RETHINKING MULTIMODAL (MIS)ALIGNMENT FOR VISION-LANGUAGE REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) commonly learn by aligning image and text representations, achieving strong performance on zero-shot classification and image-text retrieval. However, this paradigm is sensitive to imperfect image-text correspondence and can overlook information that is complementary rather than shared across modalities. We investigate whether multimodal representation learning can provide a more robust alternative while retaining strong cross-modal representations. For the first time, we systematically assess CoMM as a vision-language representation learner, studying its behavior across image-text alignment quality and pretraining scale. We find that CoMM is substantially more robust to misaligned pretraining pairs than alignment-based objectives such as CLIP, while learning strong visual and multimodal representations. At larger scale, however, CoMM lags behind CLIP on standard cross-modal tasks, revealing a trade-off between robustness to misalignment and efficient learning of explicit image-text alignment. To bridge this gap, we introduce **CoMM-VL**, which augments CoMM's multimodal objective with an explicit image-text alignment loss within the same fused representation space. CoMM-VL substantially improves zero-shot classification and image-text retrieval over CoMM and outperforms established alignment-based VLMs, including CLIP, SigLIP, and TIPS, while preserving CoMM's robustness to misaligned data and strong multimodal representations. Across small- and large-scale vision-language datasets, these results show that explicit cross-modal alignment and multimodal interaction can be complementary rather than competing objectives. Code and pretrained weights will be open-sourced.

1 INTRODUCTION

Learning from multiple modalities is fundamental to human cognition: by seeing, smelling, touching, hearing, or tasting the world around us, we build knowledge about our environment and learn to understand and navigate it. Similarly, multimodal AI aims to integrate diverse data sources and capture their interactions to learn semantically meaningful representations. Vision-language models (VLMs), which combine visual and textual information, exemplify this paradigm. Models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) showed that large-scale contrastive pretraining on image-text pairs can produce powerful representations that transfer to downstream tasks, notably zero-shot image classification and image-text retrieval. Their success has motivated alternative pretraining strategies, including image-text matching, multimodal masked modeling, and combinations of multimodal and unimodal objectives (Mu et al., 2022; Singh et al., 2022).

Despite these differences, most VLMs rely on a fundamental assumption: **the semantic information relevant for learning is sufficiently *shared* across modalities**. Cross-modal alignment objectives, like CLIP, explicitly encourage this shared information to be represented similarly. Yet this assumption is often violated in real-world data: a caption may describe only part of an image, while an image may contain information absent from its associated text. More generally, **modalities can provide complementary information that cannot be recovered from their counterparts**, such as patient history or clinical information that enrich the interpretation of a medical image.

CoMM (Dufumier et al., 2025) offers an alternative to conventional alignment-based contrastive VLMs. Rather than directly aligning image and text representations in a shared embedding space, CoMM uses a **fusion module that jointly processes the two modalities to produce a single multi-**

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

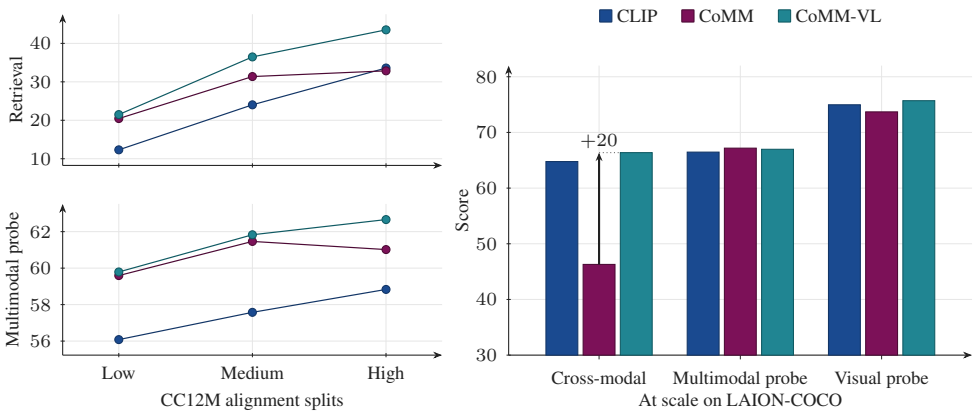


Figure 1: **Motivation for CoMM-VL.** Left: as image-text correspondence decreases on CC12M, alignment-based CLIP degrades substantially, whereas CoMM remains more robust across retrieval and multimodal evaluation. Right: when scaling to LAION-COCO, CoMM remains competitive on multimodal and visual representations but trails CLIP on cross-modal tasks. CoMM-VL addresses this trade-off, retaining CoMM's robustness while recovering strong cross-modal alignment.

modal representation. Its objective encourages this fused representation to capture information redundantly shared across modalities while retaining modality-specific information, potentially making it less dependent on precise image-text correspondence. However, its behavior across alignment regimes and pretraining scales, as well as its performance on standard cross-modal tasks such as zero-shot classification and image-text retrieval, remain unexplored.

To investigate the trade-off between robustness to imperfect image-text alignment and the ability to learn strongly aligned cross-modal representations, we study how these models behave as alignment quality and pretraining scale vary. Using CC12M (Changpinyo et al., 2021), we partition image-text pairs by alignment quality and find that alignment-based models, like CLIP, degrade substantially as correspondence decreases, whereas CoMM remains markedly more robust across retrieval and multimodal evaluation (Figure 1, left). Scaling to the larger LAION-COCO dataset (Schuhmann et al., 2022b), CoMM remains competitive on multimodal and visual representations but substantially trails CLIP on zero-shot classification and image-text retrieval (Figure 1, right). We hypothesize that CoMM recovers redundant cross-modal information, but learns explicitly aligned representations more slowly because alignment emerges only indirectly through multimodal interactions.

Motivated by these observations, we introduce **CoMM-VL**, which complements CoMM's multimodal objective with explicit image-text alignment of the image-only and text-only representations produced in the same fusion space. As shown in Figure 1, CoMM-VL retains CoMM's robustness to weak alignment and strong visual and multimodal representations while closing the gap with CLIP on cross-modal tasks. We evaluate these models under a broad and controlled protocol, varying image-text alignment quality while keeping the pretraining setting fixed, scaling from CC12M to LAION-COCO, and assessing cross-modal, multimodal, and visual representation quality, as well as transfer to a large language model. Together, these experiments show that multimodal interaction and explicit cross-modal alignment can be complementary rather than competing objectives.

