ITO: Multi-View Alignment and Training-Time Fusion for Image-Text Pretraining

Hanpeng Liu[1,2*] **Yaqian Li**[2†] **Zidan Wang**[1] **Shuoxi Zhang**[3] **Zonglin Zhao**[1]
Zihao Bo[2] **Rinyoichi Takezoe**[2] **Kaiwen Long**[2] **Kun He**[1†]

[1]School of Computer Science and Technology,
Huazhong University of Science and Technology, Wuhan, China
[2]Li Auto Inc.
[3]Institute of AI for Industries, Chinese Academy of Sciences
liyaqian@lixiang.com, brooklet60@hust.edu.cn

Abstract

Image–text contrastive pretraining has become a dominant paradigm for visual representation learning, yet existing methods often yield representations that remain partially organized by modality rather than by semantics. We propose **ITO**, a framework addressing this limitation through two complementary mechanisms with distinct roles. *Multimodal multiple alignment* enriches supervision by constructing diverse cross-modal correspondences from multi-view image augmentations, providing the primary source of discriminative gain. A lightweight *training-time multimodal fusion* module then acts as a training-time regularizer, encouraging the encoders to produce features that are compatible under fusion. Crucially, the fusion module is discarded at inference, preserving the efficiency of standard dual-encoder architectures while incurring training-time overhead only. Extensive experiments across pretraining scales from millions to billions of image–text pairs show that ITO outperforms strong contrastive baselines at moderate scale and yields consistent gains over CLIP under identical data, backbone, and compute at the 100M–1B scale, on classification, retrieval, and multimodal benchmarks. Controlled comparisons show that these gains are not explained by additional training compute or image augmentation alone. Our analysis further reveals that the contribution of fusion grows with data scale and that it stabilizes optimization, mitigating the late-stage overfitting commonly observed in aggressive contrastive learning. Code is available at https://github.com/showstarpro/ITO.

1 Introduction

Recent foundation models such as CLIP [41, 9] have fundamentally reshaped visual representation learning through large-scale image–text contrastive pretraining, demonstrating strong transferability across zero-shot classification, retrieval, and as visual backbones for multimodal large language models [33]. A growing body of follow-up work [37, 15, 50] has further strengthened alignment quality and scaling behavior through improved data curation, alternative objectives, and architectural refinements, establishing image–text contrastive learning as a dominant paradigm for large-scale visual pretraining.

Despite this success, we identify a fundamental limitation that has received insufficient attention: *strong pairwise alignment does not imply that the representation space is organized by semantics*

*Work done as a research intern at Li Auto Inc.
†Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

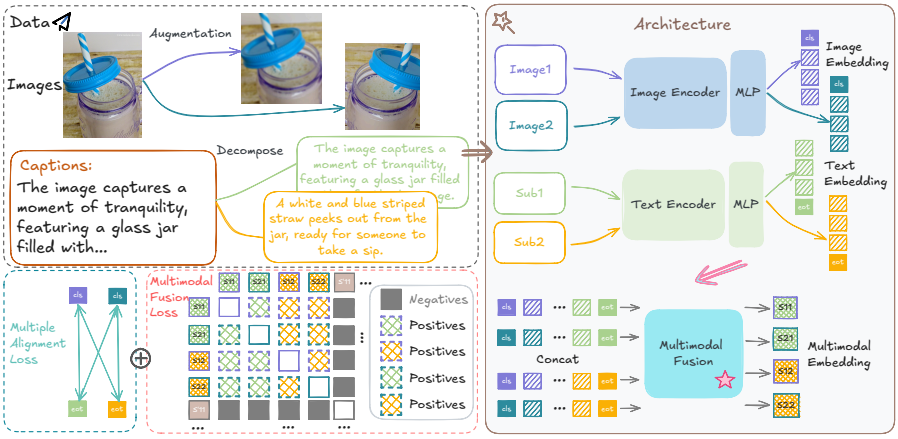


Figure 1: **Overview of the proposed ITO training framework.** Starting from standard image–text contrastive pretraining, ITO restructures supervision through multimodal multiple alignment and introduces a lightweight multimodal fusion module during training. Multiple augmented image–text pairs derived from the same sample are used to enrich instance-level alignment, while training-time fusion encourages encoders to produce features that are compatible under fusion, reducing modality-induced separation. Importantly, the fusion module is used only during training and is discarded at inference time, allowing ITO to retain a standard dual-encoder architecture for efficient deployment.

rather than by modality. Contrastive objectives optimize instance-level matching between paired images and texts, but place no explicit constraint on the global geometry of the resulting embedding space. As a consequence, image and text embeddings can remain organized by modality even when retrieval and classification metrics are strong—a tendency we term *modality-induced separation*: cross-modal semantic neighbors (an image and a text of the same concept) are not reliably closer than cross-modal non-neighbors. This notion is related to but distinct from the *modality gap* [31], a mean offset between the two modality clusters that can persist even when semantics are well organized, and from embedding anisotropy, which is a unimodal property of how features spread. We emphasize that reducing modality-induced separation does not mean collapsing the two modalities into a single geometric cluster; rather, it means organizing the shared embedding space so that semantically similar samples are close regardless of modality, while unrelated samples remain well separated.

This distinction—between *aligning* paired instances and *organizing* the representation space by semantics—is the central concern of this work. Prior work on cross-modal fusion [14, 15, 50] has explored richer multimodal interaction, but as part of the inference architecture, conflating the goal of improving representation structure with the cost of modifying the inference pipeline. A cleaner approach uses cross-modal interaction purely as a *training signal* to reshape the dual encoders, and then discards it at inference; to our knowledge, this has not been systematically explored.

This motivates **ITO**, an image–text contrastive pretraining framework built around a unified principle: *instance-level alignment and global representation organization require distinct, dedicated mechanisms.* ITO addresses each through a dedicated component. First, *Multimodal multiple alignment* restructures the cross-modal supervision signal itself: rather than pairing each image with a single caption, it constructs many-to-one (or many-to-many) correspondences from multiple augmented image views against the same caption within each batch. This densified supervision is the primary engine of discriminative gain. Second, a lightweight *training-time multimodal fusion* module acts as a training-time regularizer on the encoder representations: by applying a contrastive objective over fused multimodal tokens, it propagates gradients back to the dual encoders, which must then produce image and text tokens that remain discriminative and mutually consistent after cross-modal attention. Crucially, the fusion module is discarded entirely at inference, so ITO retains the full efficiency and scalability of a standard dual-encoder architecture. This makes ITO a drop-in replacement in settings where inference cost is the binding constraint, such as large-scale image–text retrieval and embedding serving, where cross-modal interaction at query time is prohibitive but extra one-off training compute is acceptable. The overall framework is illustrated in Figure 1.

The distinguishing features of ITO relative to prior art are two-fold. First, whereas SLIP [37] introduces intra-modal visual augmentation as an auxiliary objective and LaCLIP [17]/VeCLIP [28] expand textual views, all while preserving the one-to-one cross-modal pairing structure, ITO restructures

the cross-modal supervision itself through many-to-one correspondences—a qualitatively different form of enrichment. Second, whereas inference-time fusion architectures such as ALBEF [29], FLAIR [50], and FIBER [14] couple cross-modal interaction with deployment, and knowledge distillation [26] requires a separately trained teacher, ITO uses fusion exclusively as a training-time regularizer with no inference-time role and no teacher, separating the goal of representation shaping from the constraint of inference efficiency.

Our experiments reveal a clear division of labor. Multi-view alignment contributes the dominant share of discriminative gain across all data scales, and this gain is not explained by extra compute or augmentation: CLIP trained with twice the compute recovers less than one point of the gap, and ITO exceeds SLIP, which uses multiple image views under one-to-one supervision (Section 4.7). Training-time fusion plays a complementary role whose direct accuracy contribution is modest; we therefore ground it in optimization evidence rather than in a geometric claim: its gain grows with data scale, it suppresses the late-stage overfitting that alignment-only training still exhibits, and its effect is non-monotone in the fusion weight, as expected of a regularizer. Extensive evaluations across zero-shot classification [2, 11, 47, 48], image–text retrieval, and multimodal benchmarks confirm these gains. They come at a non-trivial training-time cost (Table 7), but with no inference-time overhead.

2 Related Work

Image–Text Contrastive Learning. CLIP [41] established image–text contrastive learning as a dominant paradigm for visual representation learning, with follow-up work improving data curation, training objectives, and scaling behavior [54, 51, 19, 55]. The visual encoders learned this way transfer effectively to zero-shot classification, retrieval, and as backbones for multimodal large language models [20]. Despite this success, most approaches focus on instance-level pairwise alignment and do not explicitly constrain how representations are globally organized in the shared embedding space, motivating recent efforts to examine deeper structural properties of learned representations.

Strengthening Alignment via View Augmentation. A line of work enriches contrastive supervision by augmenting individual modality views. SLIP [37] adds intra-modal visual augmentations as an auxiliary SimCLR-style objective, but the cross-modal contrastive signal still pairs each image with one caption. LaCLIP [17], VeCLIP [28], and DreamLIP [56] expand textual views using large language or vision–language models, again preserving instance-level one-to-one cross-modal supervision. CyCLIP [19] adds geometric consistency constraints on top of CLIP's pairing without altering the supervision structure. Multi-crop and multi-positive objectives [7, 8] draw their positives from within a modality (crop $\leftrightarrow$ crop of the same image) or from existing pairs, and densify single-modality structure. In contrast, our *multi-view alignment* restructures the *cross-modal* supervision itself: multiple augmented image views are paired against the same caption inside the image–text InfoNCE to form many-to-one (or, optionally, many-to-many) cross-modal correspondences within each batch. This is a change to the supervision structure rather than to the augmentation set, and we show in Section 4.7 that its benefit is not reproduced by adding image views under one-to-one supervision.

Cross-Modal Fusion and Auxiliary Training Signals. Another line of work introduces cross-modal interaction during pretraining. ALBEF [29], FIBER [14], and related architectures incorporate cross-attention or joint encoding modules that remain active at inference time. AlignCLIP [15] reduces the modality gap [31] by sharing parameters across encoders, while FLAIR [50] uses text-conditioned pooling that modifies the inference-time encoder behavior. In all of these, fusion is tightly coupled with the inference architecture or with task-specific designs. Knowledge distillation [26] also shapes a deployed model with an auxiliary network, but requires a separately trained teacher whose outputs the student imitates. ITO instead uses a fusion module *exclusively as a training signal*, with neither an inference-time role nor a teacher: gradients from a multimodal contrastive objective on fused tokens propagate back to reshape the dual encoders, after which the fusion module is discarded. To our knowledge, this combination—cross-modal multi-view supervision with a train-only fusion regularizer—has not been explored in prior image–text contrastive learning, and it allows ITO to retain a standard CLIP-style dual-encoder architecture at deployment.

3 Method

3.1 Preliminaries: CLIP-style Contrastive Pretraining

We build upon the standard image–text contrastive learning framework introduced by CLIP. Given a batch of B image–text pairs $\{(I^n, T^n)\}_{n=1}^{B}$, an image encoder E_I and a text encoder E_T are used to extract visual and textual representations, which are then projected into a shared embedding space via projection heads P_I and P_T:

$$Y^n = P_I(E_I(I^n)), \qquad Z^n = P_T(E_T(T^n)). \tag{1}$$

Image–text alignment is achieved through a symmetric InfoNCE objective. Specifically, the image-to-text and text-to-image losses are defined as

$$\mathcal{L}_{I\to T} = -\sum_{n=1}^{B} \log \frac{\exp\left(\langle Y^n, Z^n\rangle/\tau\right)}{\sum_{m=1}^{B} \exp\left(\langle Y^n, Z^m\rangle/\tau\right)}, \tag{2a}$$

$$\mathcal{L}_{T\to I} = -\sum_{n=1}^{B} \log \frac{\exp\left(\langle Z^n, Y^n\rangle/\tau\right)}{\sum_{m=1}^{B} \exp\left(\langle Z^n, Y^m\rangle/\tau\right)}. \tag{2b}$$

, where $\langle\cdot\rangle$ denotes cosine similarity and τ is a learnable temperature. The overall CLIP loss is given by

$$\mathcal{L}_{CLIP} = \frac{1}{2}(\mathcal{L}_{I\to T} + \mathcal{L}_{T\to I}). \tag{3}$$

This dual-encoder formulation enables efficient and scalable inference, which we preserve throughout this work.

