BALANCING CROSS-MODAL SEMANTIC ALIGNMENT AND INSTANCE-LEVEL VISUAL DISCRIMINATION BY CONTROLLING CONSTRAINT DENSITY IN VISION-LANGUAGE CONTRASTIVE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models trained on large-scale image–text pairs achieve strong cross-modal semantic alignment and have become the dominant approach for cross-modal retrieval. Despite substantial progress in CLIP and its recent variants, they remain limited in the instance-level visual discrimination required for image-to-image retrieval. We first show that distilling visual discrimination knowledge from instance-oriented teachers introduces a trade-off between cross-modal semantic alignment and instance-level visual discrimination. We then identify the density of transferred neighborhood relations (constraint density) as a key factor governing the strength of this trade-off, providing a unified perspective on representative feature-, similarity-, and distribution-based distillation objectives. Based on this principle, we propose Teacher-guided Nearest-neighbor Contrastive Distillation (TNCD) as a simple realization of the lowest-constraint-density regime. Extensive experiments demonstrate that lower constraint density consistently leads to a more favorable trade-off, with TNCD serving as a practical realization that naturally supports heterogeneous teachers.

1 INTRODUCTION

Vision-language models such as CLIP (Radford et al., 2021) have become the dominant paradigm for learning transferable visual representations from large-scale image–text pairs. Subsequent studies have improved vision-language models through larger-scale pre-training (Jia et al., 2021; Cherti et al., 2023), improved data curation (Xu et al., 2023; Gadre et al., 2023; Fang et al., 2024a), alternative pre-training objectives (Yao et al., 2022; Li et al., 2022; Zhai et al., 2023), and stronger visual encoders (Sun et al., 2023; Fang et al., 2024b; Jose et al., 2025), leading to models such as SigLIP2 (Tschannen et al., 2025) with substantially improved zero-shot classification and cross-modal retrieval performance.

Despite these advances, vision-language models trained primarily on image-text pairs, without relying on additional instance-level supervision, remain considerably less effective at instance-level image retrieval, where the objective is to retrieve images of the same object instance despite changes in viewpoint, illumination, background, or context. Figure 1 illustrates representative retrieval examples from the GPR1200 benchmark (Schall et al., 2021). While OpenCLIP (Cherti et al., 2023) and SigLIP2 retrieve images sharing semantic context or dominant visual appearance, MOFI (Wu et al., 2024b), which is pretrained with large-scale image-entity supervision for instance-level visual discrimination, consistently retrieves the same object instance despite substantial appearance changes. This qualitative difference persists even for recent vision-language models, suggesting that improving image–text supervision alone does not fully address instance-level visual discrimination.

Figure 2 further compares pairwise similarities on positive image pairs from GPR1200 to isolate differences in neighborhood relations across models. Under this controlled setting, OpenCLIP and SigLIP2 assign highly consistent similarities, whereas both differ substantially from MOFI. These observations suggest that visual similarity is inherently task-dependent: different tasks favor different notions of visual similarity, which induce different neighborhood relations in the embedding