We summarize our contributions as follows:

- We study how image-text alignment quality affects vision-language pretraining, showing that CoMM is substantially more robust than alignment-based models to misaligned training pairs.
- We evaluate CoMM on standard cross-modal tasks and across pretraining scales, from CC12M to LAION-COCO, revealing a trade-off: CoMM learns strong multimodal and visual representations but acquires explicit cross-modal alignment less efficiently than CLIP.
- We introduce **CoMM-VL**, which combines CoMM's multimodal interaction objective with explicit image-text alignment, improving zero-shot classification and retrieval while preserving its robustness and multimodal representation quality.

2 Related Work

Vision-Language models. Vision-Language Models (VLMs) combine visual and textual information to learn representations that support a wide range of tasks, including zero-shot image classification, image-text retrieval, and visual question answering (VQA) (Radford et al., 2021; Lu et al., 2019; Li et al., 2022). The success of cross-modal contrastive learning with models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) established image-text alignment as a central paradigm for VLM pretraining. Subsequent approaches have explored both alternative alignment objectives and additional sources of supervision. SigLIP (Zhai et al., 2023), for example, replaces CLIP's softmax-based contrastive objective with a pairwise sigmoid loss. More recent approaches such as SigLIP2 (Tschannen et al., 2025) and TIPS (Maninis et al., 2025) further combine image-text alignment with complementary objectives, including self-supervised visual learning and captioning, to improve the quality and transferability of the learned representations. Despite these advances, VLMs remain strongly dependent on the correspondence between the visual and textual modalities, which are typically paired at scale using web-scraped data.

Dealing with data misalignment. Web-scraped image-text datasets inevitably contain noisy, incomplete, or weakly aligned pairs, which can compromise cross-modal supervision. A common response has been to improve the quality of the training data through filtering and curation. For example, BLIP (Li et al., 2022) introduces a caption filtering mechanism (CapFilt) to remove poorly aligned image-text pairs, while other approaches construct or curate datasets to improve image-text correspondence (Schuhmann et al., 2022a; Nguyen et al., 2023). The importance of data quality and filtering has also been highlighted by DataComp (Gadre et al., 2024), which explicitly fixes the model and training procedure and instead makes dataset construction the central challenge, with participants selecting and curating image-text pairs from large-scale web data. While these approaches can improve the quality of cross-modal supervision, they treat misalignment primarily as a property of the data to be corrected or removed. This can be restrictive in applied domains, such as medical imaging or Earth observation, where large-scale, high-quality paired data are difficult to obtain, and where different modalities may naturally contain complementary information. This raises a broader question: rather than enforcing stronger alignment between modalities, can multimodal models learn to exploit their interactions and complementary information directly?

Multimodal learning. Rather than relying solely on cross-modal alignment, recent multimodal learning approaches explicitly model interactions between modalities. FactorCL (Liang et al., 2023), for example, separates shared and modality-specific information, while CoMM (Dufumier et al., 2025) was specifically designed to learn shared, unique, and synergistic information between modalities. CoMM has shown promising results on multimodal tasks, including vision and language, but how such multimodal representations translate to capabilities traditionally associated with VLMs, such as image-only understanding, zero-shot classification, and image-text retrieval, remains less explored. In this work, we investigate this question through large-scale experiments, studying how multimodal representations behave under varying degrees of image-text alignment, and how explicit alignment can be incorporated without compromising their multimodal capabilities.

3 VLM robustness to data misalignment

In this section, we study how the degree of image-text alignment in the pretraining data affects the multimodal representations learned by VLMs. To do so in a controlled setting, we first compute an alignment score for each image-text pair and use predefined thresholds on this score to partition the pretraining dataset into splits of increasing alignment (Sec. 3.2). We then train VLMs on each split and evaluate them on a range of visual and multimodal downstream tasks.

3.1 Background

Notation. Let E_I and E_T be Transformer-based image and text encoders. An image-text pair $x = (x_I, x_T)$ is mapped to unimodal latent representations $(z_I, z_T) = (E_I(x_I), E_T(x_T))$. $\mathcal{L}_{\text{NCE}}$ stands for the InfoNCE loss (Oord et al., 2018).

CLIP: aligning the two modalities. CLIP (Radford et al., 2021) directly aligns the image and text representations with a symmetric InfoNCE objective:

$$\mathcal{L}_{\text{CLIP}}(z_I, z_T) = \frac{1}{2}\left(\mathcal{L}_{\text{NCE}}(z_I, z_T) + \mathcal{L}_{\text{NCE}}(z_T, z_I)\right). \tag{1}$$

Because the only training signal is the *agreement* between an image and its caption, CLIP is encouraged to retain information that both modalities share, and has no incentive to encode what only one modality conveys.

CoMM: aligning multimodal views. Cross-modal objectives such as CLIP implicitly rely on the *multiview redundancy* assumption, i.e., that the task-relevant information is shared between modalities. However, this information is often distributed across modalities in other ways. Given a downstream task y, partial information decomposition (PID) (Williams & Beer, 2010; Bertschinger et al., 2014) decomposes the information that the two modalities carry about y into four terms, $I(x_I, x_T; y) = R + U_I + U_T + S$. *Redundancy* R is the information about y available from either modality; *uniqueness* U_i is the information available only from modality i; and *synergy* S is the information that becomes available only when both modalities are observed together. Under multiview redundancy, cross-modal contrastive methods can only learn R (Dufumier et al., 2025).

CoMM (Dufumier et al., 2025) overcomes this limitation by aligning multimodal rather than unimodal representations. A fusion Transformer f maps the unimodal representations to a single multimodal representation $z = f(z_I, z_T)$. Two label-preserving multimodal augmentations of x yield the views $z' = f(z'_I, z'_T)$ and $z'' = f(z''_I, z''_T)$, and masking all modalities but one yields the masked-modality views $z_I^M = f(z'_I, \varnothing)$ and $z_T^M = f(\varnothing, z'_T)$. CoMM then maximizes the mutual information between these views:

$$\mathcal{L}_{\text{CoMM}} = \underbrace{\mathcal{L}_{\text{NCE}}(z', z'')}_{\text{multimodal invariance}} + \sum_{i\in\{I,T\}} \underbrace{\frac{1}{2}\left(\mathcal{L}_{\text{NCE}}(z_i^M, z') + \mathcal{L}_{\text{NCE}}(z_i^M, z'')\right)}_{\text{unimodal information preservation}}. \tag{2}$$

Minimizing the first term captures all interactions $R + S + U_I + U_T$, provided the augmentations preserve task-relevant information, while each unimodal term captures the information $R + U_i$ contained in modality i. In practice, both terms are complementary: the unimodal terms accelerate the learning of modality-specific features, which the first term then combines to capture synergy.

Interestingly, CoMM never compares z_I^M and z_T^M directly, and thus has no explicit cross-modal alignment term. Redundancy is nonetheless learned, as it appears in every term of equation 2.

