss. Our analyses and experimental results demonstrate that exploiting the inter-modal alignment of off-the-shelf VLMs improves performance even on intra-modal tasks.

Limitations. Our analyses demonstrate the significance of intra-modal misalignment when exploiting pre-trained CLIP models, but fall short of offering practical alternatives. The modality inversion techniques we propose to derive inter-modal representations are computationally expensive. They are based on iterative optimization of learnable input parameters (150 optimization steps for OTI and 1000 for OVI in our experiments). This limits their practical applicability and future work should concentrate on efficient methods for deriving inter-modal features from pre-trained VLMs.

REPRODUCIBILITY STATEMENT

We have taken steps in this work to ensure the reproducibility of our results. All code, models, and datasets used in our experiments are publicly available, and we will release the complete source code after the reviewing period. In the main paper and appendices material we provide complete details of all experimental setups, including model architectures, training and evaluation protocols, and hyperparameters. All random seeds are fixed in our experiments, ensuring that others can replicate our results with the provided code. Finally, our work only relies on publicly accessible datasets, and we include clear references for any dataset-specific processing. We believe that the measures we have taken to ensure reproducibility will facilitate straightforward replication and verification of our findings, as well as allow the community to build upon our results in the future.

REFERENCES

Lorenzo Agnolucci, Alberto Baldrati, Marco Bertini, and Alberto Del Bimbo. iSEARLE: Improving Textual Inversion for Zero-Shot Composed Image Retrieval. *arXiv preprint arXiv:2405.02951*, 2024. 5

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8948–8957, 2019. 17

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-Shot Composed Image Retrieval with Textual Inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15338–15347, 2023. 2, 5

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pp. 446–461. Springer, 2014. 17

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020. 3, 9, 16

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014. 17

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009. 17

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 4

Jeremy Elson, John R Douceur, Jon Howell, and Jared Saul. Asirra: a captcha that exploits interest-aligned manual image categorization. *CCS*, 7:366–374, 2007. 3

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7346–7356, 2023. 1

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004. 17

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 7

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1

Robert Geirhos, Priyank Jaini, Austin Stone, Sourabh Medapati, Xi Yi, George Toderici, Abhijit Ogale, and Jonathon Shlens. Towards flexible perception with visual memory. *arXiv preprint arXiv:2408.08172*, 2024. 1

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 17

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021. 3

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015. 17

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013. 9, 17

Ken Lang. Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995*, pp. 331–339. Elsevier, 1995. 19

Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding CLIP latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*, 2023. 5, 20

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 3

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the Gap: Understanding the Modality Gap in Multi-Modal Contrastive Representation Learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 1, 2, 3, 4, 10

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014. 9, 10, 17, 20

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 19

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7086–7096, 2022. 1

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011. 19

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 17

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pp. 529–544. Springer, 2022. 2, 3, 9, 15

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008. 17

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4004–4012, 2016. 17

Maria Parelli, Alexandros Delitzas, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. Clip-guided vision-language pre-training for question answering in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5607–5612, 2023. 1

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012. 17

Maitreya Patel, Changhoon Kim, Sheng Cheng, Chitta Baral, and Yezhou Yang. Eclipse: A resource-efficient text-to-image prior for image generations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9069–9078, 2024. 5, 19

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015. 7, 16, 17

Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5706–5715, 2018. 17, 20

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021. 1, 3, 4, 17

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 5

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023. 1

Simon Schrodi, David T Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Representation Learning. *arXiv preprint arXiv:2404.07983*, 2024. 3

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 20

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018. 19

Peiyang Shi, Michael C Welle, Mårten Björkman, and Danica Kragic. Towards Understanding the Modality Gap in CLIP. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023. 3

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016. 4

Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*, 2022. 1

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 17

Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. SuS-X: Training-Free Name-Only Transfer of Vision-Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2725–2736, 2023. 1, 4

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 17

Chao Yi, Lu Ren, De-Chuan Zhan, and Han-Jia Ye. Leveraging Cross-Modal Neighbor Representation for Improved CLIP Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27402–27411, 2024. 1, 4

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*. 20

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18123–18133, 2022. 3

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023. 3, 15

Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and Rectifying Vision Models using Language. In *International Conference on Learning Representations (ICLR)*, 2023. 3

Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yinan Zhao, Peter Vajda, Dimitris Metaxas, and Licheng Yu. Avid: Any-length video inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7162–7172, 2024. 1

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pp. 696–712. Springer, 2022a. 1