3.2 Multimodal Multiple Alignment

To enrich contrastive supervision beyond one-to-one image–text pairing, we introduce Multimodal multiple alignment, which restructures the supervision signal within a training batch. Instead of treating each image–text pair as a single positive instance, this strategy exposes the model to multiple image–text correspondences derived from the same underlying sample, enabling more flexible instance-level alignment.

Concretely, for each original image–text pair (I_i^n, T_j^n), multiple image–text combinations are constructed through standard image and text perturbations, resulting in a set of augmented pairs $\{(I_i^n, T_j^n)\}$. By default, we use two image views ($i \in \{1, 2\}$) and a single text view ($j = 1$)—the configuration we recommend for all large-scale experiments. A variant with two text views ($j \in \{1, 2\}$), denoted *ITO_sub2*, is helpful only at moderate data scales. We present the formulation under *ITO_sub2*; the default is the special case $j = 1$.

For each augmented image–text pair (I_i^n, T_j^n), we compute a bidirectional contrastive loss following the CLIP formulation, including image-to-text and text-to-image directions. Importantly, we retain the standard batch-wise negative sampling strategy: the corresponding augmented pair is treated as the positive sample, while all other samples within the batch serve as negatives. This process is repeated for all valid combinations of image and text views, and the final alignment loss is computed as the average over these losses. Formally, the multiple alignment loss is defined as

$$\mathcal{L}_{\text{Align}} = \frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{1}{2}\left[\mathcal{L}_{I_i\to T_j} + \mathcal{L}_{T_j\to I_i}\right], \tag{4}$$

where $\mathcal{L}_{I_i\to T_j}$ and $\mathcal{L}_{T_j\to I_i}$ denote the image-to-text and text-to-image InfoNCE losses, respectively.

By increasing the diversity and number of positive image–text pairings within each batch, Multimodal multiple alignment enriches instance-level supervision and improves alignment robustness without introducing additional inference-time cost. Note that the multiplicity enters the *cross-modal* objective: each image view is a positive for the shared caption, rather than for another crop of the same image as in multi-crop self-supervised learning. However, this strategy alone still scores each image and text independently, and never requires their features to remain consistent under joint processing, which motivates the training-time fusion objective below.

3.3 Training-Time Multimodal Fusion

While multiple alignment strengthens correspondence between individual image–text pairs, it scores the two modalities only through a dot product of pooled embeddings. To add a complementary training signal, we introduce a lightweight training-time multimodal fusion module that acts as a regularizer on the encoders' representations.

Given an augmented image–text pair (I_i^n, T_j^n), we obtain their corresponding visual tokens Y_i^n and textual tokens Z_j^n. These tokens are concatenated to form a joint multimodal sequence:

$$H_{i,j}^n = \mathrm{Concat}(Y_i^n, Z_j^n), \tag{5}$$

where $\mathrm{Concat}(\cdot,\cdot)$ denotes token-wise concatenation. The joint sequence is then processed by a lightweight fusion module F_M, implemented as a two-layer Transformer with bidirectional self-attention over the joint sequence, producing fused multimodal tokens:

$$S_{i,j}^n = F_M(H_{i,j}^n). \tag{6}$$

We use the output at the first position of the joint sequence, i.e., the image `[CLS]` token, as the fused multimodal representation. Since the vision tokens come first and have a fixed length, this is a fixed read-out position that, under bidirectional self-attention, attends over the entire joint sequence, and it reuses the aggregation role `[CLS]` already plays in the ViT image encoder. We compare it with alternative read-outs in Appendix D.

During training, the fusion objective encourages consistency among multimodal representations derived from the same underlying image–text pair, while pushing apart representations from different pairs. Specifically, fused representations obtained from different augmentations of the same image–text pair are treated as positive samples, while fused representations from other samples within the batch serve as negatives. We adopt a contrastive formulation with multiple positives to account for this structure. The loss for a fused representation is defined as:

$$\mathcal{L}_{S_{1,1}^n} = -\log \frac{\exp[\langle S_{1,1}^n \cdot S_{2,1}^n \rangle/\tau] + \sum_{i=1}^{2} \exp[\langle S_{1,1}^n \cdot S_{i,2}^n \rangle/\tau]}{\sum_{m=1}^{B} \sum_{i=1}^{2} \sum_{k=1}^{2} \exp[\langle S_{1,1}^n \cdot S_{i,k}^m \rangle/\tau]}, \tag{7}$$

where the trivial self-pair $(m = n, i = 1, k = 1)$ is excluded. The overall fusion loss is:

$$\mathcal{L}_{\mathrm{Fusion}} = \frac{1}{4B} \sum_{n=1}^{B} \sum_{i=1}^{2} \sum_{j=1}^{2} \mathcal{L}_{S_{i,j}^n}. \tag{8}$$

Gradient path. The fusion module receives the image and text tokens without stop-gradient, and the fused `[CLS]` representation is a self-attention function of both. The fusion loss therefore backpropagates through F_M into both encoders. Because F_M has no private input path, it cannot reduce $\mathcal{L}_{\mathrm{Fusion}}$ on its own: the encoders must produce image and text tokens that remain discriminative and mutually consistent after cross-modal self-attention. In this way $\mathcal{L}_{\mathrm{Fusion}}$ acts as a training-time regularizer: it encourages features that are not only linearly separable (as in vanilla contrastive learning) but also compatible under fusion, while preserving the dual-encoder architecture. We characterize its effect through optimization behavior rather than through a claim about global geometry (Section 4.8).

3.4 Overall Objective and Inference

The final training objective combines the multiple alignment loss and the multimodal fusion loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{Align}} + \lambda \mathcal{L}_{\mathrm{Fusion}}. \tag{9}$$

Here, λ balances the trade-off between discriminative intensity (from alignment) and fusion regularization. At inference time, ITO reduces to a standard dual-encoder model identical to CLIP, with no fusion modules and no additional inference-time overhead, enabling direct replacement of existing image–text contrastive encoders. The improved representation quality, however, comes at a non-trivial training-time cost (approximately $2.2\times$ training time and 2.1–$2.3\times$ peak memory compared to OpenCLIP), reported in Table 7.

4 Experiments

4.1 Implementation Details

Datasets. We evaluate our method across image–text datasets spanning from millions to billions of samples in order to assess effectiveness, robustness, and scalability. Specifically, we conduct pretraining on CC3M [43], CC12M [4], and YFCC15M [46], which are widely used benchmarks for image–text contrastive pretraining. To study large-scale behavior, we further perform experiments on Laion100M (a 100M subset of Laion400M [42]) and the billion-scale DataComp-1B [18] dataset. Due to computational constraints, large-scale baselines are limited: on Laion100M we compare against CLIP and SLIP, and on DataComp-1B only against CLIP. Billion-scale reruns of SLIP, SigLIP, and FLAIR are beyond our budget, and public checkpoints trained on a matched corpus are not available. At these scales, CLIP therefore serves as a controlled reference that holds data, backbone, and schedule fixed. Detailed dataset statistics are provided in appendix Table 8.

Data Augmentation. We adopt standard image augmentations following prior work (e.g.,MoCo v3 [8] and SLIP [37]) to generate two image views per sample; the exact augmentation policy is given in Appendix B.1. For text, we consider two settings: a default configuration using a single caption, and an optional variant that samples two sub-descriptions from the original text. The latter is denoted as ITO_sub2. Sub-description sampling (ITO_sub2) is applied only to CC3M, CC12M, YFCC15M, and DataComp-1B experiments; sub-descriptions are generated following FLAIR [50] (Appendix H). These augmentations are used solely to construct multiple views for contrastive supervision.

Pretraining Settings. All models are implemented based on OpenCLIP [9]. Unless otherwise specified, we use ViT-B/16 as the vision encoder and the standard CLIP text encoder. Images are resized to 224×224, and text sequences are tokenized to a maximum length of 77 tokens. Models are optimized using AdamW with a cosine learning rate schedule and linear warmup. We train models for 30 epochs on CC3M, CC12M, YFCC15M and Laion100M. For DataComp-1B, models are trained for one epoch due to computational cost; an additional 10-epoch ViT-B/16 model is trained to study data scaling effects. Batch size and learning rate are adjusted according to the dataset scale. The fusion loss weight λ is set to 2 by default. All other hyperparameters follow OpenCLIP defaults. All training is conducted on NVIDIA A800 GPUs. To ensure fair comparison, we reproduce these methods using the versions of the datasets we have downloaded.

Evaluation Protocol. We evaluate pretrained models using only the image and text encoders, without any fusion modules, to assess representation quality under a standard dual-encoder setting. Four categories of downstream tasks are considered.

Zero-shot Image Classification. We follow the evaluation protocol of EVA-CLIP [45] and report top-1 accuracy on 26 datasets covering generic, fine-grained, and domain-specific classification tasks.

Linear Image Classification & Retrieval. We perform linear probing on frozen visual features and bidirectional zero-shot retrieval on COCO [32], Flickr30K [40], and DOCCI [38] (Recall@1/5/10).

Vision-Language Understanding. To assess transferability to multimodal systems, we follow the official LLaVA-1.5 [33] evaluation protocol on 13 benchmarks, including VQAv2 [21], MM-Vet [52], POPE [30], and MMStar [6]. These experiments evaluate the pretrained visual encoders as backbones rather than introducing new VLM architectures.

All baselines are reproduced and evaluated under the same pipeline for fair comparison and reproducibility.

4.2 Zero-shot Image Classification

We evaluate zero-shot image classification on 26 public benchmarks following the EVA-CLIP [45]. Average top-1 accuracy is reported in Table 1; full per-benchmark results are deferred to Appendix Table 12.

At moderate scale (CC3M, CC12M, YFCC15M), ITO outperforms CLIP and prior contrastive baselines including SLIP [37], SigLIP [54], and FLAIR [50]. At the 100M and 1B scales, where CLIP is the only baseline run at every scale (SLIP is additionally available on Laion100M), ITO yields consistent gains over CLIP under identical data, backbone, and compute. The benefits of

Table 1: Average top-1 zero-shot classification accuracy across 26 public benchmarks following the EVA-CLIP [45] protocol. All models use ViT-B/16 and are trained for 30 epochs unless otherwise noted. The upper block compares ITO against contrastive baselines under the same backbone and training schedule; the lower block reports scaling behavior with longer training (10 epochs) and a larger backbone (ViT-L/16) on DataComp-1B. Due to computational constraints, SigLIP and FLAIR are evaluated only on smaller datasets and SLIP up to Laion100M; CLIP is the only baseline run consistently across all pretraining scales, and at large scale the comparison is against CLIP under identical data, backbone, and compute. Per-benchmark results are provided in Appendix Table 12. Best results are in **bold**.

Pretraining Data	CLIP	SLIP	SigLIP	FLAIR	ITO	ITO_sub2
CC3M	19.3	21.4	19.3	23.2	**24.9**	**24.9**
CC12M	34.1	36.0	35.0	35.5	38.4	**40.4**
YFCC15M	30.5	–	–	–	34.5	**35.4**
Laion100M	53.5	51.3	–	–	**56.1**	–
DataComp-1B	54.2	–	–	–	56.6	**57.0**
Scaling: longer training and larger backbones on DataComp-1B						
ViT-B/16, 10 epochs	63.7	–	–	–	**64.9**	63.2
ViT-L/16, 1 epoch	60.7	–	–	–	**63.0**	61.5

Table 2: Linear probing top-1 accuracy of ITO and CLIP across pretraining scales. We report ImageNet-1K (IN-1k) and the average over five datasets (IN-1k, Flowers-102, Stanford Cars, Pets, Food-101). Per-dataset results, additional baselines, and ITO_sub2 variants are provided in Appendix Table 13. Best results in each block are in **bold**.