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

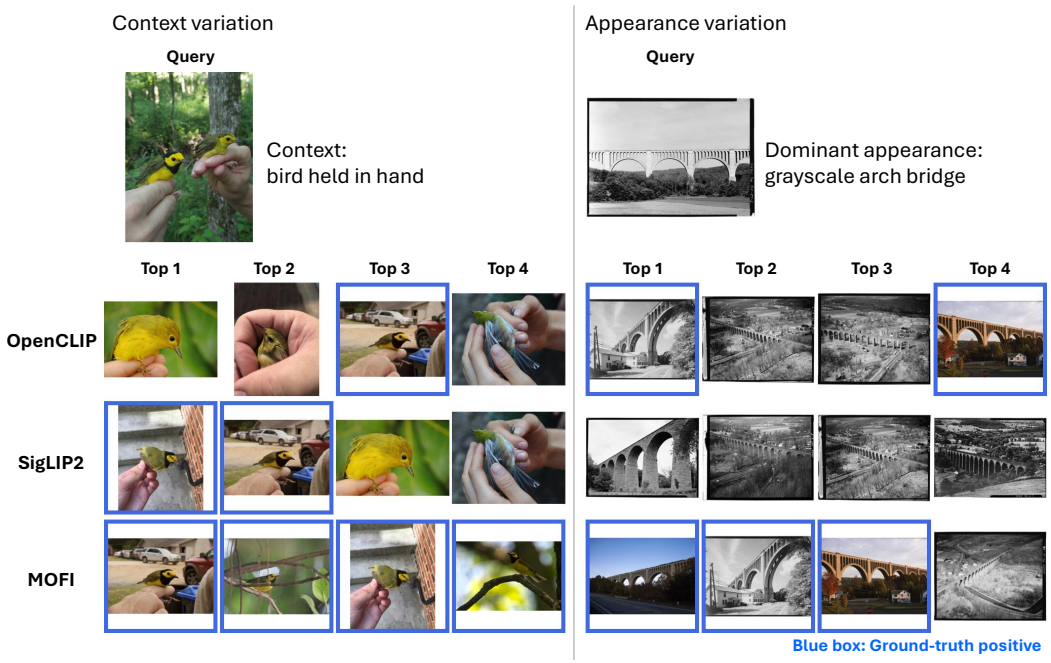


Figure 1: Representative image-to-image retrieval examples on the GPR1200 benchmark illustrating task-dependent notions of visual similarity. Colored boxes indicate ground-truth positive images of the same object instance. OpenCLIP and SigLIP2 tend to retrieve images sharing semantic context (left, bird held in hand) or dominant visual appearance (right, grayscale arch bridge), rather than the same object instance. In contrast, MOFI, trained with large-scale image–entity supervision, consistently retrieves the correct object instance despite substantial variations in viewpoint, background, illumination, and appearance.

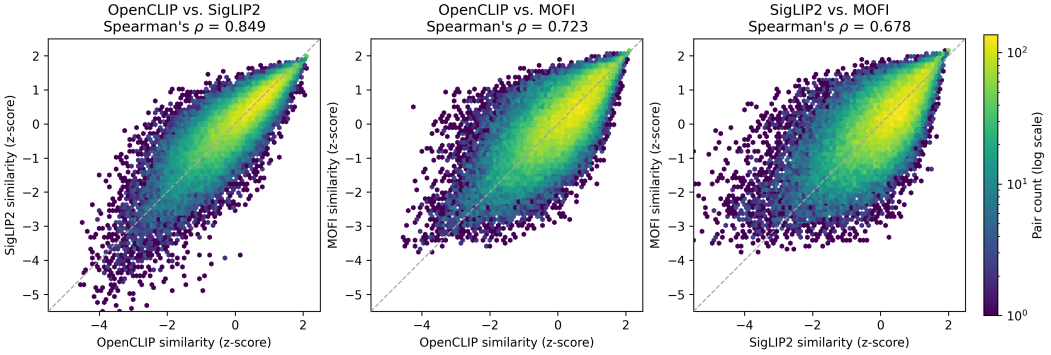


Figure 2: Pairwise similarity comparison on positive image pairs from the GPR1200 benchmark. We focus on positive image pairs depicting the same object instance to directly compare how neighborhood relations differ across models. Similarities are independently z-score normalized for each model. OpenCLIP and SigLIP2 exhibit highly consistent similarity scores (Spearman's $\rho = 0.849$), whereas both differ substantially from MOFI. Notably, image pairs assigned high similarity by MOFI span a much broader range of similarities in OpenCLIP and SigLIP2, and vice versa.

space. This naturally raises the following question: how can a single model accommodate these task-dependent neighborhood relations without sacrificing cross-modal semantic alignment?

We approach this problem from the perspective of neighborhood relation transfer. Knowledge distillation transfers neighborhood relations learned by the teacher into the student representation. When the teacher and the student learn different similarity notions, the resulting neighborhood relations are not necessarily compatible. As a consequence, transferring task-dependent neighborhood relations

through knowledge distillation gives rise to a trade-off between cross-modal semantic alignment and instance-level visual discrimination.