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b. 8, 17

Algorithm 1 Optimization-based Textual Inversion (OTI)

```
Input: Image $I$, number of pseudo-word tokens $R$, number of optimization steps $S$
Initialize $v^* = \{v_1^*, v_2^*, \dots, v_R^*\}$
Extract image features: $\psi_I = f_\theta(I)$
for $s = 1$ to $S$ do
    Form $\overline{Y}_{v^*} = [E_v(\text{"a photo of"}), v^*]$
    Extract text features: $\psi_T = g_\phi(\overline{Y}_{v^*})$
    Compute loss: $\mathcal{L}_{\text{cos}} = 1 - \cos(\psi_I, \psi_T)$
    Update $v^*$ to minimize $\mathcal{L}_{\text{cos}}$
end for
Output: OTI-inverted text features $\psi_T = g_\phi(\overline{Y}_{v^*})$
```

Algorithm 2 Optimization-based Visual Inversion (OVI)

```
Input: Text $Y$, number of pseudo-patches $P$, number of optimization steps $S$
Initialize $w^* = \{w_1^*, w_2^*, \dots, w_P^*\}$
Extract text features: $\psi_T = g_\phi(E_v(Y))$
for $s = 1$ to $S$ do
    Form input $\bar{I}_{w^*}$ using Eq. (3)
    Extract image features: $\psi_I = f_\theta(\bar{I}_{w^*})$
    Compute loss: $\mathcal{L}_{\text{cos}} = 1 - \cos(\psi_I, \psi_T)$
    Update $w^*$ to minimize $\mathcal{L}_{\text{cos}}$
end for
Output: OVI-inverted image features $\psi_I = f_\theta(\bar{I}_{w^*})$
```

Algorithms 1 and 2. **Left**: OTI maps an image into the textual embedding space by optimizing pseudo-word tokens. **Right**: OVI maps a text into the visual embedding space by optimizing pseudo-patches. Both approaches iteratively minimize the cosine distance between the feature representations of the native and complementary modality.

APPENDIX A IMPLEMENTATION DETAILS

We give the pseudo-code of Optimization-based Textual Inversion (OTI) and Optimization-based Visual Inversion (OVI) in Algorithm 1 and Algorithm 2, respectively. We use the same hyperparameters for both OTI and OVI unless stated otherwise. We employ the AdamW optimizer with learning rate equal to 0.02, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 0.01. The results presented in the main paper are evaluated at step 150 for OTI and step 1000 for OVI. The computational cost of OTI and OVI is linear with respect to the number of queries and the inversion processes can be parallelized. On average, OTI takes approximately 0.2 seconds per image, while OVI takes around 0.5 seconds per text prompt on a single A100 GPU (40GBs) with a batch size of 2048 and OpenAI ViT/B-32 as the backbone. The memory usage scales linearly with the batch size. Specifically, when using the CLIP ViT-B/32 model, OTI requires approximately 1,878 MiB plus 18.6 MiB per sample in the batch. For example, with a batch size of 128, the memory consumption is about 4,260 MiB. For OVI, the memory usage is approximately 2,218 MiB plus 16.2 MiB per sample, resulting in about 4,290 MiB with the same batch size.

APPENDIX B ADDITIONAL VLMS

In this section we provide a more detailed explanation of the SigLIP and SLIP models, and we highlight the main differences between these models and CLIP.

SigLIP. In SigLIP (Zhai et al., 2023), given a batch of image-caption pairs $B = \{(I_i, Y_i)\}_{i=1}^N$, training maximizes the cosine similarity for the N correct pairs and minimizes it for the $N^2 - N$ incorrect pairs. Unlike the softmax-based contrastive loss from Eq. (1) used in CLIP, SigLIP employs a sigmoid-based loss that avoids global normalization factors. Each image-text pair is processed independently, transforming the learning task into a binary classification problem across all pair combinations. The matching pair (I_i, Y_i) receives a positive label, while all other pairs $(I_i, Y_{j \neq i})$ receive negative labels. SigLIP consists of an image encoder f_θ and a text encoder g_ϕ. We denote the image and text embeddings as $\psi_I^i = f_\theta(I_i)$ and $\psi_T^i = g_\phi(Y_i)$, respectively. The SigLIP loss is:

$$\mathcal{L}_{\text{SigLIP}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \log \left(\frac{1}{1 + e^{z_{ij}(-c(\psi_I^i, \psi_T^j)/\tau + b)}} \right), \tag{4}$$

where $c(\cdot, \cdot)$ denotes the cosine similarity, τ is a learnable temperature parameter, b is a learnable bias, and z_{ij} is the label for a given image and text input ($z_{ij} = 1$ if $i = j$ and $z_{ij} = -1$ otherwise). Similar to CLIP, the SigLIP loss does not include explicit intra-modal constraints; the loss focuses solely on inter-modal alignment between image and text embeddings, without directly enforcing intra-modal alignment.