	YFCC15M		Laion100M		DataComp-1B (1ep)		DataComp-1B (10ep)	
Method	IN-1k	Avg	IN-1k	Avg	IN-1k	Avg	IN-1k	Avg
CLIP	51.82	55.88	67.27	81.86	67.39	82.14	74.99	87.62
ITO	**60.52**	**63.76**	**68.70**	**83.56**	**69.87**	**82.53**	**76.20**	**89.18**

multi-view alignment and training-time fusion are thus not confined to a single data regime, although we do not claim superiority over stronger baselines at billion scale.

We further analyze the effect of textual diversity through sub-description sampling. ITO_sub2 is not part of the default ITO, and the headline results do not depend on it. Its behavior reflects the data regime. On CC3M, CC12M, and YFCC15M, where native captions are short and sparse, sub-descriptions inject textual diversity the captions lack, and ITO_sub2 matches or exceeds the default ITO. On DataComp-1B, where native captions are already lexically rich, sub-descriptions become largely redundant paraphrases and act as noise: ITO_sub2 falls slightly below both ITO and CLIP at 10 epochs (63.2 vs. 63.7 for CLIP). We therefore adopt the default ITO configuration for all large-scale experiments. The scaling block of Table 1 further shows that longer training (10 epochs) and a larger backbone (ViT-L/16) both continue to improve ITO, with the gap to CLIP preserved or widened.

4.3 Linear Image Classification

We further evaluate the quality of learned visual representations via linear probing. Following standard practice, we train a linear classifier on top of frozen visual features for 30 epochs using AdamW with a cosine learning-rate schedule. Top-1 accuracy on ImageNet-1K and the average over five classification datasets are reported in Table 2 for medium- to billion-scale pretraining; results for CC3M and CC12M are visualized in Appendix Figure 2 and Figure 3, with full per-dataset numbers and additional baselines deferred to the appendix.

Across all pretraining scales, ITO consistently outperforms CLIP under linear evaluation, indicating improved linear separability of the learned visual representations. The gap is largest at medium scale (e.g., +7.9 average points on YFCC15M) and remains consistent at billion scale (+1.6 points on DataComp-1B with 10-epoch training). Importantly, the linear-probing advantage holds even though the fusion module is discarded at inference, showing that ITO training changes the visual encoder itself rather than relying on an inference-time helper. On DataComp-1B, longer training raises ITO

Table 3: Zero-shot image-text retrieval (Recall@1) on MSCOCO and Flickr30k. CC12M and DataComp-1B (ViT-B/16, 1 epoch) represent medium- and large-scale pretraining respectively. Comparisons against additional baselines (SLIP, SigLIP, FLAIR), all five pretraining scales, and full Recall@5/10 metrics are provided in Appendix Table 18. Best results are in **bold**.

	CC12M				DataComp-1B (1 epoch)			
	MSCOCO		Flickr30k		MSCOCO		Flickr30k	
Method	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I
CLIP	34.17	23.04	62.23	45.74	47.08	29.51	72.68	54.73
ITO	42.30	29.62	**72.49**	54.50	49.26	31.30	**75.94**	56.79
ITO_sub2	**43.94**	**30.51**	72.29	**56.65**	**49.50**	**31.89**	73.37	**56.79**

Table 4: The performance of ITO on a broad range of multimodal tasks. The best results are **bold**.

Dataset	Method	VQAv2	GQA	T-VQA	SciQA-I	VisWiz	MMB-en	MMB-cn	MMVet	POPE-r	POPE-p	POPE-a	MMMU	MMStar
100M	CLIP	66.53	56.36	47.55	65.44	36.80	53.69	46.13	18.30	84.09	82.90	76.40	33.60	28.87
	ITO	**68.47**	**57.23**	**48.71**	**66.29**	**44.25**	**54.12**	**46.65**	**19.80**	**84.74**	**83.63**	**77.30**	**33.90**	**29.13**
1B	CLIP	70.42	57.93	50.24	65.00	**45.28**	48.45	55.67	18.40	83.63	82.71	78.63	**34.00**	29.33
	ITO	**73.19**	**59.99**	**50.89**	**66.24**	42.92	**50.52**	**58.85**	**21.90**	**85.46**	**84.23**	**80.46**	33.90	**31.13**

from 82.53 (1 epoch) to 89.18 (10 epochs) while preserving the gain over CLIP; full results are in the appendix.

4.4 Zero-shot Image-Text Retrieval

We evaluate zero-shot image-text retrieval to assess cross-modal alignment quality on MSCOCO [32] and Flickr30k [40]. Bidirectional Recall@1 is reported in Table 3; results across all five pretraining scales, full Recall@5/10, and fine-grained retrieval on DOCCI [38] are provided in Appendix Tables 18 and Table 19.

Two findings are consistent across all retrieval settings. First, ITO outperforms CLIP, SigLIP, and FLAIR on both retrieval directions and both datasets, indicating stronger cross-modal alignment under the standard dual-encoder protocol. Second, the role of textual diversity differs sharply across data scales: at medium scale (CC12M), sub-description sampling (ITO_sub2) yields additional gains of 1.6–2.2 R@1 points, whereas at billion scale this advantage attenuates as the abundant native variation in web captions already saturates the textual supervision signal. Since retrieval depends directly on cross-modal similarity, these results indicate that the benefits of ITO extend beyond same-modality discriminability (Section 4.3) to cross-modal alignment.

4.5 Transfer to MLLM Benchmarks

Many multimodal large language models rely on a pretrained vision encoder as a fixed perceptual backbone. To evaluate the transferability of learned visual representations, we follow the LLaVA-1.5 [33] protocol and integrate Vicuna-7B [10] with vision encoders pretrained using CLIP and ITO, all based on ViT-B/16. Unless otherwise specified, the vision encoder is frozen during multimodal fine-tuning, isolating the effect of visual pretraining.

We evaluate models pretrained on Laion100M and on the 10-epoch version of DataComp-1B across 13 Multimodal Large Language Model benchmarks, including VQAv2 [21], GQA [1], T-VQA [44], SciQA [35], VizWiz [22], MMBench [34], MMVet [52], POPE [30], MMMU [53], and MMStar [6].

As shown in Table 4, vision encoders pretrained with ITO consistently outperform their CLIP counterparts, with the largest improvements on VQAv2, GQA, MMVet, and POPE, and pretraining on DataComp-1B gives the highest overall performance.

4.6 Ablation Study

We conduct ablation experiments on YFCC15M and DataComp-1B using ViT-B/16 to analyze the contribution of each component in ITO.

Table 5: Decomposing the contributions of multi-view alignment and fusion regularization on YFCC15M and DataComp-1B (ViT-B/16, 1 epoch on 1B). Δ values are gains over the previous row. Multi-view alignment provides the dominant gain at both scales; fusion regularization adds a smaller but consistent improvement that grows at billion scale.

Dataset	Setting	ZS IN-1k	Linear IN-1k	MSCOCO I→T	MSCOCO T→I
15M	OpenCLIP	36.4	51.8	26.4	15.1
	+ Multi-view alignment (λ=0)	43.7 (+7.3)	60.2 (+8.4)	30.6 (+4.2)	19.6 (+4.5)
	+ Fusion regularization (λ=2)	**44.3** (+0.6)	**60.5** (+0.3)	**30.8** (+0.2)	19.5 (−0.1)
1B	OpenCLIP	63.1	67.4	47.1	29.5
	+ Multi-view alignment (λ=0)	64.8 (+1.7)	69.6 (+2.2)	49.0 (+1.9)	30.7 (+1.2)
	+ Fusion regularization (λ=2)	**65.9** (+1.1)	**69.9** (+0.3)	**49.3** (+0.3)	**31.3** (+0.6)

Table 6: Ablation study on YFCC15M. We analyze the effect of Multimodal Multiple Alignment and training-time multimodal fusion by varying the fusion weight λ (left) and the number of Sentence Blocks (right). Results are zero-shot Top-1 accuracy (%) on ImageNet-1K using ViT-B/16. The baseline is standard CLIP.

Method	λ	Top-1
Baseline	-	36.4
ITO	0	43.7
ITO	2	**44.3**
ITO	4	44.0
ITO	6	43.2
ITO	8	43.6

Method	Blocks	Top-1
Baseline	-	36.4
ITO	1	43.9
ITO	2	**44.3**
ITO	3	43.7
ITO	4	43.5
ITO	5	44.0

Decomposing the contributions of multi-view alignment and fusion regularization. Table 5 cleanly separates the two components by progressively adding them to a CLIP baseline at both medium (15M) and large (1B) data scales. Multi-view alignment alone (ITO with λ=0) accounts for the dominant share of the improvement: at YFCC15M it adds $+7.3$ on zero-shot ImageNet-1K, $+8.4$ on linear probing, and $+4.2/+4.5$ on MSCOCO retrieval; the corresponding gains at DataComp-1B are $+1.7/+2.2/+1.9/+1.2$. On top of this, fusion regularization (λ=2) adds a smaller but consistent gain in all four metrics, and—notably—this gain *grows* with data scale: zero-shot accuracy improves by $+0.6$ at 15M and by $+1.1$ at 1B. The accuracy contribution of fusion is thus modest in absolute terms, and at 15M it slightly lowers MSCOCO T→I (-0.1). We therefore do not present fusion as a primary source of accuracy; in Section 4.8 we characterize it as an optimization regularizer.

Sensitivity to fusion weight and depth. We further sweep the fusion loss weight λ and the number of fusion blocks (Table 6). Accuracy follows an inverted U in λ: it peaks at λ=2 (44.3), and an over-strong fusion signal falls *below* the no-fusion setting (λ=6: 43.2 vs. λ=0: 43.7). This is the signature of a regularizer with a bounded optimum: moderate fusion helps, whereas a dominant fusion signal overrides modality-specific structure the dual encoder still needs. A shallow two-block design suffices: deeper fusion modules yield no additional gain, consistent with fusion serving as a training-time regularizer rather than a deep reasoning module. We further explored cross-attention as an alternative fusion architecture and found it consistently underperforms self-attention (Appendix C); alternative fusion read-outs are compared in Appendix D.

4.7 Are the Gains from Compute or Augmentation?

Multi-view alignment processes two image views per caption, so ITO costs about $2.2\times$ the training time of OpenCLIP (Table 7, right); the fusion regularizer adds negligible cost, and inference cost is unchanged. We test whether the gains come from this extra compute or from the additional augmentation.

Compute. Doubling CLIP's schedule from 30 to 60 epochs improves zero-shot ImageNet-1K by only $+0.9$ (Table 7, left). CLIP at 60 epochs already uses about 87% of the FLOPs of ITO at 30 epochs, yet remains 7.0 points behind. When both are trained for 60 epochs, the gap widens slightly (45.3 vs. 37.3), so the advantage of ITO is not erased by a longer schedule.

Augmentation. SLIP [37] also uses multiple image views but keeps one-to-one cross-modal supervision. On the 26-benchmark zero-shot average (Table 1), SLIP improves over CLIP (21.4 vs. 19.3

Table 7: **Left:** compute-matched comparison on YFCC15M (ViT-B/16); FLOPs are training FLOPs relative to CLIP at 30 epochs. **Right:** training time (h) and peak memory (GB/GPU) on 64 A800 GPUs, with ratios to OpenCLIP. Inference cost of ITO is identical to CLIP.