This perspective explains why the trade-off arises. However, a more fundamental question remains unanswered: what governs the strength of this trade-off across different distillation objectives? We identify the effective number of neighborhood relations transferred from the teacher to the student, which we refer to as *constraint density*, as a key factor governing the strength of this trade-off. This perspective provides a unified view of representative feature-, similarity-, and distribution-based distillation objectives by interpreting them as transferring neighborhood relations and characterizing their differences through constraint density. Extensive analyses across these objectives demonstrate that lower constraint density consistently leads to a more favorable trade-off between cross-modal semantic alignment and instance-level visual discrimination.

Based on this principle, we propose Teacher-guided Nearest-neighbor Contrastive Distillation (TNCD), a simple realization of the lowest-constraint-density regime. By transferring only the teacher's nearest-neighbor relation, TNCD directly realizes this principle while naturally supporting heterogeneous teachers.

The main contributions of this paper are summarized as follows.

- We show that different tasks favor different notions of visual similarity, leading models to learn distinct neighborhood relations. We further demonstrate a trade-off between cross-modal semantic alignment and instance-level visual discrimination when distilling the neighborhood relations learned by an instance-oriented teacher into a vision-language model.
- We identify the density of transferred neighborhood relations, which we refer to as constraint density, as a key factor governing the strength of this trade-off.
- We provide a unified perspective on representative feature-, similarity-, and distribution-based distillation objectives by interpreting them as transferring neighborhood relations and characterizing their differences through constraint density.
- Based on this perspective, we propose TNCD, a simple realization of the lowest-constraint-density regime, and demonstrate that lower constraint density consistently leads to more favorable trade-offs across representative distillation objectives.

2 METHOD

We study knowledge distillation from an instance-oriented teacher to a vision-language model. After formulating the problem and reviewing representative distillation objectives, we introduce constraint density as a unified perspective on neighborhood relation transfer and present TNCD as a realization of the lowest-constraint-density regime.

2.1 PROBLEM SETTING

Our objective is to improve the instance-level visual discrimination of a vision-language model while preserving its cross-modal semantic alignment. Throughout this paper, we use CLIP as the student model and an instance-oriented model (e.g., MOFI) as the teacher. We choose CLIP as a simple vision-language baseline because models trained primarily on image-text pairs consistently exhibit limited instance-level visual discrimination. Student training uses only image-text pairs together with teacher guidance, without requiring additional instance-level annotations.

Notation. Let $\boldsymbol{f}_i^m \in \mathbb{R}^D$ denote the normalized embedding of sample i produced by the student model, where $m \in v, t$ indicates the modality (v: image, t: text). Pairwise similarity is defined as $s_{ij}^{mm'} = \mathrm{sim}(\boldsymbol{f}_i^m, \boldsymbol{f}_j^{m'})$, where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. Distribution-based distillation converts pairwise similarities into probability distributions over the batch. Let $\boldsymbol{p}_i^{mm'} = (p_{i1}^{mm'}, \ldots, p_{iB}^{mm'})$ denote the corresponding student similarity distribution, where B is the batch size. The exact formulation of $\boldsymbol{p}_i^{mm'}$ is introduced in the next subsection.

Hereafter, $\boldsymbol{f}$, s, and $\boldsymbol{p}$ denote student features, pairwise similarities, and similarity distributions, respectively, and $\hat{(\cdot)}$ denotes their teacher counterparts.

2.2 Existing Distillation Objectives

The student model is optimized using the standard CLIP objective together with a distillation objective,

$$\mathcal{L} = \mathcal{L}_{\text{CLIP}} + \lambda \mathcal{L}_{\text{distill}}, \tag{1}$$

where $\mathcal{L}_{\text{CLIP}}$ denotes the standard symmetric image–text contrastive loss, $\mathcal{L}_{\text{distill}}$ denotes the distillation objective, and λ controls distillation strength. Although representative distillation objectives supervise different quantities, they can all be interpreted as transferring neighborhood relations from the teacher to the student. Below, we briefly review representative feature-, similarity-, and distribution-based objectives.