SLIP. SLIP (Mu et al., 2022) is a model trained with both language supervision and image self-supervision. It is trained with a loss consisting of two components: the first is the same loss used in

CLIP (Eq. (1)), while the second term is a self-supervised learning (SSL) term that forces the model to represent different views or augmentations of the same image similarly. In particular, for the self-supervised loss component, SLIP adopts an adaptation of SimCLR (Chen et al., 2020). At each training step, CLIP and SSL losses are computed on the relevant embeddings and then summed together. The self-supervised component of the loss, $\mathcal{L}_{\text{SimCLR}}$, aims to maximize the agreement between two augmented views of the same image:

$$\mathcal{L}_{\text{SimCLR}} = -\frac{1}{2N}\sum_{i=1}^{2N} \log \frac{\exp\left(\mathrm{c}(\psi_I^i, \psi_I^{p(i)}/\tau\right)}{\sum_{k=1,k\neq i}^{2N} \exp\left(\mathrm{c}(\psi_I^i, \psi_I^k)/\tau\right)}, \tag{5}$$

where ψ_I^i is the embedding of sample i, $p(i)$ is the other augmented view of the same image, $\mathrm{c}(\psi_I^i, \psi_I^k)$ represents the cosine similarity between ψ_I^i and ψ_I^k, and τ is a temperature parameter. The final loss used in SLIP is a combination of CLIP and self-supervised losses:

$$\mathcal{L}_{\text{SLIP}} = \mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{SimCLR}}. \tag{6}$$

By incorporating the self-supervised loss, SLIP encourages better intra-modal alignment within the image embedding space. This intra-modal constraint aims to mitigate some of the limitations seen in models like CLIP, which solely rely on inter-modal contrastive loss. We confirm this empirically in Tab. 3.

Appendix C Selecting the number of pseudo-patches for OVI

As mentioned in Sec. 5.3 our initial experiments showed that in OVI learning a single pseudo-patch ($P = 1$) often failed to adequately minimize the loss. This leads to a poor representation of the input caption. To determine the optimal number of pseudo-patches for each VLM, we conduct a text-to-text retrieval experiment using the Flickr30K (Plummer et al., 2015) validation set, varying the number of pseudo-patches P from 1 to 16.

Table A1 presents the results of this ablation study. We observe that the ideal number of pseudo-patches changes depending on the backbone. In particular, larger models – with a greater number of input patches U – tend to require more pseudo-patches. We hypothesize that this is because, as the number of patches increases, the influence of a single (pseudo-)patch decreases, necessitating a larger number of pseudo-patches to capture sufficient information. Note that in certain cases where different numbers of pseudo-patches lead to similar performance (e.g., SLIP B/16), we choose the smallest number of patches for the experiments.

Table A1: Ablation on the number of OVI pseudo-patches for text-to-text retrieval on the Flickr30K validation set. The highest mAP score in each row is highlighted in bold.

			Number of Pseudo-Patches P				
VLM	Backbone	Intra-modal	1	2	4	8	16
CLIP	B/32	51.4	**54.5**	52.9	51.8	51.6	51.6
	L/14	52.6	51.7	55.2	**56.0**	55.3	54.1
OPEN	B/32	57.3	**59.6**	57.9	57.5	57.4	57.4
	L/14	59.6	60.6	**62.5**	62.4	61.2	60.4
SigLIP	B/16	56.3	45.2	58.0	**60.1**	59.9	59.4
SLIP	B/16	45.8	**46.4**	46.4	46.1	45.9	45.9
	L/16	49.8	48.9	**50.0**	49.8	49.9	49.8

Appendix D Different VLM, Different Modality Gap

In the main paper we show how intra-modal misalignment arises from the contrastive inter-modal pre-training of CLIP-like VLMs. We also demonstrate that models introducing an intra-modal training constraint (*e.g.* SLIP) can mitigate this issue. Additionally, we highlight how this is inherently inter-connected with the modality gap, an expression of the separation between the feature distributions of the different modalities in the shared embedding space.

To facilitate a clearer comparison of modality gaps across the different VLMs, in Table A2 we report the magnitude of the modality gaps evaluated on the COCO validation split. The modality gap is defined as the difference between the two centroids of the image and text modality embeddings:

$$\Delta_{\text{gap}} = \frac{1}{N}\sum_{i=1}^{N} x_i - \frac{1}{N}\sum_{i=1}^{N} y_i, \tag{7}$$

where x_i and y_i are the L2-normalized image and text embeddings, respectively, and N is the number of image-text pairs.