Model (epochs)	FLOPs	ZS	Lin.	T→I
CLIP (30)	1.0×	36.4	51.8	15.1
CLIP (60)	2.0×	37.3	52.8	15.4
ITO, λ=0 (30)	2.3×	43.7	60.2	19.6
ITO, λ=2 (30)	2.3×	44.3	60.5	19.5
ITO, λ=0 (60)	4.6×	44.6	61.7	18.7
ITO, λ=2 (60)	4.6×	45.3	62.1	19.6

Data	Method	Time	Mem.
15M	OpenCLIP	3.6	7.9
	ITO	8.2 (2.3×)	16.8 (2.1×)
1B	OpenCLIP	8.1	30.0
	ITO	∼18 (2.2×)	68.6 (2.3×)

on CC3M; 36.0 vs. 34.1 on CC12M), and ITO exceeds SLIP by a further $+3.5$ and $+2.4$ points. As SLIP also trains a self-supervised image branch, it is a conservative one-to-one reference; the residual margin is attributable to the many-to-one restructuring rather than to more views.

4.8 What Does Training-Time Fusion Do?

Fusion as an optimization regularizer. Three observations, each comparing ITO with λ=2 against the alignment-only variant (λ=0), which differ only in the fusion module, support this interpretation. *(i) Its gain grows with data scale and training length:* zero-shot ImageNet-1K improves by $+0.6$ at 15M and $+1.1$ at 1B, and MSCOCO T→I changes from -0.1 to $+0.6$ (Table 5); on YFCC15M, extending training from 30 to 60 epochs raises the T→I gain from -0.1 to $+0.9$ (Table 7). *(ii) It suppresses late-stage overfitting:* on YFCC15M, CLIP and SLIP degrade in later epochs, and alignment alone raises accuracy without preventing this decline, whereas adding fusion holds zero-shot accuracy flat through the end of training (Appendix Figure 8). *(iii) Its effect is non-monotone in λ:* an over-strong fusion signal falls below the no-fusion setting (Section 4.6), as expected of a regularizer rather than of an additional source of supervision.

Effect on representation geometry. Fusion improves the cross-modal consistency score of CyCLIP [19], which measures the agreement between top-K neighbors retrieved in image and in text space, at every K and at both scales (e.g.,$+0.6$ and $+0.8$ at top-1 on YFCC15M and DataComp-1B; Appendix Table 20). Global quantities, however, tell a less uniform story (Appendix K): fusion narrows the image–text centroid gap [31] at CC3M (centroid cosine $0.423 \rightarrow 0.466$) but widens it at DataComp-1B ($0.336 \rightarrow 0.314$), while cross-modal MMD and the same-/different-class distance ratio change only at the third to fourth decimal. No single global quantity tracks the downstream gains, consistent with the complex relation between the modality gap and downstream performance reported in [31]. We therefore do not claim that fusion works by improving global geometry, and rest its justification on the optimization evidence above.

5 Conclusion

We proposed **ITO**, whose two components play distinct roles. Multi-view alignment restructures cross-modal supervision from one-to-one to many-to-one and provides the dominant share of discriminative gain, which is not explained by extra compute or augmentation. A training-only fusion module acts as an optimization regularizer whose contribution grows with data scale and which suppresses late-stage overfitting. Because the fusion module is discarded at inference, ITO keeps the cost of a standard dual encoder, suiting inference-cost-constrained settings such as large-scale retrieval and embedding serving.

Limitations. ITO requires about $2.2\times$ the training time and 2.1–$2.3\times$ the peak memory of CLIP, almost entirely due to multi-view alignment. At 100M–1B scale we compare mainly against CLIP, so we do not establish superiority over stronger baselines there. The accuracy gain from fusion is modest, and its effect on global geometry is small and scale-dependent; the consistency score is an indirect measure. Results may depend on the augmentation policy and text-view construction (as ITO_sub2 shows). Gains hold across corpora of varying curation (Appendix F), but we have not run a controlled noisy-versus-curated ablation. Finally, an over-strong fusion signal hurts accuracy, so reducing modality-induced separation can conflict with preserving modality-specific information.

Acknowledgments and Disclosure of Funding

This work is supported by National Natural Science Foundation of China (U22B2017), and International Cooperation Foundation of Hubei Province, China (2024EHA032).

References

[1] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: training generalized multi-query transformer models from multi-head checkpoints. In *EMNLP*, pages 4895–4901, 2023.

[2] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *NeurIPS*, 2019.

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014.

[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 *CVPR*, pages 3558–3568, 2021.

[5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, pages 370–387, 2024.

[6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020.

[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9620–9629, 2021.

[9] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023.

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

[11] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

[12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[14] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, Jianfeng Gao, and Lijuan Wang. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *NeurIPS*, 2022.

[15] Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Investigating approaches for improving cross-modal alignment in CLIP. In *ICLR*, 2025.

[16] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015.

[17] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving CLIP training with language rewrites. In *NeurIPS*, 2023.

[18] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M. Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander J. Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. In *NeurIPS 2023*, 2023.

[19] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. In *NeurIPS*, 2022.

[20] 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 *CVPR*, pages 6325–6334, 2017.

[21] 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 *CVPR*, pages 6325–6334, 2017.

[22] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617, 2018.

[23] 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 J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2019.

[24] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *CVPR*, 2021.

[25] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021.

[26] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[27] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[28] Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, and Meng Cao. Veclip: Improving CLIP training via visual-enriched captions. In *ECCV*, pages 111–127, 2024.

[29] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, pages 9694–9705, 2021.

[30] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pages 292–305, 2023.

[31] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022.

[32] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014.

[33] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26286–26296, 2024.

[34] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, pages 216–233, 2024.

[35] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.

[36] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013.

[37] Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. SLIP: self-supervision meets language-image pre-training. In *ECCV*, pages 529–544, 2022.

[38] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. DOCCI: descriptions of connected and contrasting images. In *ECCV*, pages 291–309, 2024.

[39] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

[40] 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 *ICCV*, pages 2641–2649, 2015.

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[42] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114, 2021.

[43] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018.

[44] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, pages 8317–8326, 2019.

[45] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389, 2023.

[46] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528, 2011.

[47] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *MICCAI*, 2018.

[48] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 2019.

[49] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[50] Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. Flair: Vlm with fine-grained language-informed image representations. In *CVPR*, 2025.

[51] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *ICLR*. OpenReview.net, 2024.

[52] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *ICML*, 2024.

[53] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, pages 9556–9567, 2024.

[54] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11941–11952, 2023.

[55] Weiheng Zhao, Zilong Huang, Jiashi Feng, and Xinggang Wang. Superclip: Clip with simple classification supervision, 2025.

[56] Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. In *ECCV*, pages 73–90, 2024.

A Datasets

Table 8: The original and actual downloaded number of image-text pairs in pretraining datasets.

Dataset	CC3M [43]	CC12M [4]	YFCC15M [46]	Laion100M [42]	DataComp-1B [18]
Origin	3.32M	12.42M	15.39M	361.02M	1.40B
Download	2.91M	10.97M	14.08M	107.75M	1.03B

We conduct pretraining experiments on several widely used image–text datasets spanning different scales. All web-scale datasets, including CC3M [43], CC12M [4], Laion100M (a 100M subset of LAION-400M [42]), and DataComp-1B [18], are downloaded following the official `img2dataset` pipeline[3] with default configurations. For YFCC15M, we use the filtered English subset released by [41], which can be obtained via the public downloader[4].

Due to network availability and filtering, the number of successfully retrieved image–text pairs may differ from the nominal dataset size. The exact number of samples used for each dataset is summarized in Table 8.

B Implementation Details

Table 9 summarizes the common hyperparameter settings used for pretraining ITO on the billion-scale DataComp-1B dataset. Unless otherwise specified, these configurations are shared across all ITO variants and baselines to ensure fair comparison.

All models are trained using the OpenCLIP framework. We adopt AdamW as the optimizer with standard momentum parameters and apply cosine learning-rate decay with linear warmup. The fusion loss weight λ is set to its default value as determined in the ablation study, and the fusion module consists of two lightweight Transformer blocks. Image resolution, tokenizer settings, batch size, and training schedules follow the standard CLIP protocol.

For baselines, we strictly follow the officially recommended hyperparameter settings. When reproducing baseline results, we use the same dataset versions and training pipelines as those used for ITO.

B.1 Image Augmentation Policy

Each image is augmented independently twice with the same stochastic policy to produce the two views used in multi-view alignment. The policy applies random resized cropping to 224×224 with scale range $(0.5, 1.0)$, color jitter (brightness 0.4, contrast 0.4, saturation 0.4, hue 0.1) with probability 0.8, and random grayscale conversion with probability 0.2, followed by normalization with the CLIP mean and standard deviation. The CLIP and SLIP baselines use their official augmentation settings.

C Fusion Architecture: Self- vs Cross-Attention

A natural alternative to our self-attention fusion module is cross-attention, where one modality's tokens attend to the other. We compare both designs by replacing the two-layer self-attention block with a two-layer cross-attention block that uses the text EOS token as the query, keeping all other settings unchanged. Table 10 shows that cross-attention consistently underperforms self-attention; on YFCC15M, it actually drops below the no-fusion baseline (λ=0) on zero-shot accuracy.

We attribute this to the asymmetric, compressive nature of single-query cross-attention: routing all cross-modal information through one query token weakens fine-grained bidirectional token mixing. Self-attention over concatenated tokens, in contrast, lets all multimodal tokens interact symmetrically, which is better suited to lightweight structural regularization where no single token should dominate the bottleneck. We therefore adopt self-attention as the default fusion architecture in ITO.

[3] `https://github.com/rom1504/img2dataset`

[4] `https://github.com/AdamRain/YFCC15M_downloader.git`

Table 9: Common ITO hyperparameters on Datacomp-1B.

Config	Value
Optimizer	AdamW
Learning rate	5e-4
Weight decay	0.1
Optimizer momentum	β_1=0.9, β_2=0.98
Batch size	16384(ViT-B), 8192(ViT-L)
Warm-up iterations	500
λ	2
Scale	(0.5,1.0)
Gray_scale_prob	0.2
Color_jitter	[0.4, 0.4, 0.4, 0.1]
Color_jitter_prob	0.8

Table 10: Self-attention vs cross-attention fusion (two-layer, λ=2). Cross-attention with an EOS-query design consistently underperforms self-attention and, on YFCC15M, falls below the no-fusion variant. Best results are in **bold**.

Dataset	Fusion	ZS IN-1k	Linear IN-1k	MSCOCO I→T	MSCOCO T→I
CC3M	no fusion (λ=0)	23.0	—	—	—
	cross-attention	23.0	47.5	19.7	13.6
	self-attention (ours)	**23.3**	**48.1**	**21.6**	**14.6**
YFCC15M	no fusion (λ=0)	43.7	60.2	30.6	19.6
	cross-attention	43.1	59.9	30.8	19.4
	self-attention (ours)	**44.3**	**60.5**	**30.8**	**19.6**

D Fusion Read-out Ablation

The fusion module reads out the output at the first position of the joint sequence, i.e., the image `[CLS]` token, as the fused representation (Section 3.3). The submitted version and our rebuttal described this read-out as the end-of-text (EOT) token; the implementation used for all reported results reads out the image `[CLS]` token, and we correct the description here. No reported number is affected.

We compare this choice with alternative read-outs, changing only the read-out and keeping all other settings unchanged: mean-pooling over all joint tokens, and an additional learnable `[FUSE]` token. The `[FUSE]` token is prepended at the first position of the joint sequence, before the image `[CLS]` token, without positional embedding, and its output is used as the fused representation. These runs use YFCC15M, ViT-B/16, λ=2, and 30 epochs with the same hyperparameters as the main experiments, but were trained on 16 NVIDIA H200 GPUs (global batch size 4096), whereas the main experiments were trained on NVIDIA A800 GPUs. The numbers in Table 11 are therefore comparable with each other but not directly with the main tables.