3.2 Misalignment protocol

To study the effect of data misalignment on the performance of VLMs, we create three splits of CC12M with controlled levels of image-text alignment. We define an alignment score between the image and text representations coming from a Perception Encoder G14 (Bolya et al., 2026) by measuring their cosine similarity. We define a **LowAlign**, **MediumAlign** and **HighAlign** split by selecting 3 million image-text pairs in each portion of the distribution, see Fig. 2. All splits contain the same number of training pairs, so the comparison is not confounded by the amount of pretraining data, and the rest of the training setup is held fixed across splits. Differences in downstream performance therefore reflect the alignment of the image-text pairs rather than the data scale or the training budget. This lets us compare directly how different multimodal objectives respond to increasingly well-aligned supervision.

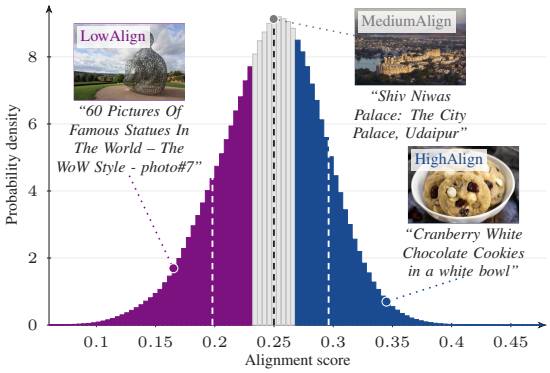


Figure 2: **CC12M alignment-score distribution and its three splits.** PE-Core-G14 image-text alignment score over CC12M pairs; the splits are tertiles of it, with means $\mu = 0.20$ / 0.25 / 0.30 (dashed).

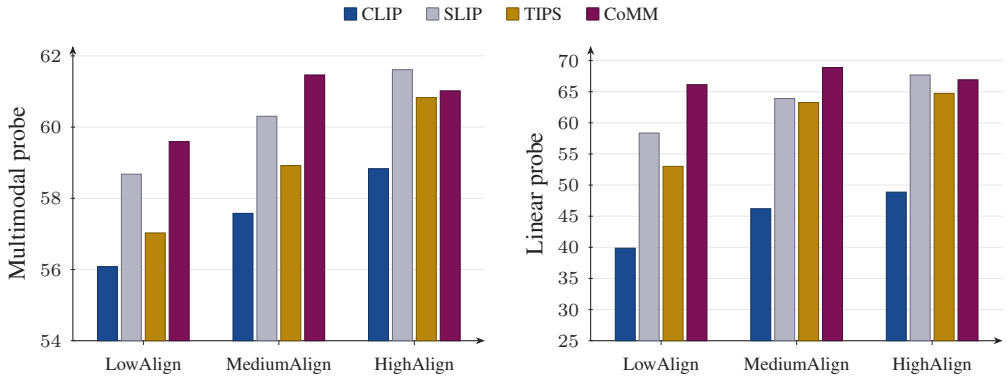


Figure 3: **Multimodal and visual representations across the CC12M alignment splits.**

3.3 CoMM IS ROBUST TO DATA MISALIGNMENT FOR MULTIMODAL AND VISUAL TASKS

We evaluate the multimodal representations of state-of-the-art VLMs, including CLIP (Radford et al., 2021), SLIP (Mu et al., 2022), TIPS (Maninis et al., 2025), and CoMM (Dufumier et al., 2025), across five multimodal tasks (see Section 5.2 for more details) when pretrained on the three splits of CC12M. We also evaluate their visual representations on ImageNet using linear probing to evaluate the capacity of VLMs to extract unique information from a single modality (images, in this case).

Looking at Fig. 3, we show that CLIP is the most sensitive model to data misalignment, underperforming compared to CoMM on the low-align split by -26% linear probing accuracy on ImageNet and -3.5% on multimodal tasks. More recent approaches such as TIPS or SLIP highly improve the visual and multimodal representations over CLIP but still lag behind CoMM on the low-align split (-8% accuracy on ImageNet and -1% on multimodal tasks for SLIP). This result holds also across the medium-align split, suggesting that learning only redundant information across modalities is suboptimal when data are misaligned and learning unique, synergistic and redundant information across modalities is key for strong multimodal learning (Dufumier et al., 2025). On the high-align split, CoMM still outperforms the recent TIPS approach while being slightly worse than SLIP in this case (-0.8% accuracy on ImageNet and -0.6% on multimodal tasks). This is more expected in this case since mostly redundant information is present across modalities, favoring alignment-based models.

3.4 CoMM'S LIMITATIONS ON CROSS-MODAL TASKS

While CoMM produces strong multimodal and visual representations, it is not clear how to perform cross-modal tasks such as zero-shot classification or image-text retrieval from the original architecture (Dufumier et al., 2025). The natural approach is to compare the "fused" representations (z_I^M, z_T^M), which are obtained by passing each modality through the fusion module with the other one masked. Another approach would be to compare the "native" representations (z_I, z_T) (before fusion) but the two unimodal spaces are not comparable in CoMM's original architecture (different dimensionalities and no projections). We thus keep the first approach.

Table 1: **Zero-shot and retrieval results on CC12M and LAION-COCO.**

Method	CC12M			LAION-COCO		
	ZS	i2t	t2i	ZS	i2t	t2i
CLIP	**39.3**	**41.5**	**29.5**	57.1	**63.9**	46.1
SigLIP	25.8	29.7	20.2	**57.7**	**63.9**	**47.2**
CoMM	33.6	33.7	24.0	40.2	39.6	25.1

We report the performance in Table 1 on CC12M (full split) and at large-scale on LAION-COCO (402M pairs, see Sec. 5.1), using CLIP and SigLIP models as baselines[1]. We observe that CoMM performs poorly while having very good unimodal (visual) representations: 73.9% and 73.7% linear probe accuracy on ImageNet on CC12M and LAION-COCO respectively. We hypothesize that the

[1]In order to avoid early over-fit of CLIP on CC12M, we use strong SimCLR image augmentations.

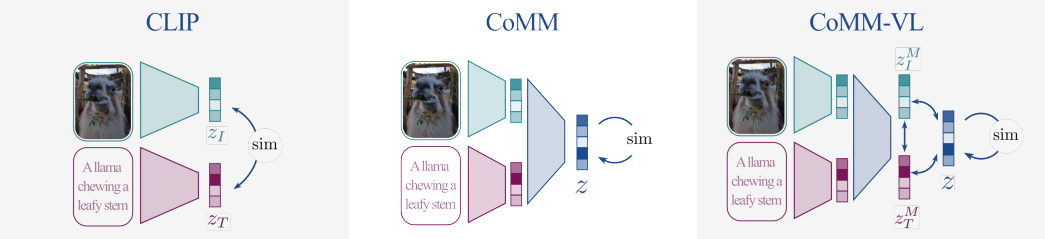


Figure 4: **Overview of the CoMM-VL architecture.** Compared to CoMM, it adds a cross-modal alignment objective in the fused space, which lets the model produce consistent visual and textual representations for downstream zero-shot transfer and image-text retrieval tasks while keeping its strong multimodal capacity.

visual and language space are rich but not aligned enough to perform these tasks, leading us to ask: can we combine the best of both multimodal and cross-modal worlds?