In addition to CLIP, OpenCLIP, SigLIP, and SLIP, we fine-tuned only the final projections layers of two OpenAI B/32 models using AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ and weight decay of 0.2, learning rate of 1e-6 and batch size of 512 for more than 30k training steps and with softmax temperatures $\tau = 0.01$ and $\tau = 1$ on the COCO training set (Lin et al., 2014). We refer to these two fine-tuned CLIP models as *Cold CLIP* and *Hot CLIP*, respectively, in relation to their fine-tuning temperatures.

Table A2: $\|\Delta_{\text{gap}}\|$ for different VLMs on COCO.

VLM	Backbone	Loss	$\|\Delta_{\text{gap}}\|$
CLIP	B/32	$\mathcal{L}_{\text{CLIP}}$	0.82
	L/14		0.82
OPEN	B/32	$\mathcal{L}_{\text{CLIP}}$	0.82
	L/14		0.80
SigLIP	B/16	$\mathcal{L}_{\text{SigLIP}}$	**1.05**
SLIP	B16	$\mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{SimCLR}}$	0.57
	L/16		0.49
Hot/Cold CLIP	B/32	$\mathcal{L}_{\text{CLIP}}(\tau = 1)$	**0.007**
		$\mathcal{L}_{\text{CLIP}}(\tau = 0.01)$	0.88

These results show that models trained with an additional intra-modal constraint (*i.e.* SLIP), or fine-tuned with a higher temperature (*i.e.* our B/32 we fine-tuned on COCO with temperature $\tau = 1$), significantly reduce the modality gap. Notably there seems to be a correlation between the magnitude of the modality gap and the improvement in approaching intra-modal tasks inter-modally using OTI (or OVI).

APPENDIX E DATASET DETAILS

Our experimental evaluation is performed on 18 datasets. Here we report all the evaluated splits and details of the datasets used in our experiments.

Zero-shot Image Classification. Following Zhou et al. (2022b), we validate our zero-shot image classification experiments on 11 publicly available datasets with diverse characteristics: ImageNet (Deng et al., 2009) for large-scale object classification; Caltech101 (Fei-Fei et al., 2004) for general object classification; EuroSAT (Helber et al., 2019) for satellite image recognition; Food101 (Bossard et al., 2014), FGVCAircraft (Maji et al., 2013), OxfordPets (Parkhi et al., 2012), Flowers102 (Nilsback & Zisserman, 2008), and StanfordCars (Krause et al., 2013) for fine-grained classification; UCF101 (Soomro et al., 2012) for action recognition; and the Describable Textures Dataset (DTD) (Cimpoi et al., 2014) for texture classification. Following Zhou et al. (2022b), we discard the “BACKGROUND Google” and “Faces easy” classes from Caltech101. For UCF101, a video dataset, we follow Radford et al. (2021) and use the middle frame of each video clip as the input image. In all classification experiments, we report the accuracy results on the test set.

Image-to-Image Retrieval. For Image-to-Image retrieval we use the 11 datasets used for zero-shot image classification (*i.e.* using the test set as the query set and the training set as the gallery) and four widely used datasets commonly used for metric learning and image retrieval: CUB-200-2011 (CUB) (Wah et al., 2011), Stanford Online Products (SOP) (Oh Song et al., 2016), $\mathcal{R}$Oxford (Radenović et al., 2018), and $\mathcal{R}$Paris (Radenović et al., 2018), for a total of 15 datasets. For CUB we use the entire dataset as both the query and gallery sets. In SOP, we use the test set for both query and gallery sets. In all experiments involving $\mathcal{R}$Oxford and $\mathcal{R}$Paris, following the standard benchmark we include the $\mathcal{R}$1M distractor set, containing 1 million images, as negative samples for all the queries. For brevity in the paper we report only the metric calculated on the Easy setting. For image-to-image retrieval evaluation, we use the standard mean Average Precision (mAP) metric. Note that reusing the 11 classification datasets allows us to evaluate and compare the same OTI features in both the classification and retrieval tasks.