Table 11: Fusion read-out ablation on YFCC15M (ViT-B/16, λ=2, 30 epochs, trained on H200 GPUs). Zero-shot ImageNet-1K top-1 accuracy (%).

Read-out	ZS IN-1k
Image `[CLS]` token (default)	**45.06**
Mean-pooling over joint tokens	43.90
Learnable `[FUSE]` token	**45.06**

A dedicated learnable token performs on par with the image `[CLS]` token, indicating that the choice between single-token read-outs is not load-bearing. Mean-pooling is 1.2 points lower. We attribute this to the imbalance of the joint sequence: the 196 image patch tokens outnumber the text tokens, so an unweighted average is dominated by visual content and dilutes the cross-modal signal that a single attending token can aggregate.

Table 12: Top-1 zero-shot classification accuracy on 26 public benchmarks following the EVA-CLIP [45] protocol. Benchmarks include ImageNet-1K [13], ImageNet-A [25], ImageNet-R [24], CIFAR-10/100 [27], Food-101 [3], Pets [39], SUN397 [49], FGVC Aircraft [36], EuroSAT [23], VOC2007 [16], etc. The best results are highlighted in **bold**.

	ImageNet-1k	ImageNet-A	ImageNet-R	ImageNet-S	ImageNet-V2	ObjectNet	CIFAR-10	CIFAR-100	Flowers-102	Food-101	Pets	Stanford Cars	MNIST	Caltech	SUN397	FGVC Aircraft	Country-211	DTD	EuroSAT	FER2013	GTSRB	PCam	Rendered SST2	Resisc45	STL10	VOC2007	Avg
(a) *Results on CC3M (ViT-B/16, 30 epochs)*																											
CLIP	14.8	3.2	16.0	6.0	12.7	7.8	45.5	17.8	11.6	9.4	10.4	0.7	10.2	49.8	22.3	1.6	0.5	12.5	6.0	27.3	4.5	55.6	49.9	17.4	70.3	18.5	19.3
SLIP	18.3	5.1	19.6	8.1	15.5	10.5	47.0	19.0	10.9	13.0	10.4	1.2	10.1	56.6	31.2	1.6	0.7	12.6	6.4	18.1	2.8	55.9	50.1	17.5	88.1	27.2	21.4
SigLIP	16.6	3.6	19.6	7.6	14.6	9.8	51.3	16.8	10.6	10.1	8.0	1.0	14.0	46.2	21.1	0.9	0.6	16.4	5.6	16.4	11.0	43.2	50.1	24.4	66.6	16.6	19.3
FLAIR	19.3	4.6	19.4	8.5	16.6	11.3	63.2	29.6	13.5	12.8	9.9	1.0	8.2	61.6	30.6	0.9	0.7	12.5	4.7	26.8	5.2	57.2	50.1	27.6	83.1	25.2	23.2
ITO	23.3	5.6	29.0	16.2	19.9	11.8	65.0	26.1	17.1	13.7	14.3	1.7	17.9	62.9	32.4	1.1	0.7	12.7	13.8	21.4	9.3	48.4	50.1	25.9	85.6	21.9	**24.9**
ITO_sub2	23.1	4.7	28.1	16.2	19.8	12.2	57.7	26.9	13.6	13.6	13.7	1.7	10.1	66.1	31.5	1.0	0.8	11.1	14.9	26.9	6.8	59.6	50.0	27.6	88.1	21.4	**24.9**
(b) *Results on CC12M (ViT-B/16, 30 epochs)*																											
CLIP	36.7	8.5	45.6	24.5	31.5	22.4	70.4	38.1	32.5	45.6	59.7	21.0	9.6	72.0	42.6	2.2	4.7	17.6	5.7	29.6	12.9	54.3	50.0	36.4	88.0	24.2	34.1
SLIP	40.9	12.4	50.7	28.8	34.7	27.6	72.6	46.6	32.9	48.8	54.0	20.7	9.0	74.9	43.8	2.7	5.4	22.6	11.5	31.7	7.9	50.0	48.8	32.1	91.3	33.2	36.0
SigLIP	40.6	11.3	53.0	28.7	35.1	25.2	76.5	41.0	27.8	38.1	48.4	20.9	12.9	71.4	42.1	2.6	5.0	27.1	10.2	20.9	11.4	53.5	49.6	41.2	88.7	27.5	35.0
FLAIR	41.5	10.4	51.2	30.2	34.9	25.5	73.5	38.3	30.5	47.6	60.8	22.2	4.6	75.4	47.9	2.0	6.2	19.4	7.2	22.8	10.7	51.8	49.9	39.3	89.1	29.6	35.5
ITO	45.5	12.7	60.6	36.0	39.3	28.6	76.7	47.5	34.9	54.1	62.0	30.0	10.3	79.1	50.8	3.2	5.7	23.7	1.2	19.7	7.7	50.0	50.2	44.9	95.4	27.7	38.4
ITO_sub2	47.0	13.1	60.9	38.4	40.6	29.5	84.2	51.2	39.9	57.0	69.7	33.1	11.9	77.4	53.3	3.7	7.8	25.5	4.2	17.6	13.1	50.9	49.9	45.3	94.8	30.4	**40.4**
(c) *Results on YFCC15M (ViT-B/16, 30 epochs)*																											
CLIP	36.4	20.2	22.9	9.4	31.9	17.6	71.6	36.9	53.6	41.6	22.6	2.7	10.2	65.1	37.6	2.7	7.1	16.7	16.6	26.3	4.4	51.9	50.0	24.6	86.1	25.6	30.5
ITO	44.3	25.6	32.7	16.8	38.6	22.8	70.4	40.6	60.6	50.8	30.9	4.5	10.5	72.8	47.2	2.4	8.4	22.4	7.1	24.3	7.2	55.0	50.4	31.7	94.0	23.8	34.5
ITO_sub2	44.4	25.2	34.1	16.0	39.1	22.5	75.5	38.3	63.2	54	36.9	3.9	10.0	74.0	48.9	3.8	8.8	26.2	7.8	22.2	8.2	53.3	50.0	29.4	95.2	28.8	**35.4**
(d) *Results on Laion100M (ViT-B/16, 30 epochs)*																											
CLIP	59.0	21.8	69.2	44.1	51.4	43.7	91.0	69.8	54.8	79.4	82.7	76.6	63.6	81.5	61.7	11.1	13.5	45.4	3.1	36.4	38.2	51.0	54.6	56.0	95.5	34.9	53.5
SLIP	55.4	21.0	65.7	41.7	48.5	43.1	89.9	68.9	51.5	75.3	74.3	72.7	57.0	81.2	58.6	9.2	11.5	45.3	1.8	42.9	35.5	50.3	50.6	50.7	93.8	36.2	51.3
ITO	62.8	25.6	75.5	50.2	54.8	46.2	93.8	75.0	61.4	81.9	83.6	81.7	58.1	83.8	64.6	13.7	13.9	50.1	2.7	40.6	40.7	50.3	54.4	60.2	97.0	35.8	**56.1**
(e) *Results on DataComp-1B (ViT-B/16, 1 epoch)*																											
CLIP	63.1	24.1	67.9	47.3	54.7	50.0	92.1	74.6	63.7	83.1	84.0	76.9	55.1	82.1	62.8	14.2	13.9	42.1	3.1	25.0	38.8	50.8	49.8	58.1	95.3	35.4	54.2
ITO	65.9	25.0	72.9	53.3	57.2	51.9	94.9	78.3	70.5	83.9	86.6	83.5	57.3	83.4	64.9	18.1	14.7	46.0	3.9	24.4	35.5	53.5	51.7	60.2	96.4	37.8	56.6
ITO_sub2	65.7	26.9	74.0	53.5	57.0	52.2	95.1	78.5	67.0	84.5	85.8	82.3	57.0	82.0	64.7	16.3	14.9	47.8	3.2	29.5	42.1	58.4	51.4	57.8	96.7	37.4	**57.0**
(f) *Results on DataComp-1B (ViT-B/16, 10 epochs)*																											
CLIP	72.0	45.7	81.2	58.6	64.7	62.7	95.7	80.8	74.4	90.2	92.0	86.7	79.3	84.7	69.7	25.4	21.4	59.4	2.2	39.9	54.1	62.8	51.3	66.1	97.7	37.3	63.7
ITO	73.5	45.7	83.2	62.0	66.3	63.6	96.9	83.5	75.8	90.8	91.1	89.2	81.1	85.1	70.3	30.8	21.8	61.4	7.3	37.8	55.5	57.7	50.5	69.9	98.3	38.3	**64.9**
ITO_sub2	72.1	42.8	82.9	60.7	64.4	62.9	96.7	82.7	74.0	89.7	90.8	87.3	76.6	85.4	69.8	27.2	21.0	60.4	4.8	32.4	45.9	56.5	52.4	66.7	98.1	38.1	63.2
(g) *Results on DataComp-1B (ViT-L/16, 1 epoch)*																											
CLIP	69.5	39.8	78.0	56.5	61.1	60.1	96.1	80.9	71.5	88.5	88.1	83.2	70.1	83.2	68.9	16.2	19.0	51.1	2.4	33.6	52.1	59.8	51.8	63.1	97.5	37.2	60.7
ITO	72.2	44.5	82.9	62.0	64.3	64.1	97.6	83.8	72.0	89.8	90.8	87.2	74.8	84.6	70.7	27.4	20.3	56.0	2.2	31.4	49.0	50.4	53.7	67.9	98.6	38.9	**63.0**
ITO_sub2	70.7	38.8	80.4	59.4	62.8	60.3	97.0	82.4	73.4	87.9	90.7	86.8	66.7	83.4	69.6	21.5	19.1	53.9	3.4	33.7	47.3	52.8	52.0	67.2	98.2	38.3	61.5

E Zero-shot Results on DataComp-1B

We provide additional zero-shot classification results on the billion-scale DataComp-1B dataset to analyze the effect of training duration and model scale. All evaluations follow the same protocol as in the main paper and use standard CLIP-style prompts.

Effect of Training Epochs. We first study the impact of training duration using ViT-B/16. In addition to the 1-epoch setting reported in the main paper, we further train models for 10 epochs to examine scaling behavior with respect to optimization. As shown in Table 12, extending training consistently improves zero-shot performance across most benchmarks. Compared with CLIP, ITO exhibits more stable gains as training progresses, indicating improved utilization of large-scale data.

Effect of Model Scale. We further evaluate a larger ViT-L/16 model trained for 1 epoch on DataComp-1B. Despite the shorter training schedule, the ViT-L/16 variant achieves competitive or superior performance compared with ViT-B/16 trained for longer durations. This suggests that ITO benefits from model scaling and that training-time fusion remains effective when increasing model capacity.

Overall, these results demonstrate that ITO scales favorably with both training epochs and model size on billion-scale image–text data, while preserving the standard dual-encoder inference architecture.

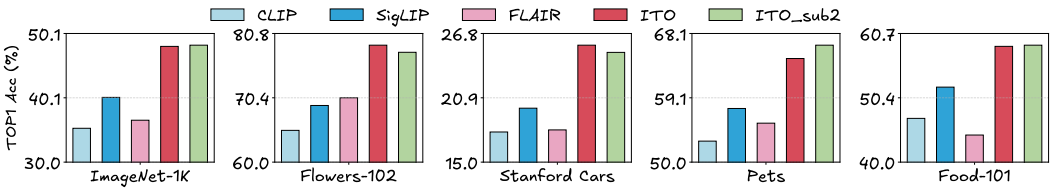


Figure 2: Linear image classification of ITO and its variants pretrained on CC3M.