4 CoMM-VL: Aligning Modalities in the Multimodal Space

As shown in the previous section, CoMM learns strong multimodal representations when pretrained on vision-language data. However, it never aligns the image and text representations with each other (see Section 3.1). As a result, it underperforms on tasks that require comparing an image and a text directly, such as cross-modal retrieval or zero-shot classification. In this section, we introduce CoMM-VL, which extends CoMM with an explicit cross-modal alignment in the multimodal space, so that a single model provides both multimodal representations and aligned unimodal representations. Fig. 4 gives an overview of our approach, compared to CoMM and CLIP.

Architecture and training. CoMM-VL follows CoMM's architecture. The image and text encoders E_I and E_T map an image-text pair $x = (x_I, x_T)$ to sequences of token embeddings (z_I, z_T). These are processed by a fusion Transformer f, which outputs the multimodal representation $z = f(z_I, z_T)$.

During training, two augmentations t' and t'' are drawn and applied to each image-text pair x, obtaining $x' = t'(x)$ and $x'' = t''(x)$. Each augmentation combines SimCLR image transformations (Chen et al., 2020) with random text masking. We also consider the masked-modality inputs $(x'_I, \varnothing)$ and $(\varnothing, x'_T)$, where only one modality is kept. These four inputs are encoded by the network to obtain four embeddings, namely z', z'', z_I^M and z_T^M.

Loss function. CoMM's objective aligns the multimodal views z' and z'' with each other and each masked-modality view z_i^M with both multimodal views (see equation 2), but it never aligns image and text representations with each other. We propose to additionally align the masked-modality representations (z_I^M, z_T^M) in the multimodal space consistently with CoMM's objective. Of note, the additional computational cost is negligible, since these representations are already computed during CoMM's training. We thus propose an additional symmetric InfoNCE loss between representations:

$$\mathcal{L}_{\text{CoMM-VL}} = \mathcal{L}_{\text{CoMM}} + \mathcal{L}_{\text{CLIP}}(z_I^M, z_T^M) \tag{3}$$

Inference. At inference, no augmentations are applied, and CoMM-VL can be used in two ways depending on the task. For multimodal tasks, the image-text pair $x = (x_I, x_T)$ is processed through the full network to obtain the multimodal representation $z = f(z_I, z_T)$, which can be transferred by linear probing or by fine-tuning the whole architecture, as in CoMM. For cross-modal tasks, images and texts are instead embedded separately through the masked-modality paths, yielding $z_I^M = f(z_I, \varnothing)$ and $z_T^M = f(\varnothing, z_T)$, and compared with cosine similarity.

5 EXPERIMENTS

5.1 SETUP

We evaluate CoMM-VL when pretrained at different data scales and we compare it against the previous VLMs. We pretrain at two scales: first on CC12M (full split) for 50k steps and we evaluate the representation on our complete evaluation suite (see Sec. 5.2). Second, we perform pretraining for 100k steps at a larger scale on LAION-COCO (Schuhmann et al., 2022b) containing 402M images from LAION-5B (Schuhmann et al., 2022a) captioned with synthetically generated COCO-style captions. All models are pretrained from scratch with the same batch size 16,384, a ViT-B/16 (Dosovitskiy et al., 2021) as vision encoder and a 12-layer Transformer (Radford et al., 2021) as text encoder (see Supplementary A.1 for more details).

Implementation details. For the fusion encoder, we use a slightly deeper 2-layer Transformer architecture compared to CoMM. We use token pooling before fusion for small-to-medium scale pretraining data (full CC12M and its subparts). For larger-scale experiments, we do not use any pooling and we encode all visual and textual tokens along with their CLS tokens. In this case, we use attention normalization to stabilize the training. We use SimCLR's image augmentation (Chen et al., 2020) for training CoMM and CoMM-VL and we use random text masking with probability $p = 0.15$ for all experiments except on LAION-COCO for CoMM-VL where we set $p = 0$ (no masking).

5.2 EVALUATION SUITE

Multimodal tasks. To assess VLM representations, we consider diverse multimodal downstream tasks requiring reasoning across modalities, with a focus on datasets exhibiting strong cross-modal synergy (Liang et al., 2024). We evaluate on MM-IMDb (Arevalo et al., 2017) for multi-label movie genre classification from posters and plot summaries; Hateful Memes (Kiela et al., 2020) for hateful content detection from memes combining images and overlaid text; SNLI-VE (Xie et al., 2019) for visual natural language inference, where an image and hypothesis must be classified as entailment, contradiction, or neutrality; VQAv2 (Goyal et al., 2017) for visual question answering requiring reasoning about objects, attributes, and relationships in images; and Fashion-Gen (Rostamzadeh et al., 2018) for product-category prediction from clothing images and their textual descriptions. In what follows, we denote MM the linear probe performance average over these five tasks; the per-task results are given in Table 8.

Vision-Language alignment tasks. We further evaluate the cross-modal capacities of VLMs through zero-shot accuracy (ZS) on ImageNet (how well a textual concept is aligned with its visual representation?), denoted ZS, and through image-to-text and text-to-image retrieval tasks averaged on Flickr30k (Young et al., 2014) and MS-COCO (Lin et al., 2014), denoted i2t and t2i. The per-dataset retrieval results are given in Table 9.

Visual tasks: in-domain, robustness and fine-grained zero-shot transfer. We evaluate the visual representation using linear probing on ImageNet (LP). Zero-shot accuracy on ImageNet measures transfer to a distribution close to the web images the models are pre-trained on. Two further zero-shot suites probe transfer away from it. The first is three ImageNet robustness variants: ImageNet-R (Hendrycks et al., 2021) (artistic renditions), ImageNet-Sketch (Wang et al., 2019) (hand drawings) and ImageNetV2 (Recht et al., 2019) (an independent re-collection of the original test set). All three live in ImageNet's label space, so they reuse its class names and its prompt ensemble, and each is scored over the subset of classes it covers; the tables report the average over the three, denoted IN-var. The second is nine fine-grained classification datasets: Caltech101, Oxford Pets, Cars (Krause et al., 2013), Flowers 102, Food-101, FGVC-Aircraft, SUN397 (Xiao et al., 2010), DTD (Cimpoi et al., 2014) and UCF101, using the same test splits as Zhou et al. (2022) and a single handcrafted prompt per dataset, which makes the per-dataset numbers comparable with published ones; this is therefore a different measurement from the linear probes above, even for a dataset that appears in both suites. The tables report the average over the nine, denoted FG, and the per-dataset numbers are given in Sec. A.2.