Text-to-Text Retrieval. We perform our text-to-text retrieval experiments using three image-caption datasets: COCO (Lin et al., 2014), Flickr30K (Plummer et al., 2015), and NoCaps (Agrawal et al., 2019). We selected these datasets for two reasons: they contain short, descriptive text similar to the ones used to train VLMs, and they provide multiple captions for each image. In our evaluation, we use the first caption of each image as the query and aim to retrieve the other captions associated with the same image from a gallery of all captions in the dataset. On average, COCO and Flickr30K images have 5 captions each, while NoCaps images have 10. We use the Karpathy split (Karpathy & Fei-Fei, 2015) for both COCO and Flickr30K and report results using captions from the test split. For NoCaps, we report results on the validation split. Although these datasets contain images associated with captions, we ignore the images in this setting. For a fair comparison we report mean Average Precision (mAP) for all the Text-to-Text retrieval datasets.

Table A3: Performance (accuracy) evaluation on the zero-shot image classification task. ✓ and X denote inter-modal and intra-modal approaches, respectively. Purple rows indicate the usage of OVI-inverted features, while white rows refer to the inter-modal baseline.

	Backbone	Inter modal	Cars	Pets	Flowers	Aircraft	DTD	EuroSAT	Food101	SUN397	Caltech	UCF101	ImageNet	*Average*
CLIP	B/32	✓	**60.4**	**87.5**	**67.0**	**19.1**	**43.6**	**45.2**	**80.5**	**62.0**	**91.2**	**62.0**	**62.1**	**61.9**
		X	37.4	59.9	35.0	9.2	26.2	18.9	65.1	44.1	83.9	51.2	42.9	**43.1**
	L/14	✓	**76.8**	**93.6**	**79.3**	**32.5**	**53.0**	**58.1**	**91.0**	**67.6**	**94.9**	**74.2**	**73.5**	**72.2**
		X	46.9	71.1	65.1	23.3	41.4	23.8	73.6	46.2	41.6	63.5	54.8	**50.1**
OPEN	B/32	✓	**88.4**	**90.3**	**73.5**	**24.4**	**53.9**	**56.5**	**83.0**	**67.0**	**96.2**	**61.6**	**68.6**	**69.4**
		X	81.4	82.1	62.4	17.9	45.8	36.6	76.1	56.9	93.6	55.1	59.6	**60.7**
	L/14	✓	**93.7**	**95.0**	**82.5**	**47.6**	**62.7**	**68.0**	**92.3**	**74.2**	**97.6**	**75.0**	**78.9**	**78.9**
		X	78.6	85.3	71.1	35.9	48.6	47.9	86.2	50.7	92.9	62.4	67.3	**66.1**
Sig LIP	B/16	✓	**90.7**	**94.1**	**85.8**	**43.9**	**62.0**	**42.3**	**89.2**	**69.6**	**97.4**	**74.9**	**75.7**	**75.1**
		X	67.2	68.9	32.6	23.5	40.5	14.2	59.6	27.8	35.1	21.0	22.1	**37.5**

Table A4: Performance evaluation on the image-to-text and on the text-to-image retrieval task. ✓ and X denote inter-modal and intra-modal approaches, respectively. Blue rows and Purple rows indicate the usage of OTI- and OVI-inverted features, respectively. White rows refer to the inter-modal baselines.

			Image-to-Text						Text-to-Image					
			Flickr30k			COCO			Flickr30k			COCO		
	Backbone	Inter modal	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	B/32	✓	**78.8**	**94.9**	**98.2**	**50.1**	**75.0**	**83.5**	**58.8**	**83.5**	**90.0**	**30.5**	**56.0**	**66.9**
		X	64.5	86.6	92.5	39.8	64.5	74.6	52.7	77.9	86.2	25.6	49.1	60.5
	L/14	✓	**85.2**	**97.4**	**99.2**	**56.3**	**79.3**	**86.6**	**64.9**	**87.3**	**92.0**	**36.5**	**61.0**	**71.1**
		X	75.8	92.9	95.9	49.0	72.8	81.2	60.7	84.8	90.3	33.2	55.1	67.7
OPEN	B/32	✓	**79.2**	**93.8**	**96.2**	**53.5**	**77.7**	**86.0**	**61.1**	**84.9**	**90.9**	**37.1**	**62.4**	**72.7**
		X	72.8	90.3	94.1	49.2	73.4	82.0	57.4	81.5	88.4	33.1	58.0	68.4
	L/14	✓	**89.1**	**98.6**	**99.7**	**63.3**	**84.2**	**90.4**	**73.4**	**91.8**	**95.5**	**45.7**	**70.1**	**79.2**
		X	86.0	97.7	98.9	60.8	81.5	88.3	67.4	88.1	93.0	39.0	63.4	73.2
Sig LIP	B/16	✓	**89.0**	**98.0**	**99.2**	**65.7**	**85.4**	**91.2**	**74.6**	**92.3**	**95.6**	**47.8**	**72.4**	**81.0**
		X	81.8	95.5	97.3	57.0	79.0	86.2	57.9	82.6	88.7	33.7	58.2	68.9

APPENDIX F ADDITIONAL EXPERIMENTS

Here we report on additional experiments that support our claims about the importance of inter-modal representations for intra-modal problems when using contrastively-trained VLMs.