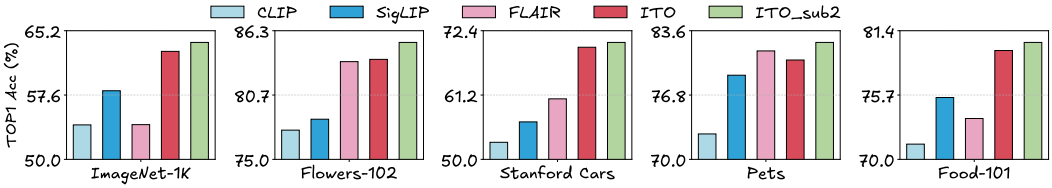


Figure 3: Linear image classification of ITO and its variants pretrained on CC12M.

F Robustness Across Data-Quality Regimes

Image–text pretraining is sensitive to data curation. We have not run a controlled ablation that varies caption quality on a fixed image set, but our pretraining corpora already span a wide range of curation: CC3M and CC12M are alt-text corpora built with automatic filtering pipelines, stricter for CC3M and relaxed for CC12M; YFCC15M consists of noisy user-provided titles and descriptions; Laion100M is loosely filtered by CLIP similarity; and DataComp-1B is aggressively filtered. ITO improves the 26-benchmark zero-shot average over CLIP in every one of these regimes (Table 1): $+5.6$ on CC3M, $+4.3$ on CC12M, $+4.0$ on YFCC15M, $+2.6$ on Laion100M, and $+2.4$ on DataComp-1B. This is indirect but broad evidence that the method is not tied to a particular curation level. A controlled study that trains on curated and noisy subsets of the same corpus would isolate this factor more cleanly, and we leave it to future work.

G Linear Probing: Full Results

This section provides additional linear probing results that complement Table 2 in the main paper. We report (i) per-dataset results on five classification benchmarks (ImageNet-1K, Flowers-102, Stanford Cars, Pets, Food-101) across all pretraining scales (Table 13), and (ii) results on CC3M and CC12M, which are not included in the main-paper table due to space constraints (Figure 2 and Figure 3).

H Sub-description Protocol

Sub-descriptions are generated and sampled following FLAIR [50]; at each training step, two sub-descriptions are sampled per caption to form the two text views of ITO_sub2.

I Additional Experiments on CC3M-recap

To further investigate the impact of text augmentation, we conduct supplementary experiments on the CC3M-recap dataset, a synthetic dataset provided by Dreamlip [56]. In CC3M-recap, each original image–text pair (I, T) is processed by Vision–Language Models (VLMs) using two prompts—"Describe the image in detail" and "Describe the image in short"—to generate a long and a short caption, respectively. Moreover, by employing three different VLMs (InstructBLIP [12], LLaVA-1.5 [33], and ShareGPT4V [5]), each original pair yields six additional synthetic textual descriptions. This setup allows us to examine whether richer textual supervision can further enhance the performance of the proposed multimodal fusion framework.

Experimental Setup. Following the same training configurations as in the main experiments, we compare FLAIR [50] with our text-augmented variants ITO_sub2 and ITO_sub3 on the CC3M-recap dataset. The original CC3M-recap dataset contains approximately 2.8 million image–text pairs, while

Table 13: Linear image classification of ITO and its variants pretrained on YFCC15M, Laion100M and DataComp-1B. The best results are highlighted in **bold**. † indicates models trained for 10 epochs using ViT-B/16. * indicates models trained for 1 epochs using ViT-L/16.

Dataset	Method	IN-1k	Flowers	Cars	Pets	Food	Avg
15M	CLIP	51.82	85.17	19.60	53.61	69.18	55.88
	ITO	**60.52**	92.18	**28.37**	61.22	**76.50**	63.76
	ITO_sub2	60.40	**92.37**	27.73	**62.82**	76.22	**63.91**
100M	CLIP	67.27	86.65	87.17	82.88	85.35	81.86
	ITO	**68.70**	**90.21**	**88.16**	**84.49**	**86.22**	**83.56**
1B	CLIP	67.39	90.08	87.14	79.61	86.47	82.14
	ITO	69.87	**92.47**	**90.70**	81.88	87.75	82.53
	ITO_sub2	**70.76**	92.24	90.34	**83.16**	**88.51**	**83.00**
	CLIP†	74.99	94.67	91.64	87.90	91.89	87.62
	ITO†	**76.20**	**96.16**	**92.89**	88.09	**92.54**	**89.18**
	ITO_sub2†	75.83	94.96	92.23	**89.37**	92.19	88.92
	CLIP*	73.78	93.04	90.23	86.54	90.83	86.88
	ITO*	**75.97**	**95.80**	92.28	87.84	**92.27**	**88.83**
	ITO_sub2*	74.88	94.89	**92.29**	**88.93**	91.43	88.48

Table 14: Top-1 zero-shot classification accuracy on 26 public benchmarks following the EVA-CLIP [45] protocol. Results are reported for models pretrained on CC3M-recap with both ViT-B/16 as vision encoders. The best results are highlighted in **bold**.

	ImageNet-1k	ImageNet-A	ImageNet-R	ImageNet-S	ImageNet-V2	ObjectNet	CIFAR-10	CIFAR-100	Flowers-102	Food-101	Pets	Stanford Cars	MNIST	Caltech	SUN397	FGVC Aircraft	Country-211	DTD	EuroSAT	FER2013	GTSRB	PCam	Rendered SST2	Resisc45	STL10	VOC2007	Avg
(a) CC3M Pre-training ViT-B																											
FLAIR	27.7	7.6	31.4	15.0	23.8	15.5	73.9	44.3	18.9	19.8	25.9	2.5	13.1	69.8	42.3	1.6	2.6	18.8	7.4	30.6	11.4	45.7	50.1	34.6	92.9	26.6	29.0
ITO_sub2	30.4	8.8	38.8	19.9	27.2	16.1	62.7	38.0	19.7	23.1	29.9	3.6	18.5	71.4	46.0	2.6	2.6	20.5	7.1	22.4	11.9	50.0	50.1	37.3	94.4	27.1	**30.0**
ITO_sub3	30.9	9.4	39.5	20.7	26.7	16.3	62.2	36.1	20.0	23.1	30.2	4.2	16.8	72.1	46.0	1.2	2.6	21.3	8.5	19.6	10.2	50.0	50.1	35.0	94.0	22.5	29.6

we use about 1.7 million successfully downloaded samples in our experiments. For ITO_sub2 and ITO_sub3, we randomly sample two or three sub-descriptions from the long captions to construct multiple text views, respectively. All models are trained for 30 epochs using ViT-B/16 as the vision encoder. We evaluate them on the same downstream tasks as in the main paper, including 26 zero-shot classification datasets, linear probing for image classification, and image–text retrieval benchmarks.

Results. The results are summarized in Tables 14 to 17. Across all downstream tasks, incorporating richer textual supervision from CC3M-recap leads to consistent improvements over FLAIR [50].

In the 26-dataset zero-shot classification benchmark (Table 14), both ITO_sub2 and ITO_sub3 outperform FLAIR by a clear margin, achieving average Top-1 accuracies of 30.0% and 29.6%, respectively. This demonstrates that our contrastive–fusion framework benefits from more diverse textual views even when trained on synthetic captions.

For linear probing (Table 15), ITO_sub2 and ITO_sub3 again show notable gains on ImageNet-1K and several fine-grained datasets, with ITO_sub3 achieving the best overall results (e.g., +7.4 points over FLAIR on ImageNet-1K). These findings indicate that long-caption sampling not only enhances zero-shot transfer but also improves the linear separability of learned visual features.

In zero-shot image–text retrieval (Table 16 and Table 17), our models consistently surpass FLAIR on both coarse-grained (MSCOCO, Flickr30k) and fine-grained (DOCCI) benchmarks. For instance,

Table 15: Linear image classification of ITO and its variants pretrained on CC3M-recap.

Dataset	Method	ImageNet-1k	Flowers-102	Stanford Cars	Pets	Food-101
CC3M-recap	FLAIR	45.65	**76.57**	25.05	**67.05**	59.38
	ITO_sub2	52.43	73.64	23.96	65.09	64.15
	ITO_sub3	**53.05**	74.30	**25.32**	65.55	**64.70**

Table 16: Zero-shot image-text retrieval on validation splits for standard benchmarks (MSCOCO and Flickr30k).

Dataset	Method	MSCOCO						Flickr30k					
		I → T			T → I			I → T			T → I		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CC3M-recap	FLAIR	37.54	64.38	75.88	29.76	55.53	66.69	65.19	87.08	92.41	53.25	78.01	85.48
	ITO_sub2	**47.76**	73.76	82.64	31.59	58.54	**69.98**	74.56	93.89	96.75	57.32	82.23	88.38
	ITO_sub3	47.40	**74.94**	**83.86**	**32.38**	**58.74**	69.77	**74.95**	**94.38**	**96.75**	**59.13**	**83.41**	**89.80**

Table 17: Zero-shot image-text retrieval on validation splits for fine-grained retrieval benchmarks (DOCCI).

Dataset	Method	DOCCI					
		I → T			T → I		
		R@1	R@5	R@10	R@1	R@5	R@10
CC3M-recap	FLAIR	20.16	43.56	54.66	10.65	24.23	31.60
	ITO_sub2	**27.82**	**54.14**	**65.90**	**10.70**	23.69	30.84
	ITO_sub3	27.54	53.68	65.30	10.68	**24.04**	**31.27**

ITO_sub3 improves Recall@1 on MSCOCO I → T from 37.54% to 47.40%, and on Flickr30k from $65.19\%$to 74.95%. Similar gains are observed on DOCCI, confirming that text augmentation yields more robust and fine-grained alignment between images and captions.

Overall, these results validate the effectiveness of our text sampling strategy under the VLM-recaptioned setting. Richer and more diverse textual descriptions help our method capture finer semantic correspondences, while maintaining strong generalization across classification and retrieval tasks.

Discussion. These findings suggest that when captions are linguistically rich, sub-description sampling further enhances cross-modal learning by emphasizing diverse semantic aspects of the same image. Compared with the models pretrained on the original CC3M dataset, those trained on CC3M-recap achieve consistently higher performance across classification and retrieval benchmarks, confirming that richer textual supervision facilitates more effective image–text alignment. However, as shown in Section 4.2, this advantage attenuates on DataComp-1B, where native captions are already lexically diverse and sub-descriptions become largely redundant.

J DOCCI and Full Zero-shot Retrieval Results

We provide additional zero-shot image–text retrieval results to complement the main paper. This section reports the complete Recall@1/5/10 metrics on MSCOCO and Flickr30k, as well as fine-grained retrieval performance on DOCCI.

Complete Retrieval Results on MSCOCO and Flickr30k. Table 18 report full bidirectional retrieval results (Image→Text and Text→Image) on the validation splits of MSCOCO and Flickr30k. These results extend the main paper, where only Recall@1 and Recall@5 are shown for clarity. The trends are consistent with those reported in the main text: ITO achieves strong and stable improvements over CLIP, SLIP, and FLAIR across datasets and retrieval directions, and the relative ordering of methods remains unchanged when considering Recall@10.

Fine-grained Retrieval on DOCCI. We further evaluate zero-shot retrieval on the DOCCI dataset, which provides an average of seven sentence-level descriptions per image and emphasizes fine-grained cross-modal matching. As shown in Table 19, ITO remains competitive across all recall metrics, demonstrating that representations learned via training-time fusion generalize well to detailed and

Table 18: Zero-shot image-text retrieval on validation splits for standard benchmarks (MSCOCO and Flickr30k). The best results are highlighted in **bold**. † indicates models trained for 10 epochs using ViT-B/16. * indicates models trained for 1 epochs using ViT-L/16.