Table 2: **CoMM-VL results when pre-training on CC12M and LAION-COCO.** All models use the same ViT-B/16 architecture.

	Method	Zero-shot			Linear probe	MM-LP	Retrieval	
		IN-1k	IN-var	FG	top-1	avg	i2t	t2i
CC12M	CLIP	27.2	25.4	25.9	58.2	60.1	33.8	22.4
	CLIP†	39.3	37.0	31.8	68.9	61.4	41.5	29.5
	SigLIP	25.8	23.4	23.0	57.0	59.4	29.7	20.2
	SLIP	34.6	31.1	29.5	66.2	61.5	44.9	31.1
	TIPS	37.7	35.8	32.2	70.7	61.2	48.8	34.8
	CoMM	33.6	29.4	31.9	73.9	**64.4**	33.7	24.0
	CoMM-VL	**41.8**	**39.4**	**38.0**	**74.2**	64.1	**50.1**	**36.5**
LAION-COCO	CLIP	57.1	54.0	61.6	75.0	66.5	63.9	46.1
	SigLIP	57.7	55.2	61.4	72.6	66.6	63.9	47.2
	SLIP	56.5	51.9	60.8	75.3	66.5	63.1	45.5
	TIPS	58.8	56.4	61.9	75.0	66.4	**65.6**	**48.2**
	CoMM	40.2	36.9	45.9	73.7	**67.2**	39.6	25.1
	CoMM-VL	**59.0**	**58.1**	**62.3**	**75.7**	67.0	**65.6**	47.8

5.3 Main results at medium scale

We present the results on CC12M in Table 2. CoMM-VL is the clear headline: it tops every alignment-sensitive metric while also posting the best linear probe and matching CoMM's fused MM-LP, confirming that adding a cross-modal alignment objective to CoMM closes its original weakness without sacrificing its multimodal strength. Plain CoMM tells the flip side of that story: best on concat MM-LP and a close second on ImageNet linear probe, but well behind on retrieval (33.7/24.0 vs. CoMM-VL's 50.1/36.5) and zero-shot (33.6 vs. 41.8), the exact gap CoMM-VL was built to fix. Among the alignment-only baselines, strong augmentation is doing a lot of the work for CLIP: +12.1 ZS (27.2→39.3) and +10.7 LP (58.2→68.9) over vanilla CLIP, but only +1.3 on MM-LP (60.1→61.4), suggesting it mainly improves representation quality rather than cross-modal complementarity.

5.4 Does it hold at scale?

On LAION-COCO at larger scale, CoMM-VL is essentially on par with CoMM on multimodal tasks (67.0 vs. 67.2 MM-LP) while decisively closing the gap everywhere else: it reaches the best zero-shot IN-1k (59.0), IN-var (58.1) and FG (62.3) among all methods, best LP (75.7), and retrieval on par with TIPS (65.6 i2t, 47.8 t2i vs. 65.6/48.2) — all without any teacher-student distillation, unlike TIPS. This is a sharp reversal from CoMM alone, which collapses on cross-modal tasks when tokens are pooled before fusion (39.6 i2t, 25.1 t2i, IN-1k 40.2): at this scale, keeping visual and textual tokens unpooled before fusion becomes clearly beneficial, a scaling trend we did not observe on the smaller CC12M splits. Notably, CoMM-VL's fusion space is only 512-d, half the 1024-d concatenated representation used by every baseline (including CoMM), yet it still yields the strongest or near-strongest multimodal probe accuracy in the table — evidence that at scale the fusion encoder learns to non-linearly aggregate multimodal information rather than relying on the raw dimensionality of a concatenated representation.

5.5 Frozen-LLM bridge evaluation.

A key downstream role of vision-language encoders is to serve as frozen visual front-ends for multimodal LLMs, where their patch tokens are mapped into the language model embedding space through a lightweight projector. We therefore evaluate whether the representations learned by CoMM and CoMM-VL transfer to this setting. We freeze both the visual encoder and the LLM, train only the projector on VQAv2, and evaluate object hallu-

Table 3: **POPE object hallucination under a frozen LLM.**

Method	Llama-3.2-1B			Qwen2-1.5B		
	Acc	F1	Yes-ratio	Acc	F1	Yes-ratio
Random	49.8	61.8	81.2	53.6	63.8	78.3
CLIP	62.8	64.7	55.3	79.0	78.0	45.2
SigLIP	76.6	74.7	42.6	79.0	77.6	44.1
CoMM	**77.3**	74.6	39.5	79.7	78.3	43.3
CoMM-VL	77.2	**75.4**	43.0	**80.8**	**79.3**	42.8

cination on POPE. As shown in Table 3, CoMM and CoMM-VL consistently outperform the CLIP and SigLIP baselines across both Llama-3.2-1B and Qwen2-1.5B. The gains are particularly pronounced with Llama, where CoMM-VL improves over CLIP by 14.4 points in accuracy and 10.7 points in F1, while matching CoMM in accuracy and achieving the highest F1. With Qwen, despite substantially stronger CLIP and SigLIP baselines, CoMM-VL further improves over CoMM by roughly one point in both accuracy and F1, reaching 80.8 and 79.3, respectively. Overall, these results show that the gains of CoMM-VL extend beyond standalone vision-language evaluation and transfer to multimodal LLMs through a lightweight learned bridge.

5.6 Ablation study

We ablate the key components in CoMM-VL compared to CoMM, in particular its cross-modal alignment objective and the fusion encoder. In these experiments, we test two settings for MM evaluations: we either concatenate the fused image and text representations (z_I^M, z_T^M) ("concat") or we use the multimodal fused readout ("fused"). In Table 4, we show that adding a cross-modal objective to CoMM in the encoder's native space improves zero-shot and image retrieval performance (cross-modal tasks), while preserving the MM performance in the fused representation but degrades the linear probe performance. Interestingly, adding this term makes the fused representation richer compared to concat while being of smaller dimension ($512d$ vs $1024d$ for concat). A deeper fusion module (2-layer Transformer) is beneficial for zero-shot and image-retrieval tasks while slightly degrading MM and LP performance. CoMM-VL is the best trade-off offering good performance across all tasks and enjoying good fused representation on MM tasks thanks to the cross-modal objective placed in the multimodal space.

Table 4: **Effect of cross-modal alignment and fusion depth on CC12M.** We ablate the placement of the cross-modal loss in CoMM-VL and the depth of the fusion encoder.