Feature matching. Feature distillation (Romero et al., 2015) directly matches the student and teacher embeddings using $\mathcal{L}_{\text{feat}} = \frac{1}{B}\sum_{i=1}^{B} \left\| \boldsymbol{f}_i^v - \hat{\boldsymbol{f}}_i^v \right\|_2^2$. Matching feature embeddings implicitly transfers the teacher's neighborhood relations to the student.

Similarity matching. Similarity-based distillation directly transfers pairwise similarity relations from the teacher to the student. Following previous work (Wu et al., 2024a), we adopt the ℓ_1 objective $\mathcal{L}_{\text{sim}} = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B} \left| s_{ij}^{vv} - \hat{s}_{ij}^{vv} \right|$

Distribution matching. Distribution-based distillation converts pairwise similarities into probability distributions and minimizes the divergence between teacher and student distributions. This formulation has been widely adopted in recent CLIP distillation methods (Wu et al., 2023; Vasu et al., 2024; Yang et al., 2024; Udandarao et al., 2025). Following the terminology of Yang et al. (2024), we refer to this objective as Contrastive Relational Distillation (CRD) throughout the paper.

Unlike image–text distillation, where self-pairs correspond to positive image–text pairs, our image-to-image setting yields trivial self-pairs with similarity equal to one. To avoid the resulting degenerate distribution, we define the probability distribution over all non-self samples within the batch as

$$p_{ij}^{vv} = \frac{\exp(s_{ij}^{vv}/\tau)}{\sum_{k\neq i}\exp(s_{ik}^{vv}/\tau)}, \qquad \hat{p}_{ij}^{vv} = \frac{\exp(\hat{s}_{ij}^{vv}/\hat{\tau})}{\sum_{k\neq i}\exp(\hat{s}_{ik}^{vv}/\hat{\tau})}, \qquad j \neq i, \tag{2}$$

where τ is the learnable student temperature and $\hat{\tau}$ is the teacher temperature controlling the sharpness of the teacher distribution. CRD transfers the teacher's neighborhood relations by matching the teacher and student probability distributions. The corresponding distillation objective is expressed as $\mathcal{L}_{\text{dist}} = \frac{1}{B}\sum_{i=1}^{B} \text{CE}\left(\hat{\boldsymbol{p}}_i^{vv}, \boldsymbol{p}_i^{vv}\right)$, where $\text{CE}(\cdot,\cdot)$ denotes the cross-entropy loss.

Thus, these seemingly different objectives can all be interpreted as transferring neighborhood relations from the teacher to the student.

2.3 Constraint Density: A Perspective on Knowledge Distillation

Since representative distillation objectives can all be interpreted as transferring neighborhood relations, a key distinction lies not in *what* is transferred, but in *how many* neighborhood relations effectively constrain the student during training. To characterize this distinction, we define *constraint density* as the effective number of neighborhood relations transferred from the teacher to the student during distillation. Higher constraint density transfers a larger number of neighborhood relations, whereas lower constraint density concentrates supervision on only a few neighborhood relations.

Feature matching can be regarded as a dense objective because matching feature embeddings implicitly constrains all pairwise neighborhood relations within the batch. Similarity matching explicitly constrains all pairwise neighborhood relations within the batch, likewise resulting in dense neighborhood relation transfer. For distribution-based objectives, the effective number of transferred

neighborhood relations for each sample can be quantified by the perplexity of the teacher distribution,

$$\text{Perplexity}_i = \exp(H(\hat{\boldsymbol{p}}_i)), \quad \text{where} \quad H(\hat{\boldsymbol{p}}_i) = -\sum_j \hat{p}_{ij} \log \hat{p}_{ij}. \tag{3}$$

A lower perplexity corresponds to a lower constraint density. Unlike feature- and similarity-based objectives, distribution-based distillation provides an explicit mechanism for continuously controlling constraint density through the teacher temperature: a lower teacher temperature produces a sharper distribution that concentrates supervision on only a few neighbors, whereas a higher temperature produces a smoother distribution that spreads supervision over a larger neighborhood. Thus, constraint density provides a common design axis for characterizing representative distillation objectives.