Dataset	Method	MSCOCO						Flickr30k					
		I → T			T → I			I → T			T → I		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
3M	CLIP	12.36	30.98	41.76	8.14	22.68	31.98	23.57	50.69	62.03	16.98	37.65	48.68
	SigLIP	13.70	32.66	43.16	9.28	23.48	32.60	28.90	53.25	65.09	18.32	39.43	49.45
	FLAIR	17.86	38.90	51.18	12.59	30.38	40.94	35.40	62.13	74.46	25.74	49.86	60.81
	ITO	**21.56**	45.36	57.22	**14.56**	33.71	44.90	**42.6**	**71.2**	**80.87**	**29.45**	**53.89**	**64.36**
	ITO_sub2	21.08	**45.62**	**57.42**	14.01	**33.88**	**45.06**	42.11	69.43	78.11	28.26	53.08	64.16
12M	CLIP	34.17	61.20	72.52	23.04	47.58	59.39	62.23	86.29	92.11	45.74	73.02	82.23
	SigLIP	39.98	67.28	77.82	26.85	51.48	63.20	68.34	89.45	93.29	50.87	75.96	84.18
	FLAIR	36.20	63.16	74.38	24.62	48.79	60.57	62.92	88.56	93.20	47.36	74.77	83.21
	ITO	42.30	70.00	79.46	29.62	55.74	67.09	**72.49**	90.83	95.07	54.50	79.98	86.86
	ITO_sub2	**43.94**	**70.40**	**80.26**	**30.51**	**56.76**	**68.29**	72.29	**92.80**	**96.35**	**56.65**	**82.54**	**88.58**
15M	CLIP	26.38	50.86	62.86	15.06	34.82	46.43	47.14	74.46	83.23	30.77	56.82	67.14
	ITO	30.76	57.44	68.82	19.57	41.61	53.03	54.83	82.74	88.56	35.72	62.60	73.06
	ITO_sub2	**31.42**	**58.02**	**69.06**	**20.04**	**42.88**	**54.61**	**56.21**	**83.14**	**90.43**	**37.95**	**65.46**	**75.50**
100M	CLIP	49.12	74.66	83.76	31.92	57.32	68.40	75.94	93.59	96.75	60.04	84.08	89.90
	ITO	**52.08**	**76.04**	**84.24**	**34.26**	**59.57**	**70.07**	**79.59**	**95.66**	**98.42**	**63.00**	**86.21**	**91.42**
1B	CLIP	47.08	72.70	82.64	29.51	54.84	66.05	72.68	90.63	94.48	54.73	80.49	87.53
	ITO	49.26	**74.76**	**83.62**	31.30	56.63	68.22	**75.94**	**94.08**	**97.14**	**56.79**	82.17	88.56
	ITO_sub2	**49.50**	74.12	83.06	**31.89**	**57.26**	**68.36**	73.37	93.00	96.15	55.70	**82.33**	**89.33**
	CLIP†	57.70	80.86	87.72	37.76	63.58	73.50	82.25	96.65	98.52	66.57	87.71	92.60
	ITO†	**58.14**	**81.76**	**88.88**	**38.97**	**64.45**	**74.63**	**84.81**	**96.45**	**98.52**	**67.10**	**88.38**	**93.18**
	ITO_sub2†	55.88	79.90	87.84	38.33	64.03	74.03	81.95	95.36	98.03	65.88	87.67	92.49
	CLIP*	53.40	77.76	86.62	34.88	60.16	71.04	78.40	95.07	97.63	62.60	85.62	91.26
	ITO*	**56.58**	**80.46**	**88.12**	**38.12**	**63.72**	**73.62**	**83.14**	**96.35**	**98.62**	**66.33**	**87.69**	**92.96**
	ITO_sub2*	52.12	77.44	85.96	35.69	61.62	71.86	78.99	95.27	98.03	62.88	85.42	91.20

compositional image–text retrieval scenarios. These results are consistent with the conclusions drawn from COCO and Flickr30k, and further validate the robustness of ITO under diverse retrieval settings.

K Consistency Score and Global Geometry Probes

The consistency score (Table 20) is a neighborhood-level measure: it improves with fusion at every K and at both scales, but it does not directly characterize global geometry. To complement it, we measured three global quantities on ImageNet embeddings for ITO with and without fusion across three pretraining scales: the image–text centroid gap (cosine similarity and L2 distance between the two modality centroids, following [31]), the cross-modal MMD, and the ratio of same-class to different-class cross-modal distances, with standard deviations estimated from 50 bootstrap resamples. The effect of fusion on these quantities is small and not directionally consistent across scales. Fusion narrows the centroid gap at CC3M (centroid cosine $0.423 \rightarrow 0.466$) but widens it at DataComp-1B ($0.336 \rightarrow 0.314$). MMD and the class-distance ratio change only at the third to fourth decimal; with a bootstrap standard deviation of about 3×10^{-4}, these changes are genuinely small rather than measurement noise. No single global quantity tracks the downstream gains, which is why we rest the justification of fusion on the optimization evidence in Section 4.8.

L UMAP Visualization

L.1 CLIP vs. FLAIR vs. ITO: Modality Separation vs. Integration

UMAP projections are qualitative: a two-dimensional embedding can exaggerate or hide structure, and the global geometry probes in Appendix K show that the effect of fusion on global quantities is small and scale-dependent. The observations below should be read as illustrations rather than as evidence of a geometric mechanism.

Table 19: Zero-shot image-text retrieval on validation splits for fine-grained retrieval benchmarks (DOCCI). The best results are highlighted in **bold**. † indicates models trained for 10 epochs using ViT-B/16. * indicates models trained for 1 epochs using ViT-L/16.

Dataset	Method	DOCCI					
		I → T			T → I		
		R@1	R@5	R@10	R@1	R@5	R@10
3M	CLIP	4.10	12.56	18.72	2.13	6.71	10.07
	SigLIP	4.98	14.38	20.88	2.28	6.91	10.32
	FLAIR	7.12	19.12	28.00	3.23	9.70	14.13
	ITO	8.24	22.00	**31.64**	3.48	**9.97**	14.39
	ITO_sub2	**9.18**	**22.74**	31.20	**3.49**	9.96	**14.41**
12M	CLIP	20.14	43.40	53.54	7.67	17.96	24.03
	SigLIP	25.00	49.66	61.72	9.38	20.59	26.94
	FLAIR	24.24	48.50	59.12	9.36	21.20	28.11
	ITO	26.62	51.00	61.88	9.39	20.51	26.92
	ITO_sub2	**28.26**	**53.92**	**64.46**	**10.86**	**23.60**	**30.44**
15M	CLIP	15.94	35.26	45.70	5.45	13.54	18.81
	ITO	19.06	40.28	51.90	6.82	16.45	22.45
	ITO_sub2	**19.82**	**43.06**	**53.70**	**6.86**	**17.52**	**23.60**
100M	CLIP	39.28	65.30	74.68	13.04	25.45	31.81
	ITO	**41.54**	**67.48**	**76.52**	**13.68**	**26.48**	**32.80**
1B	CLIP	43.02	69.48	77.88	13.01	25.71	32.28
	ITO	43.24	69.98	79.26	13.75	26.78	33.27
	ITO_sub2	**44.08**	**70.86**	**79.56**	**14.47**	**28.31**	**35.01**
	CLIP†	52.58	76.98	84.78	16.66	30.49	37.09
	ITO†	**52.78**	**78.20**	**85.72**	17.30	31.60	38.30
	ITO_sub2†	50.78	76.18	83.78	**17.79**	**32.73**	**39.90**
	CLIP*	48.98	75.26	83.12	15.59	29.16	35.49
	ITO*	**50.94**	**76.56**	**84.36**	**16.77**	30.75	37.60
	ITO_sub2*	48.34	75.02	82.68	16.67	**30.95**	**38.01**

Table 20: Cross-modal consistency score (%) at top-K following CyCLIP [19]. Δ is the change from adding fusion regularization (λ=2) to the alignment-only variant (λ=0). ZS IN-1k is reported alongside for reference.

Dataset	Setting	Top-1	Top-3	Top-5	Top-10	ZS IN-1k
15M	ITO (λ=0)	32.1	31.9	33.7	35.3	43.7
	ITO (λ=2)	**32.7**	**32.4**	**34.3**	**35.7**	**44.3**
	Δ	+0.6	+0.5	+0.6	+0.4	+0.6
1B	ITO (λ=0)	46.1	45.1	47.5	49.1	64.8
	ITO (λ=2)	**46.9**	**45.7**	**48.0**	**49.8**	**65.9**
	Δ	+0.8	+0.6	+0.5	+0.7	+1.1

We first compare the representation structures learned by CLIP, FLAIR, and ITO using UMAP visualization under identical settings. All models are trained on the CC3M dataset. For visualization, we randomly sample 8,192 image–text pairs from CC12M and project their embeddings into two dimensions using UMAP. Blue points represent image embeddings, and red points represent text embeddings.

As shown in Figure 4a, CLIP exhibits a clear separation between image and text embeddings, forming two modality-specific clusters with a distinct boundary. This observation is consistent with prior findings that instance-level contrastive alignment alone does not eliminate modality-induced organization in the embedding space.

Figure 4b presents the visualization for FLAIR. Compared with CLIP, FLAIR produces more compact text embeddings that are partially surrounded by image embeddings. This behavior is likely attributed to its text-conditioned pooling mechanism, which introduces localized fusion effects during training. However, the overall structure remains asymmetric, with text embeddings forming a concentrated region and images distributed more broadly, indicating that modality separation is alleviated but not fully resolved.

(a) CLIP (b) FLAIR (c) **ITO**

Figure 4: **UMAP visualization.** All models are trained on the CC3M dataset. For visualization, 8,192 image-text pairs are randomly sampled from CC12M. Blue points represent images, and red points represent texts. (a): CLIP exhibits a clear separation between modalities, with a distinct boundary between image and text embeddings. (b): FLAIR shows more compact text embeddings surrounded by image embeddings, likely due to its text-conditioned fusion mechanism. (c): ITO shows a star-shaped distribution in which image and text embeddings are more closely interleaved and the boundary between modalities is less pronounced in this projection.

In contrast, Figure 4c shows the visualization for ITO. Image and text embeddings are more closely interwoven, forming a star-shaped distribution in which both modalities are mixed within shared neighborhoods. Notably, this visualization is obtained using only the standalone dual-encoder at inference time, without any fusion module. This suggests that the reduced modality separation is not an artifact of architectural unification, but rather reflects a change in the representation structure learned during training.

Taken together, in these projections ITO appears less structured by modality than CLIP and FLAIR.

L.2 Effect of Training-Time Fusion ($\lambda = 0$ vs. $\lambda > 0$)

(a) ITO ($\lambda = 0$) (b) ITO ($\lambda = 2$) (c) ITO_sub2

Figure 5: **UMAP visualization.** All models are trained on the CC3M dataset. For visualization, 8,192 image-text pairs are randomly sampled from CC12M. Blue points represent images, and red points represent texts.

To disentangle the effect of Multimodal Multiple Alignment from that of training-time multimodal fusion, we visualize the representation structures learned with different fusion weights. All models are trained on CC3M, and 8,192 image–text pairs are randomly sampled from CC12M for UMAP visualization. Inference is performed using the standalone dual-encoder in all cases.

Figure 5a shows the visualization for ITO with $\lambda = 0$, where only Multimodal Multiple Alignment is applied. Compared with CLIP, the separation between image and text embeddings is reduced, indicating improved instance-level alignment. However, the embedding space remains partially organized by modality, and a visible boundary between image and text representations persists.

Figure 5b presents the full ITO model with $\lambda = 2$. With training-time multimodal fusion enabled, image and text embeddings appear more interwoven and the boundary between modalities is weaker in this projection on CC3M, where the centroid-gap probe also narrows (Appendix K).

Figure 5c further shows ITO_sub2, which incorporates both training-time fusion and sub-description sampling. The resulting embedding space exhibits the strongest degree of cross-modal mixing, with image and text representations densely interleaved within shared neighborhoods. This is consistent with textual diversity being most useful in the moderate data regime, as discussed in Section 4.2.