Method	Zero-shot	Linear probe	MM-LP		Retrieval	
			concat	fused	i2t	t2i
CoMM	33.6	73.9	64.4	64.1	33.7	24.0
+ 2-layer fusion	35.1	73.4	63.7	63.1	33.7	24.8
+ CLIP (native space)	37.1	71.5	63.3	64.2	37.6	28.3
CoMM-VL	**41.8**	**74.2**	63.6	64.1	**50.1**	**36.5**

6 Conclusion

In this work, we have investigated the robustness of VLMs to vision-language data misalignment that naturally exists in real-world datasets. While it should be at the core of multimodal learning, we showed in controlled experiments that state-of-the-art models (such as CLIP) are very sensitive to the misalignment gap which induces poor multimodal representations when pretrained on low-aligned data. On the other hand, we verified our prediction that modeling other multimodal interactions such as synergy and uniqueness using CoMM's architecture significantly increases the downstream performance on vision-language tasks. Nonetheless, we identified a key limitation in CoMM's design that highly limits its performance on cross-modal tasks such as zero-shot classification or image-text retrieval. We provided a simple fix, namely CoMM-VL, in the original loss function by adding a cross-modal alignment term in the fused space to align both the vision and the language space at training time. This design provided an important boost on cross-modal tasks while preserving the multimodal capacity of the original architecture. We assessed the performance of CoMM-VL at scale using 402M data pairs and across a large suite of in-domain and out-of-domain vision-language tasks, and found that it performs best overall among the models compared at a matched training budget.

Limitations and future work. There are several limitations in this work. First, CoMM's computational cost is higher at training time (but not inference) compared to CLIP because it requires two augmented image-text pairs for each sample and CoMM-VL still inherits this additional cost. An important direction for future work is to investigate teacher-student architectures to lower this computational cost. Second, studying the scaling of CoMM-VL as we increase the dataset size/model size/training time even further would be very valuable to validate the framework with frontier VLMs. It is now possible with the availability of web-scale open datasets (such as LAION). Finally, we envision that adding other modalities such as audio or videos (Zhu et al., 2024) to CoMM's training would further improve its multimodal representations, and we leave this for future work.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. Our study does not involve human subjects, and we did not collect any new data or personally identifiable information. All experiments use publicly available benchmarks, which we use in accordance with their respective licenses and terms of use. We acknowledge that these benchmarks may contain biases inherited from their original data sources, and our results should be interpreted with this in mind. We have no conflicts of interest to declare.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we have used publicly available datasets. Section 5.1 and Appendix A.1 describe neural network architectures, data augmentation strategies, and experimental settings used in all our experiments. Our code will be made available on GitHub and the pre-trained weights will be released.

AI USE STATEMENT

In this work, we **did not** use generative AI tools for generating data, developing theoretical models, formulating or proving mathematical claims, proposing hypotheses, designing experiments, implementing methods, translation, data cleaning, data analysis, or interpreting results. We used generative AI tools to proofread the manuscript for grammar, spelling, and readability, and to assist in creating code for some of the figures. **We have reviewed all AI-assisted work**: all suggested edits were reviewed and accepted or revised by the authors, and all plotting code was checked by the authors to ensure the figures accurately reflect our experimental results. We take responsibility for the final content of this manuscript, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. In *International Conference on Learning Representations - Workshops (ICLR-W)*, 2017. 7

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. 4

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, et al. Perception encoder: The best visual embeddings are not at the output of the network. *Advances in Neural Information Processing Systems*, 38:60884–60937, 2026. 4

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3558–3568, 2021. 2

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. 6, 7

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 7

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy. 7

Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? *International Conference on Learning Representations*, 2025. 1, 3, 4, 5

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. 3

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017. 7

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, 2021. 7

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. 1, 3

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020. 7

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 7

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022. 3

Paul Pu Liang, Zihao Deng, Martin Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:32971–32998, 2023. 3

Paul Pu Liang, Akshay Goindani, Talha Chafekar, Leena Mathur, Haofei Yu, Ruslan Salakhutdinov, and Louis-Philippe Morency. HEMM: Holistic evaluation of multimodal foundation models. *Advances in Neural Information Processing Systems*, 37:42899–42940, 2024. 7

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. 7

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, et al. TIPS: Text-image pretraining with spatial awareness. In *International Conference on Learning Representations*, 2025. 3, 5

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. 1, 5

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. *Advances in neural information processing systems*, 36:22047–22069, 2023. 3

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

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, 5, 7

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400, 2019. 7

Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018. 7

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, 2022a. 3, 7

Christoph Schuhmann, Andreas Köpf, Richard Vencu, Theo Coombes, and Romain Beaumont. Laion-coco: 600m synthetic captions from laion2b-en. https://laion.ai/blog/laion-coco/, 2022b. Accessed: 2026-09-25. 2, 7

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15638–15650, 2022. 1

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 3

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 7

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010. 4

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010. 7

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 7

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 7

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

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, 2022. 7, 14

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations*, volume 2024, pp. 9588–9608, 2024. 9

A Appendix

This appendix is organized in three parts. Sec. A.1 gives the experimental setup: the baselines, the pre-training recipe they share and the evaluation protocols. Sec. A.2 breaks the four summary columns of Table 2 down dataset by dataset, in both its CC12M and its LAION-COCO panel. Sec. A.3 reports experiments that correspond to no column of the main tables.

A.1 Experimental setup 13
A.1.1 Baselines 13
A.1.2 Training details 13
A.1.3 Evaluation protocols 14
A.2 Per-dataset results 15
A.2.1 Zero-shot on the ImageNet robustness variants 15
A.2.2 Zero-shot on the fine-grained datasets 15
A.2.3 Multimodal linear probe 15
A.2.4 Image-text retrieval 15
A.3 Additional analyses 16
A.3.1 CLIP and CoMM-VL at approximately matched parameter count 16

A.1 Experimental setup

Sec. A.1.1 describes the compared methods, Sec. A.1.2 the recipe they are all trained under, and Sec. A.1.3 the protocols every number in this appendix is measured with.

A.1.1 Baselines

Six methods are compared: CLIP, SLIP, SigLIP, TIPS, CoMM and CoMM-VL. All are pre-trained by us, from scratch, under the recipe of Sec. A.1.2. Each keeps the loss weights of its own formulation: SLIP weights its self-supervised term at 1.0 against the contrastive one, and TIPS weights its contrastive, DINO and iBOT terms at 1.0, 1.0 and 2.0, with an EMA teacher and centering of the teacher targets. SigLIP is the one method that needs gradient clipping, a global-norm clip at 1.0, to stay stable at this learning rate. The CC12M tables also carry a CLIP† row, which is the same CLIP trained with the SimCLR image augmentations, since plain CLIP over-fits early on a corpus of this size.