To understand how constraint density affects the strength of the trade-off, we consider the similarity notions underlying the two training objectives. We use the term *similarity notion* to refer to the task-dependent criterion for determining whether two samples should be regarded as similar. Learning such a similarity notion induces local *neighborhood relations*, namely the neighborhood structure that emerges in the learned embedding space. The supervision provided by $\mathcal{L}_{\text{CLIP}}$ and $\mathcal{L}_{\text{distill}}$ is based on different similarity notions and consequently different neighborhood relations. As constraint density increases, a larger number of neighborhood relations are transferred through $\mathcal{L}_{\text{distill}}$, increasing the potential for conflict with those encouraged by $\mathcal{L}_{\text{CLIP}}$ and leading to a stronger trade-off between cross-modal semantic alignment and instance-level visual discrimination. This perspective suggests that controlling constraint density provides a principled way to balance these competing objectives.

2.4 Teacher-guided Nearest-neighbor Contrastive Distillation

Based on this principle, we propose TNCD, which realizes the lowest-constraint-density regime by transferring only the teacher's nearest-neighbor relation. For each image embedding $\hat{\boldsymbol{f}}_i^v$, we identify its nearest neighbor in the teacher embedding space as $n_i = \arg\max_{j \neq i} \hat{s}_{ij}^{vv}$, where n_i denotes the teacher's nearest neighbor for sample i. The student is trained to retrieve the same nearest neighbor using

$$\mathcal{L}_{\text{TNCD}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(s_{in_i}^{vv}/\tau)}{\sum_{j \neq i} \exp(s_{ij}^{vv}/\tau)}. \tag{4}$$

By requiring only the teacher's nearest-neighbor relation, TNCD directly realizes the lowest-constraint-density regime. Unlike distribution-based distillation, it does not require teacher similarity distributions, naturally enabling knowledge distillation from heterogeneous teachers trained with different objectives.

3 Experiments

We first validate constraint density as a key factor governing the strength of the trade-off between cross-modal semantic alignment and instance-level visual discrimination through controlled finetuning experiments. We then evaluate TNCD as a practical realization of the lowest-constraint-density regime through large-scale scratch training with heterogeneous teachers.

3.1 Experimental Setup

Evaluation. We evaluate three representative capabilities of vision-language models: zero-shot image classification (10 benchmarks), cross-modal retrieval (COCO (Lin et al., 2014) and Flickr30K (Young et al., 2014)), and image-to-image retrieval on GPR1200 (Schall et al., 2021), a general-purpose retrieval benchmark constructed from six existing retrieval datasets (Wang & Jiang, 2015; Song et al., 2016; Van Horn et al., 2018; Rothe et al., 2018; Wang et al., 2019; Weyand et al., 2020) spanning diverse visual domains. Zero-shot classification is evaluated by top-1 accuracy using the dataset-specific prompt templates from the official CLIP implementation[1]. Cross-modal retrieval

[1] https://github.com/openai/CLIP/blob/main/data/prompts.md

is evaluated by the average Recall@1 of image-to-text and text-to-image retrieval, while image-to-image retrieval is evaluated by mean Average Precision (mAP). Unless otherwise specified, we focus on the trade-off between COCO retrieval and GPR1200, which directly reflects the trade-off between cross-modal semantic alignment and instance-level visual discrimination studied in this paper. Overall performance across all evaluation benchmarks is presented in Section 3.4, with detailed benchmark-wise results provided in the Appendix (Tables 8–11).

Finetuning. The finetuning experiments in Sections 3.2 and 3.3 use an OpenCLIP ViT-B/16 model initialized from the `laion400m_e32` checkpoint. We use MOFI-B/16 as the teacher and train on the LAION-400M (Schuhmann et al., 2021) image-text dataset. We optimize the student model using Eq. (1) for 50k iterations with AdamW (Loshchilov & Hutter, 2019) and a batch size of 16,384. The initial learning rate is set to 5×10^{-5} and follows a cosine decay schedule with linear warmup over the first 1k iterations.

Scratch Training. To evaluate the generality of TNCD beyond controlled finetuning, Section 3.4 additionally considers large-scale scratch training using heterogeneous teachers. Experiments follow the standard OpenCLIP training protocol. Detailed hyperparameters are provided in Appendix A.