L.3 CC3M-recap: FLAIR vs. ITO_sub2 vs. ITO_sub3

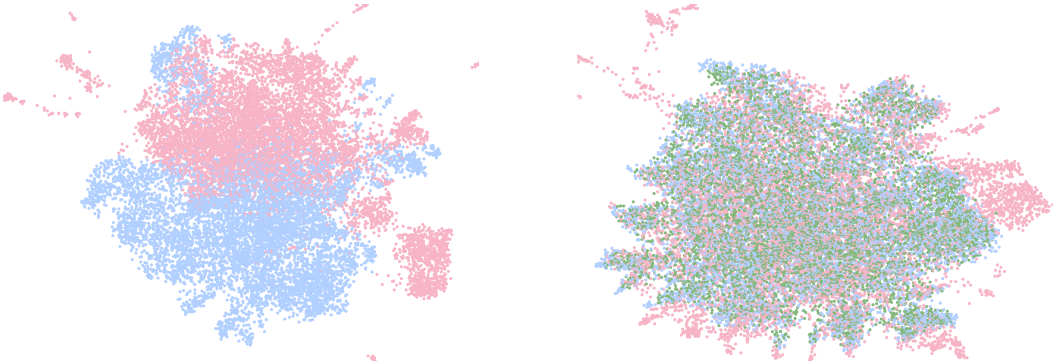
(a) FLAIR (b) ITO_sub2 (c) ITO_sub3

Figure 6: **UMAP visualization.** All models are trained on the CC3M-recap dataset. For visualization, 8,192 image-text pairs are randomly sampled from CC12M. Blue points represent images, and red points represent texts.

We conduct additional visualization analyses to further examine the behavior of our model across different fusion strategies and caption granularities. We first compare FLAIR [50], ITO_sub2, and ITO_sub3 trained on the CC3M-recap dataset. As shown in Figure 6, FLAIR remains partially organized by modality on CC3M-recap, exhibiting limited cross-modal mixing, in contrast to its slightly more compact structure observed on the original CC3M dataset. This difference may be attributed to the fact that all subclauses in CC3M-recap originate from the same source caption, reducing the diversity of textual supervision.

In contrast, both ITO_sub2 and ITO_sub3 produce embedding spaces where image and text representations are more tightly interleaved. In these projections, representations appear less separated by modality than those of FLAIR.

L.4 DataComp-1B: CLIP vs. ITO



(a) CLIP (b) ITO

Figure 7: **UMAP visualization for DataComp-1B.** All models are trained on the DataComp-1B dataset for 10 epochs. For visualization, 8,192 image-text pairs are randomly sampled from CC12M. Blue points represent images, red points represent texts, and green points denote unified multimodal tokens. .

To investigate scalability, we further visualize models pretrained on the billion-scale DataComp-1B dataset. Specifically, we compare CLIP [41] and ITO using ViT-B/16, with both models pretrained for 10 epochs. For visualization, 8,192 image–text pairs are randomly sampled from CC12M to ensure

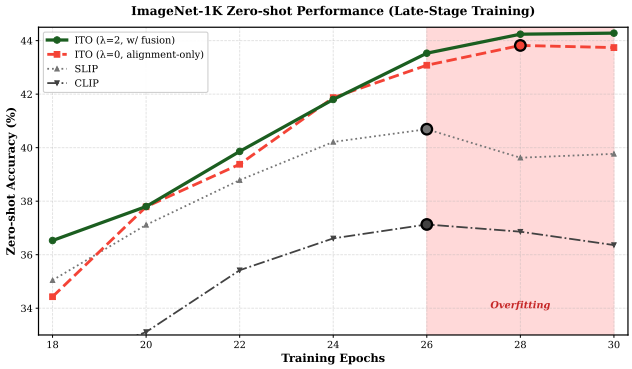


Figure 8: **Training dynamics and overfitting behavior on YFCC15M.** Zero-shot ImageNet-1K accuracy as a function of training epochs for CLIP, SLIP, and ITO variants using ViT-B/16. Both CLIP and SLIP exhibit late-stage performance degradation, indicating overfitting on moderately sized datasets. Introducing Multimodal Multiple Alignment alone (λ=0) improves overall accuracy but does not fully prevent overfitting. In contrast, enabling training-time multimodal fusion (λ=2) stabilizes training and maintains consistent zero-shot performance in later epochs, demonstrating the regularizing effect of fusion-based cross-modal interaction.

stable and consistent comparison across different pretraining scales. Blue points represent image embeddings, red points denote text embeddings, and green points correspond to unified multimodal tokens.

As shown in Figure 7, CLIP exhibits a clear modality-induced separation, with image and text embeddings forming two distinct clusters. In contrast, ITO produces more intertwined distributions in which representations appear less structured by modality. Note that this comparison is between CLIP and the full ITO; isolating the fusion module at this scale (λ=0 vs. λ=2), the centroid-gap probe does not narrow (Appendix K), so we do not attribute this visual difference to fusion alone.

M Training Dynamics and Overfitting Analysis

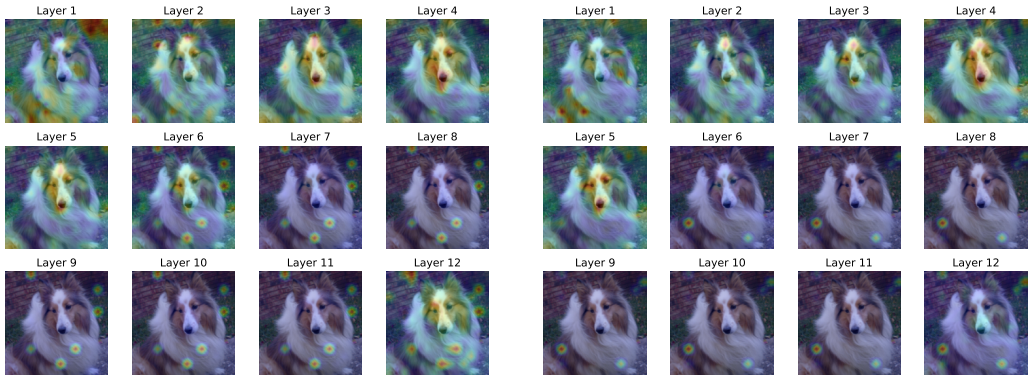
To better understand the role of training-time multimodal fusion in image–text contrastive pretraining, we analyze the training dynamics of different methods by tracking zero-shot ImageNet-1K accuracy throughout training. All models are trained on YFCC15M using ViT-B/16, and evaluated under the standard CLIP zero-shot protocol.

Figure 8 reports the evolution of zero-shot accuracy for CLIP, SLIP, and our proposed ITO variants. Both CLIP and SLIP exhibit a clear late-stage performance degradation: after reaching peak accuracy, their zero-shot performance declines as training proceeds, indicating overfitting despite continued optimization. This behavior is consistent with prior observations in image–text contrastive learning on moderately sized datasets.

Introducing Multimodal Multiple Alignment alone (λ=0) substantially improves overall accuracy compared with CLIP and SLIP. However, this alignment-only variant still suffers from noticeable late-stage overfitting, as reflected by a similar peak-and-decline pattern in the training curve. This suggests that enriching instance-level contrastive supervision, while beneficial, is insufficient to fully stabilize training.

In contrast, enabling training-time multimodal fusion (λ=2) fundamentally changes the training dynamics. ITO exhibits consistently stable zero-shot performance in the later stages of training, without observable degradation. This indicates that the fusion objective provides an additional regularizing effect, guiding the encoders toward representations that generalize better and remain robust as training progresses.

Taken together, since the λ=0 and λ=2 variants differ only in the fusion module, these results isolate overfitting suppression as a direct effect of fusion rather than of multi-view alignment. Importantly, these benefits are achieved without introducing any additional inference-time cost, as the fusion module is discarded during deployment.



(a) CLIP (c) ITO



(b) CLIP (d) ITO

Figure 9: **Layer-wise attention visualization of CLIP and ITO on DataComp-1B.** Each image corresponds to a different transformer layer, with the layer index indicated above each visualization. ITO shows progressively more focused and semantically consistent attention distributions than CLIP. Note that all visualizations are derived from held-out samples rather than those seen during pretraining.

N Layer-wise attention visualization

We visualize the attention maps across all 12 layers of the vision encoders from CLIP [41] and our ITO (trained for 10 epochs on DataComp-1B) to analyze their visual reasoning behaviors. As shown in Figure 9, ITO exhibits more concentrated and semantically coherent attention distributions across layers. From shallow to deep layers, the attention in ITO progressively evolves from dispersed low-level responses to high-level representations that focus on the semantic subject of the scene, while effectively suppressing background noise. This indicates that ITO forms a clearer hierarchical structure for semantic information aggregation and achieves stronger alignment between visual features and semantic content. In contrast, CLIP shows less stable attention evolution, with certain layers focusing on irrelevant background regions and exhibiting weaker inter-layer consistency. Overall, the representations learned by ITO are more structured and interpretable, demonstrating enhanced semantic consistency and better generalization capability.

O Inference Efficiency.

ITO introduces a multimodal fusion module *only during training*. At inference time, the fusion module is entirely removed, and ITO reduces to a standard dual-encoder architecture identical to CLIP. As a result, ITO has the same number of parameters, computational cost, and inference latency as CLIP, and can be used as a drop-in replacement for existing image–text contrastive encoders.

P Training Overhead Analysis

Table 7 in the main paper summarizes the training cost of ITO; Table 21 gives the full breakdown, including the alignment-only variant. All measurements use 64 NVIDIA A800 GPUs with matched batch size and optimizer settings.

Table 21: Training time (hours) and peak GPU memory (GB per GPU) of OpenCLIP and ITO. Ratios are relative to OpenCLIP at the same scale.

Dataset	Method	Time (h)	Time ratio	Memory (GB)	Memory ratio
YFCC15M	OpenCLIP	3.6	1.0×	7.9	1.0×
	ITO (λ=0)	8.2	2.3×	16.8	2.1×
	ITO (λ=2)	8.2	2.3×	16.8	2.1×
DataComp-1B	OpenCLIP	8.1	1.0×	30.0	1.0×
	ITO (λ=0)	$\sim$18	2.2×	68.6	2.3×
	ITO (λ=2)	$\sim$18	2.2×	68.6	2.3×

Two observations are worth noting. First, the overhead introduced by the fusion regularizer itself is negligible: ITO with λ=0 (alignment only) and λ=2 (alignment plus fusion) have indistinguishable cost at both scales. The training overhead almost entirely comes from the multi-view alignment pipeline. Second, this training-time cost does not propagate to inference: ITO has identical parameter count, FLOPs, and latency to a standard CLIP dual-encoder at deployment.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We summarize the contributions of the paper in the introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Limitations paragraph in Section 5 discusses the training-time overhead (Table 7), the limited set of baselines at 100M–1B scale, the modest accuracy gain and indirect geometric evidence for fusion, the sensitivity to augmentation and text-view construction, and the absence of a data-quality ablation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: There are no theoretical results in this paper.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In Section 4.1, we have provided the detailed experimental settings for results reproducing.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code is publicly available at `https://github.com/showstarpro/ITO`. All pretraining datasets are public, and the experimental setup is detailed in Section 4.1 and Appendix B.1.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We have provided the detailed experimental settings in Section 4.1.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Each configuration is trained once, since repeating large-scale image–text pretraining over multiple random seeds is computationally prohibitive (up to billion-scale data). We follow the standard evaluation protocols of prior work and report results over many benchmarks (26 zero-shot classification datasets, retrieval on three datasets, and 13 multimodal benchmarks) and multiple pretraining scales to reduce the dependence on any single measurement.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the hardware setup, software environment, and codebase in Section 4.1, and report training time and peak GPU memory in Table 7 and Appendix P.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper is conducted with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational research on training objectives for image–text contrastive pretraining, trained on public datasets and not tied to a particular application or deployment, so we do not discuss societal impacts specific to this work.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not release generative models. And this paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We conduct experiments in this paper using publicly available datasets and codebases, and cite relevant references.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.