Zero-shot Image Classification with OVI. In the main paper we transform zero-shot image classification from being natively inter-modal to intra-modal using OVI (see the right section of Tab. 2). As expected, this approach decreased performance. Similarly, to further confirm our results, in Tab. A3 we transform it to intra-modal but using OVI instead. Consistent with our findings, approaching classification in an intra-modal manner also decreases performance.

Image-to-Text and Text-to-Image Retrieval with OTI and OVI. To provide additional experimental evidence that transforming inter-modal tasks in intra-modal decreases performance, we perform experiments on image-text retrieval benchmarks using both OTI and OVI. Specifically, we apply OTI to the query image and use the resulting features to retrieve from the text gallery. Conversely, we apply OVI to the query text and use its features to retrieve from the image gallery. For image-to-text, and text-to-image retrieval we adhere to the standard benchmark and we report recall at 1 (R@1), recall at 5 (R@5), and recall at 10 (R@10). In Tab. A4 we show the quantitative results of these experiments, which confirm our findings from the zero-shot image classification setting: transforming an inter-modal task into an intra-modal one always leads to performance degradation due to intra-modal misalignment.

Table A5: Performance (mAP) evaluation on the text-to-text retrieval task using purely textual datasets. ✓ and ✗ denote inter-modal and intra-modal approaches, respectively. Purple rows indicate the usage of OVI-inverted features.

Method	Inter modal	IMDB	20News.	Climate	DBPedia	FEVER	NFCorpus	NQ	SciDocs	SciFact	*Average*
Baseline	✗	52.2	19.2	11.2	30.3	58.4	8.9	23.3	13.5	26.3	27.0
OVI	✓	**52.3**	**33.1**	**15.3**	**39.1**	**70.5**	**12.2**	**33.6**	**16.8**	**33.2**	**34.0**

Table A6: Performance (mAP) comparison between the proposed modality inversion techniques and the adapter-based approaches on the image-to-image (**left**) and text-to-text (**right**) retrieval tasks. ✓ and ✗ denote inter-modal and intra-modal approaches, respectively. Blue rows and Purple rows indicate the usage of OTI- and OVI-inverted features, respectively.

Method	Inter modal	CUB	SOP	$\mathcal{R}$Oxford	$\mathcal{R}$Paris	Cars	*Average*
Baseline	✗	22.9	34.4	42.6	67.9	24.6	38.5
Adapter	✓	23.7	35.0	**44.3**	69.5	25.5	39.6
OTI	✓	**24.6**	**35.1**	43.0	**70.3**	**28.0**	**40.2**

Method	Inter modal	Flickr30k	COCO	nocaps	*Average*
Baseline	✗	51.7	26.2	35.1	37.7
Adapter	✓	51.9	**28.3**	37.8	39.3
OVI	✓	**54.8**	**28.3**	**38.8**	**40.6**

Text-to-text Retrieval on Purely Textual Datasets. In Sec. 6.2 we conduct a text-to-text retrieval experiment using image captioning datasets to avoid a mismatch with VLMs pre-training data. In this section, we evaluate the performance of OVI on purely textual datasets using the CLIP ViT B/32 model. Specifically, we select seven datasets from the NanoBEIR[1] benchmark, spanning diverse domains such as scientific documents (SciDOCS) and climate-related texts (ClimateFEVER). We discard Question-Answering (QA) datasets and those with queries whose average length exceeds the context length of CLIP's text encoder (77 tokens). Additionally, we include the IMDB Reviews (Maas et al., 2011) and 20 Newsgroups (Lang, 1995) datasets. All selected datasets comprise texts that cannot be easily represented visually. Examples include "Learning Actionable Representations with Goal-Conditioned Policies" (SciDocs), "Atheism, philosophy, and the absence of belief in deities" (20 Newsgroup), and "The carbon footprint on wind energy is significant" (ClimateFEVER). Since gallery texts often exceed CLIP's context length, we employ a Large-Language Model (Llama-3.2-1B-Instruct[2]) to summarize them to fit within the token limit. We report the results in Tab. A5. OVI achieves a significant performance improvement over the intra-modal baseline. This outcome demonstrates that OVI is effective even when considering texts that can not be easily represented visually.