3.2 Constraint Density Governs the Trade-off

If constraint density is a key factor governing the strength of the trade-off between cross-modal semantic alignment and instance-level visual discrimination, lower-constraint-density objectives should exhibit more favorable trade-offs across different distillation formulations.

Figure 3a compares representative distillation objectives with different levels of constraint density. For each objective, we vary the distillation weight λ and evaluate the resulting trade-off between COCO retrieval and image retrieval on GPR1200. Feature matching and similarity matching, both corresponding to dense neighborhood relation transfer, exhibit the strongest trade-offs while reaching the highest peak performance on GPR1200. CRD achieves a more favorable trade-off, with a sharper teacher distribution ($\hat{\tau} = 0.01$) outperforming a smoother one ($\hat{\tau} = 0.05$). TNCD exhibits a trade-off nearly identical to that of CRD with a sharp teacher distribution.

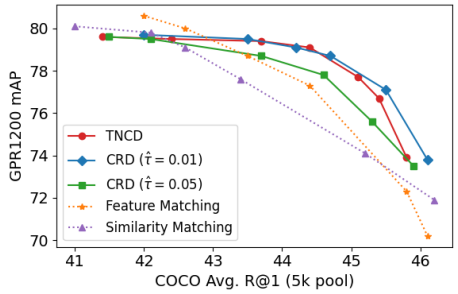
Across these representative distillation objectives, a consistent trend emerges: lower constraint density is associated with a more favorable trade-off between cross-modal semantic alignment and instance-level visual discrimination. We note, however, that feature matching and similarity matching exhibit different trade-off curves despite both corresponding to dense transfer, suggesting that *how* neighborhood relations are constrained, i.e., the specific formulation of the distillation objective, also affects the resulting trade-off. Nevertheless, the distinction between density regimes remains clear: the low-constraint-density objectives (TNCD and CRD with $\hat{\tau} = 0.01$) exhibit substantially more favorable trade-offs than the high-constraint-density objectives (feature and similarity matching), while CRD with $\hat{\tau} = 0.05$ exhibits intermediate behavior. This systematic trend across density regimes supports the proposed perspective that constraint density is a key factor governing the strength of the trade-off.

Effect of Retrieval Difficulty. We investigate how retrieval difficulty influences the observed trade-off by enlarging the COCO retrieval pool from 5k to 50k images while keeping the same 5k query set. As shown in Fig. 3b, the separation between low- and high-constraint-density objectives becomes more pronounced under the larger retrieval pool.

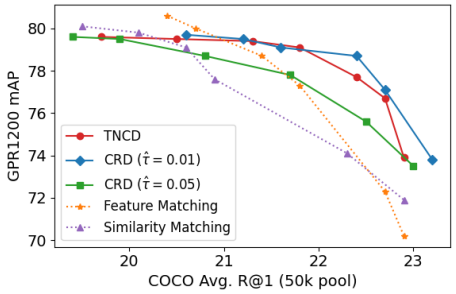
3.3 Isolating the Effect of Constraint Density

Section 3.2 reveals a consistent trend across different distillation formulations: lower constraint density is associated with a more favorable trade-off between cross-modal semantic alignment and instance-level visual discrimination. In this subsection, to isolate the effect of constraint density, we fix the distillation formulation to CRD and vary the teacher temperature.

Within CRD, the teacher temperature provides a direct mechanism for controlling constraint density through the sharpness of the teacher distribution. To quantify the resulting constraint density at each teacher temperature, we compute sample-wise perplexity according to Eq. (3) and summarize it by

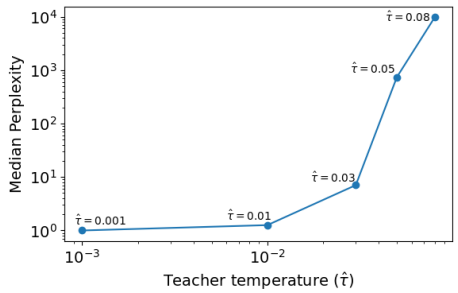


(a) Trade-off under the standard 5k-image retrieval pool.