A.1.2 Training details

Every model is pre-trained under one recipe: the encoders, the optimizer, the learning-rate schedule and the global batch size are the same for all six, so that a difference in the results follows from the objective and from the pre-training corpus rather than from the training budget. Table 5 lists the settings, and the only per-method deviations are the ones given in Sec. A.1.1. What follows the corpus is the number of steps, 50k on the CC12M splits and 100k on LAION-COCO, with the warmup and drop phases as fixed fractions of it.

Training runs on 64 H100s with a local batch of 256, thus a global batch size of $16{,}384$ using a peak learning rate of 5×10^{-4}.

Weight decay applies to the weight matrices only. Every parameter of at most one dimension is excluded, which covers the biases, the normalization gains and the learnable logit scale, and so are

Table 5: **Training hyperparameters.** Settings for every model pre-trained in this work, on the CC12M alignment splits and on LAION-COCO. The settings of the other baselines that differ from this recipe are given in the text.

Hyperparameter	**CC12M**	**LAION-COCO**
Data		
Image–text pairs	3M per split, 9M full	402M
Captions	original web alt-text	original web alt-text
Image crop scale	$[0.9, 1.0]$	$[0.9, 1.0]$
CoMM augmented views	$[0.08, 1.0]$	$[0.08, 1.0]$
Model		
Image encoder	ViT-B/16	ViT-B/16
Text encoder	12 layers, width 512	12 layers, width 512
Context length / vocabulary	77 / 49,408	77 / 49,408
Embedding dimension	512	512
Initialization	from scratch	from scratch
Optimization		
Optimizer	AdamW	AdamW
β_1, β_2	0.9, 0.98	0.9, 0.98
ϵ	10^{-6}	10^{-6}
Peak learning rate	5×10^{-4}	5×10^{-4}
Weight decay	0.2	0.2
Weight decay exclusions	biases, norms, learned tokens	biases, norms, learned tokens
LR schedule	WSD	WSD
Warmup steps	2,500 (5%)	5,000 (5%)
Drop phase steps	7,500 (15%)	15,000 (15%)
Gradient clipping	NA	NA
Local batch size	256	256
Global batch size	16,384	16,384
Training steps	50k	100k
Samples seen	819M	1.64B
Precision	bfloat16 mixed	bfloat16 mixed
Hardware	64 × H100	64 × H100
Fusion module		
projection head	$512 \to 4096 \to 256$	$512 \to 4096 \to 256$
CoMM objective temperature	0.1	0.1
alignment temperature	0.01	0.01
alignment space	fused	fused
fusion input	pooled	pooled + tokens
fusion depth	2 layers	2 layers

the learned input tokens that are not one-dimensional: the CLS tokens, the register tokens and the per-modality position tables of the fusion module. The position embeddings of the two backbones are not excluded.

A.1.3 Evaluation protocols

Every number reported in this appendix comes from the final checkpoint of the corresponding run, on a single seed, and is measured on frozen features.

Zero-shot classification. The three ImageNet robustness variants share ImageNet-1k's label space and are scored with its class names and its 80-template prompt ensemble, over the subset of classes each one covers. The nine fine-grained datasets use the test splits of Zhou et al. (2022) with a single handcrafted prompt per dataset, which is what makes their numbers comparable with published ones.

Linear probes. Each probe is a linear classifier fit on frozen features, one per task, with its regularization strength selected on the validation split. The visual probe reads the native CLS token, except for CoMM-VL on LAION-COCO: that model fuses unpooled tokens, which leaves its back-

bone CLS untrained, so its row is probed in the fused space instead. The multimodal probe reads the concatenated 1024-d $[z_I; z_T]$ representation for every model except CoMM-VL, which is read in its 512-d fused space. Each multimodal task keeps the metric its own literature uses, so the five-task average mixes accuracy, AUROC and macro-F1 and is a summary rather than a quantity with a unit.

Retrieval. Both datasets use their Karpathy test split, and retrieval is reported at Recall@1 in each direction over the full test set, with no re-ranking.

A.2 Per-dataset results

The four sections below give the per-dataset breakdown of the four summary columns of Table 2, for both of its panels: IN-var in Sec. A.2.1, FG in Sec. A.2.2, MM-LP in Sec. A.2.3 and retrieval in Sec. A.2.4.

A.2.1 Zero-shot on the ImageNet robustness variants

Table 6 breaks the IN-var column down over ImageNet-R, ImageNet-Sketch and ImageNetV2. Each is scored over the subset of ImageNet-1k classes it covers, so the three columns do not cover the same class set and are best compared within a column.

Table 6: **Zero-shot top-1 on three ImageNet robustness variants.** Per-dataset breakdown of the IN-var column of Table 2. Best per column and panel in **bold**, second underlined.

Method	IN-R	IN-Sketch	IN-V2	Avg
CC12M, 50k steps				
CLIP	35.5	17.9	22.7	25.4
CLIP†	49.6	28.2	33.3	37.0
SigLIP	32.0	16.4	21.9	23.4
SLIP	42.0	22.1	29.1	31.1
TIPS	49.2	26.4	31.8	35.8
CoMM	35.7	23.6	28.9	29.4
CoMM-VL	**52.4**	**30.5**	**35.4**	**39.4**
LAION-COCO, 100k steps				
CLIP	68.2	44.8	49.1	54.0
SigLIP	69.7	45.9	50.1	55.2
SLIP	65.7	40.8	49.1	51.9
TIPS	71.2	47.0	50.8	56.4
CoMM	45.2	31.7	33.7	36.9
CoMM-VL	**73.3**	**49.7**	**51.2**	**58.1**

A.2.2 Zero-shot on the fine-grained datasets

Table 7 breaks the FG column down over its nine datasets. The gains are not carried by a few datasets: on CC12M, CoMM-VL is best on all nine.

A.2.3 Multimodal linear probe

Table 8 breaks the MM-LP column down over its five tasks. The average is not driven by a single one, but Fashion-Gen is close to saturation for every model at both scales and so compresses the differences the other four show: on CC12M the spread between the weakest and the strongest row is 1.7 points on Fashion-Gen against 9.2 on Hateful Memes and 9.8 on MM-IMDb.

A.2.4 Image-text retrieval

Table 9 breaks the i2t and t2i columns down over MS-COCO and Flickr30k. Recall is uniformly higher on Flickr30k, whose Karpathy test gallery holds a fifth of the images, but the two datasets separate the methods the same way: the two strongest rows of a panel are the same pair on both

Table 7: **Zero-shot top-1 on nine fine-grained classification datasets.** Per-dataset breakdown of the FG column of Table 2. Best per column and panel in **bold**, second underlined.