Inter-modal Representation via Adapters. To broaden our comparative analysis we conduct an additional experiment where we train two single-layer linear adapters: one maps image features to text features (aligned with the goal of OTI), and the other maps text features to image features (aligned with the goal of OVI). For training, we leverage the LLaVA-CC3M[3] dataset (Liu et al., 2024), which comprises 595K image-text pairs. This dataset is derived by filtering the CC3M dataset (Sharma et al., 2018) to achieve a more balanced distribution of concept coverage. We train each adapter using a cosine loss that minimizes the distance between the adapter output and the corresponding complementary features. Additionally, following Patel et al. (2024), we also employ a CLIP-based contrastive loss component. Table A6 presents the results for image-to-image and text-to-text retrieval tasks using the CLIP ViT-B/32 model. The adapter-based approach improves performance over the intra-modal baseline for both tasks. These findings support our claim that leveraging inter-modal representations for intra-modal tasks enhances performance thanks to CLIP's inherent inter-modal alignment. Nevertheless, we observe that OTI and OVI outperform the adapter-based approach in most scenarios. This result emphasizes the effectiveness of OTI and OVI, as they do not require a training dataset but rather map individual features directly to the complementary modality without relying on external resources.

---

[1] https://huggingface.co/collections/zeta-alpha-ai/nanobeir-66e1a0af21dfd93e620cd9f6

[2] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct

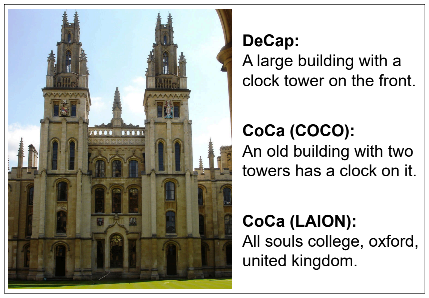
[3] https://huggingface.co/datasets/liuhaotian/LLaVA-CC3M-Pretrain-595K

Table A7: Performance (mAP) comparison between the proposed OTI technique and the captioning-based approach on the image-to-image retrieval task. ✓ and X denote inter-modal and intra-modal approaches, respectively. Blue rows indicate the usage of OTI-inverted features.

Method	Inter modal	CUB	SOP	$\mathcal{R}$Oxford	$\mathcal{R}$Paris	Cars	*Average*
Baseline	X	22.9	34.4	42.6	67.9	24.6	38.5
DeCap	✓	4.4	2.0	0.1	1.2	2.5	2.0
CoCa (COCO)	✓	3.5	0.8	0.0	0.7	1.8	1.4
CoCa (LAION)	✓	17.6	3.9	8.4	28.2	23.6	16.3
OTI	✓	**24.6**	**35.1**	43.0	**70.3**	**28.0**	**40.2**

Inter-modal Representations via Captioning. We compare the performance of OTI on image-to-image retrieval with a captioning-based approach using the CLIP ViT-B/32 model. Specifically, given a query image, such an approach involves generating a caption with a pre-trained captioning model, leveraging CLIP's text encoder to extract text features from the caption and using them to perform retrieval. We experiment with three pre-trained captioning models: 1) DeCap (Li et al., 2023), which directly generates captions from CLIP image features; 2) CoCa (LAION)[4] (Yu et al.), trained on the Laion2B (Schuhmann et al., 2022) dataset; and 3) CoCa (COCO)[5] (Yu et al.), pre-trained on Laion2B and fine-tuned on COCO (Lin et al., 2014).

Table A7 shows the results. Regardless of the captioning model, the captioning-based approach achieves unsatisfactory performance, even falling short of the intra-modal baseline despite leveraging CLIP's inter-modal alignment. This outcome stems from the fact that the generated captions are not discriminative enough to perform image retrieval. This is particularly evident in fine-grained domains such as the buildings of the $\mathcal{R}$Oxford and $\mathcal{R}$Paris datasets (Radenović et al., 2018). Figure A1 shows an example of generated captions for a randomly chosen image from the $\mathcal{R}$Oxford dataset. We observe that all the captioning models generate generic and not sufficiently discriminative captions. CoCa (LAION) produces a more precise description than the other models, reflecting its higher performance. Nevertheless, OTI obtains better results than the captioning-based approach and the intra-modal baseline.



DeCap:
A large building with a clock tower on the front.

CoCa (COCO):
An old building with two towers has a clock on it.

CoCa (LAION):
All souls college, oxford, united kingdom.

Figure A1: Captions generated by pre-trained captioning models for an image from the $\mathcal{R}$Oxford dataset.