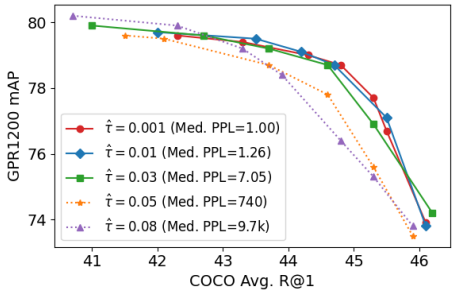


(b) Trade-off under the enlarged 50k-image retrieval pool.

Figure 3: Trade-off between COCO Avg. R@1 and GPR1200 mAP under 5k and 50k COCO retrieval pools. The retrieval pool is enlarged from 5k to 50k images while keeping the same 5k query set. The larger retrieval pool makes the trade-off between low- and high-constraint-density objectives more apparent.



(a) Median perplexity of the teacher distribution as a function of teacher temperature. Increasing teacher temperature increases median perplexity, corresponding to a higher constraint density.



(b) Trade-off between COCO Avg. R@1 and GPR1200 mAP for different teacher temperatures. The legend reports the corresponding median perplexity (Med. PPL). Higher constraint density yields a progressively stronger trade-off within the same CRD formulation.

Figure 4: Isolating the effect of constraint density within a fixed distillation formulation. (a) Teacher temperature controls constraint density through the sharpness of the teacher distribution. (b) Within CRD, increasing constraint density progressively strengthens the trade-off between cross-modal semantic alignment and instance-level visual discrimination. Nearly identical trade-off curves at $\hat{\tau} = 0.001$ and 0.01 are consistent with their similar constraint densities.

the median perplexity over all samples. The reported statistics are computed from teacher distributions collected over 100 iterations using the same batch size and training data as the finetuning experiments. Figure 4a shows the resulting median perplexity as the teacher temperature varies. At low temperatures ($\hat{\tau} = 0.001$ and 0.01), the median perplexity remains close to one, indicating that supervision is concentrated almost entirely on the teacher's nearest neighbor. As the teacher temperature increases, the median perplexity rises rapidly, indicating that supervision spreads to a larger number of neighbors.

Figure 4b shows the corresponding trade-off curves obtained using the same teacher temperatures. When the median perplexity remains close to one ($\hat{\tau} = 0.001$ and 0.01), the resulting trade-off curves are nearly identical. As the teacher temperature increases and the corresponding constraint density rises, the trade-off progressively becomes less favorable. At $\hat{\tau} = 0.08$, the trade-off becomes substantially stronger while reaching the highest peak GPR1200 performance across teacher temperatures, exhibiting behavior similar to that of dense objectives such as feature matching and similarity matching.

These results show that, within a fixed distillation formulation, increasing constraint density systematically strengthens the trade-off between cross-modal semantic alignment and instance-level

Table 1: Comparison with publicly available OpenCLIP models after large-scale scratch training. The OpenCLIP ViT-B/16 and ViT-L/14 baselines are publicly released models trained on LAION-400M for 13B seen samples, whereas our ViT-B/16 model is trained from scratch under the same dataset and sample budget. The average is computed over all evaluation benchmarks.

Model	Avg.	ImageNet-1K	COCO	GPR1200
OpenCLIP ViT-B/16 (Cherti et al., 2023)	71.2	67.1	46.9	71.3
OpenCLIP ViT-L/14 (Cherti et al., 2023)	75.9	72.8	51.3	74.7
Dual-Teacher TNCD (ViT-B/16)	75.6	72.7	53.1	79.7

visual discrimination. This also explains why TNCD and CRD with $\hat{\tau} = 0.01$ exhibit nearly identical trade-off curves in Section 3.2: when the teacher distribution has a perplexity close to one, CRD effectively operates in the same low-constraint-density regime as TNCD. Additional evidence across retrieval benchmarks and from sample-wise analyses of teacher distributions is provided in Appendix B.1.

3.4 Generalizing TNCD Beyond Controlled Finetuning

Having established the role of constraint density through controlled finetuning experiments, we next evaluate TNCD, our realization of the lowest-constraint-density regime, in large-scale scratch training. Unlike distribution-based distillation, TNCD transfers only nearest-neighbor relations and therefore does not require teacher similarity distributions, naturally enabling knowledge distillation from heterogeneous teachers.