Method	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN397	DTD	UCF101	Avg
CC12M, 50k steps										
CLIP	56.9	45.0	10.9	25.3	26.7	1.9	30.1	11.6	24.6	25.9
CLIP†	71.7	47.8	16.4	30.5	36.1	2.7	36.4	17.9	27.2	31.8
SigLIP	52.5	38.7	8.2	22.7	22.3	2.6	27.2	10.8	22.1	23.0
SLIP	67.9	45.6	13.6	25.7	34.6	1.6	36.1	13.1	27.6	29.5
TIPS	70.3	54.5	17.2	27.7	38.5	1.7	38.3	13.1	28.6	32.2
CoMM	77.6	45.7	9.8	25.8	36.1	3.1	44.0	11.7	32.9	31.9
CoMM-VL	**81.2**	**61.5**	**20.4**	**33.7**	**42.8**	**3.2**	**46.9**	**19.0**	**33.6**	**38.0**
LAION-COCO, 100k steps										
CLIP	91.8	**82.3**	77.8	**55.8**	70.7	8.9	62.7	47.5	56.4	61.6
SigLIP	93.0	82.2	79.1	53.8	70.3	**11.6**	62.2	44.2	55.8	61.4
SLIP	92.9	78.5	77.6	53.7	71.1	9.2	62.4	46.1	55.7	60.8
TIPS	**93.6**	81.1	78.0	55.5	**72.8**	9.8	62.1	43.7	**60.8**	61.9
CoMM	84.3	45.0	57.6	30.6	48.2	8.5	51.9	42.3	44.4	45.9
CoMM-VL	92.3	82.0	**81.7**	53.8	72.5	9.8	**62.9**	**47.7**	57.9	**62.3**

Table 8: **Per-dataset multimodal linear-probe results.** Per-task breakdown of the MM-LP column of Table 2. Metrics: AUROC on Hateful Memes, macro-F1 on MM-IMDb, top-1 accuracy on SNLI-VE and Fashion-Gen, soft VQA accuracy over the 3,129-answer vocabulary on VQAv2. Every row is probed on the concatenated 1024-d $[z_I; z_T]$ representation except CoMM-VL, which is probed in its 512-d fused space. Best per column and panel in **bold**, second underlined.

Method	Fashion-Gen	Hateful Memes	MM-IMDb	SNLI-VE	VQAv2	Avg
CC12M, 50k steps						
CLIP	95.6	56.2	40.6	56.8	51.4	60.1
CLIP†	95.7	59.4	44.8	55.3	51.8	61.4
SigLIP	94.7	56.7	38.6	56.2	50.8	59.4
SLIP	96.0	58.1	43.7	56.6	53.2	61.5
TIPS	95.8	55.2	45.2	56.5	53.3	61.2
CoMM	**96.4**	63.3	**48.4**	59.2	**54.6**	**64.4**
CoMM-VL	94.8	**64.4**	48.1	**60.6**	52.5	64.1
LAION-COCO, 100k steps						
CLIP	97.3	63.8	54.7	60.3	56.2	66.5
SigLIP	97.0	65.2	54.0	60.5	56.2	66.6
SLIP	97.3	64.1	55.1	60.1	56.2	66.5
TIPS	97.2	62.9	54.7	60.5	**56.5**	66.4
CoMM	**97.6**	**66.2**	**55.7**	61.0	55.4	**67.2**
CoMM-VL	97.1	65.5	54.6	**62.3**	55.4	67.0

datasets and in both directions, and CoMM's weakness on cross-modal tasks is as visible on one as on the other (29.1 i2t on MS-COCO and 50.2 on Flickr30k at LAION scale, against CoMM-VL's 52.0 and 79.2).

A.3 Additional analyses

A.3.1 CLIP and CoMM-VL at approximately matched parameter count

CoMM-VL adds a fusion module on top of the two CLIP towers, so part of its advantage over CLIP could come from the extra parameters rather than from where the alignment loss is applied.

Table 9: **Per-dataset image-text retrieval.** Recall@1 behind the i2t and t2i columns of Table 2, which report the mean of the two datasets. i2t is image-to-text, t2i text-to-image, both over the full test set of each dataset with no re-ranking. Best per column and panel in **bold**, second underlined.

Method	MS-COCO		Flickr30k		Average	
	i2t	t2i	i2t	t2i	i2t	t2i
CC12M, 50k steps						
CLIP	22.9	15.0	44.8	29.9	33.8	22.4
CLIP†	29.4	20.7	53.6	38.4	41.5	29.5
SigLIP	20.9	13.5	38.4	26.9	29.7	20.2
SLIP	32.5	21.3	57.4	40.8	44.9	31.1
TIPS	35.2	23.9	62.5	45.6	48.8	34.8
CoMM	23.1	16.5	44.2	31.5	33.7	24.0
CoMM-VL	**36.4**	**25.7**	**63.8**	**47.4**	**50.1**	**36.5**
LAION-COCO, 100k steps						
CLIP	49.6	33.2	78.3	59.0	63.9	46.1
SigLIP	50.7	34.1	77.2	60.3	63.9	47.2
SLIP	50.5	32.8	75.7	58.3	63.1	45.5
TIPS	**52.0**	**35.0**	79.1	**61.5**	**65.6**	**48.2**
CoMM	29.1	17.9	50.2	32.4	39.6	25.1
CoMM-VL	**52.0**	34.8	**79.2**	60.9	**65.6**	47.8

Table 10: **A deeper CLIP against CoMM-VL on LAION-COCO.** Columns as in Table 2, best in **bold**.

	Zero-shot			LP	MM-LP	Retrieval	
Method	IN-1k	IN-var	FG	top-1	avg	i2t	t2i
CLIP + 2 blocks	57.4	54.8	61.4	75.5	66.6	64.0	46.5
CoMM-VL	**59.0**	**58.1**	**62.3**	**75.7**	**67.0**	**65.6**	**47.8**

We therefore train a deeper CLIP, with one additional transformer block in the vision tower and one in the text tower, under the same recipe and for the same 100k steps on LAION-COCO. The two blocks add 10.2M parameters to CLIP's 149.6M, more than the 7.0M of CoMM-VL's fusion module, so the control is not the smaller model. Table 10 reports both. The extra depth buys CLIP little: against the CLIP row of Table 2 it gains 0.3 points of ImageNet zero-shot accuracy and 0.1 on the multimodal probe, and it loses 0.2 on fine-grained zero-shot. CoMM-VL leads it on all seven metrics, by 1.6 points on ImageNet zero-shot, 3.3 under distribution shift and 1.6 on image-to-text retrieval. Capacity therefore does not account for the gap.