Intra-OTI Similarity Comparisons. We conduct an experiment on image-to-image retrieval where we apply OTI to both the query and gallery images and perform retrieval via the similarity between the OTI-inverted features. Since OTI maps image features into text features, this intra-OTI strategy involves intra-modal similarity comparisons within the text embedding space.

Table A8 shows the results on image retrieval datasets using the CLIP ViT-B/32 model. We observe that employing inter-modal similarity comparisons by applying OTI only to the query images achieves better performance than using intra-modal similarities with the intra-OTI approach. These findings confirm that modality inversion techniques do not inherently improve performance. Instead, their effectiveness lies in leveraging CLIP's inter-modal alignment by transforming intra-modal tasks into inter-modal ones.

Table A8: Performance (mAP) evaluation on the image-to-image retrieval task. ✓ and X denote inter-modal and intra-modal approaches, respectively.

Method	Inter modal	CUB	SOP	$\mathcal{R}$Oxford	$\mathcal{R}$Paris	Cars	*Average*
Baseline	X	22.9	34.4	42.6	67.9	24.6	38.5
Intra-OTI	X	21.3	31.9	42.3	68.2	24.9	37.7
OTI **(ours)**	✓	**24.6**	**35.1**	43.0	**70.3**	**28.0**	**40.2**

---

[4] https://huggingface.co/laion/CoCa-ViT-B-32-laion2B-s13B-b90k

[5] https://huggingface.co/laion/mscoco_finetuned_CoCa-ViT-B-32-laion2B-s13B-b90k

Table A9: Impact on the performance (mAP) of the OTI template sentence on the image-to-image retrieval task. Each prompt is given by the combination of a template sentence with the pseudo-word token v^*.

OTI Prompt	CUB	SOP	$\mathcal{R}$Oxford	$\mathcal{R}$Paris	Cars	*Average*
"v^*" (*empty prompt*)	24.0	34.6	**43.7**	69.6	28.2	40.0
"We see v^* in this photo"	24.5	34.7	43.0	69.7	**28.3**	40.0
"An image of v^*"	24.0	34.8	43.1	**70.7**	28.3	**40.2**
"A photo of v^*" **(ours)**	**24.6**	**35.1**	43.0	70.3	28.0	**40.2**

Impact of the OTI Template Sentence. As detailed in Sec. 5.1, for OTI we concatenate the template sentence "a photo of" with the pseudo-word token v^* to craft the prompt "a photo of v^*". To study the impact on the performance of the template sentence, we test the following prompts: 1) "an image of v^*"; 2) "we see v^* in this photo"; and 3) "v^*" (the *empty prompt*). Table A9 reports the image-to-image retrieval results using the CLIP ViT-B/32 model. We observe that all the considered prompts achieve comparable performance. These results demonstrate the robustness of the OTI technique to the template sentence.

Combining Native and Inverted Features. We conduct an experiment on image-to-image retrieval to assess whether combining native image features with the corresponding OTI-inverted features improves the performance. Let $\psi_I = f_\theta(I)$ be the native image features and $\psi_T = g_\phi(\overline{Y}_{v^*})$ be the OTI-inverted features. We query the gallery using a weighted combination of native and OTI-inverted representations:

$$\psi_{IT} = \alpha * \psi_T + (1 - \alpha) * \psi_I, \tag{8}$$

where $\alpha \in [0, 1]$ is a weighting factor that controls the contribution of each component.

Table A10 reports the results on image-to-image retrieval datasets for varying values of α using the CLIP ViT-B/32 model. Interestingly, for α large enough, combining native and inverted features obtains better results than relying solely on either of them. Nevertheless, regardless of the α value, we observe that employing inter-modal representations always improves the performance over the intra-modal baseline. We will further investigate the combination of intra- and inter-modal representations in future work.

Table A10: Performance (mAP) evaluation of the combination between native and OTI-inverted features for varying weighting factors α for image-to-image retrieval.

Method	CUB	SOP	$\mathcal{R}$Oxford	$\mathcal{R}$Paris	Cars	*Average*
Baseline ($\alpha = 0$)	22.9	34.4	42.6	67.9	24.6	38.5
OTI ($\alpha = 0.25$)	24.0	35.6	44.9	70.1	25.9	40.1
OTI ($\alpha = 0.50$)	24.6	**36.1**	**46.7**	71.0	27.0	41.1
OTI ($\alpha = 0.75$)	**24.8**	35.9	46.3	**71.1**	27.7	**41.2**
OTI ($\alpha = 1$) **(ours)**	24.6	35.1	43.0	70.3	**28.0**	40.2