We train an OpenCLIP ViT-B/16 model from random initialization on LAION-400M for 13B seen samples. The student is trained solely with TNCD using two heterogeneous teachers: SigLIP2-Base for cross-modal semantic supervision and MOFI-B/16 for instance-level visual discrimination. Detailed training hyperparameters are provided in Appendix A. Table 1 compares our scratch-trained model with publicly available OpenCLIP models evaluated under the same protocol. Despite using the smaller ViT-B/16 backbone, our model achieves overall performance comparable to OpenCLIP ViT-L/14 while outperforming it on both COCO retrieval and GPR1200.

Figure 5 revisits the representative failure cases shown in Figure 1. For the left example, the first four retrieved images coincide with those retrieved by MOFI, while the fifth is a semantically related image ("bird held in hand"). Similarly, for the right example, the retrieved images largely follow those of MOFI while ranking the visually closest grayscale bridge first among the positive instances. These qualitative results suggest that TNCD successfully incorporates instance-level neighborhood relations while preserving semantic preferences learned from image–text supervision.

Taken together, the quantitative and qualitative results demonstrate that TNCD naturally supports heterogeneous teachers and generalizes beyond controlled finetuning to large-scale scratch training. Additional analyses of TNCD are provided in Appendix B.2.

4 Related Work

Vision-language Models. CLIP (Radford et al., 2021) established image–text contrastive learning as the dominant paradigm for learning transferable vision-language representations, leading to rapid advances in vision-language models (Jia et al., 2021; Cherti et al., 2023; Xu et al., 2023; Gadre et al., 2023; Fang et al., 2024a; Yao et al., 2022; Li et al., 2022; Zhai et al., 2023; Sun et al., 2023; Fang et al., 2024b; Jose et al., 2025; Tschannen et al., 2025). Nevertheless, image–text supervision alone remains insufficient for learning strong instance-level visual discrimination, motivating our study of knowledge distillation for transferring instance-level knowledge to vision-language models.

Knowledge Distillation. Knowledge distillation has been widely adopted for vision-language models, including model compression (Wu et al., 2023; Vasu et al., 2024; Yang et al., 2024; Udandarao et al., 2025; Li et al., 2025) and preserving pretrained capabilities during model adaptation (Zheng et al., 2023; Yu et al., 2024; Feng et al., 2025; Yamaguchi et al., 2025). Representative objectives include feature matching (Romero et al., 2015; Yang et al., 2024), similarity or relational

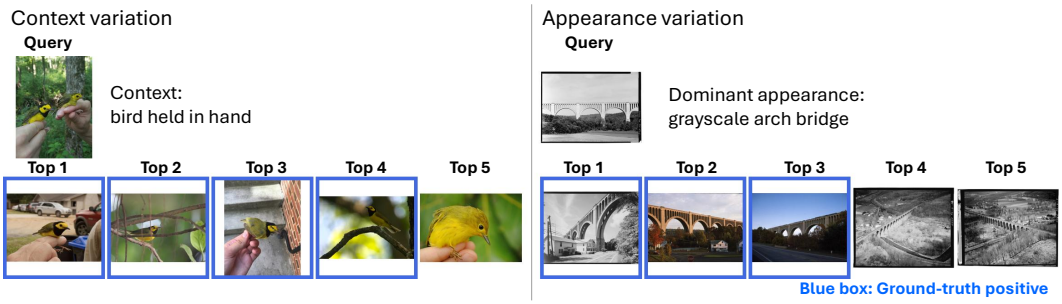


Figure 5: Representative image retrieval results of the scratch-trained model. Blue boxes indicate ground-truth positive images of the same object instance. For both examples, the model consistently retrieves the correct object instance despite substantial variations in context and appearance. Notably, the fifth retrieved image in the left example is semantically related to the query (bird held in hand), while the top-ranked positive image in the right example preserves the query's distinctive grayscale appearance.

matching (Park et al., 2019; Tung & Mori, 2019; Wu et al., 2024a), and distribution matching (Fang et al., 2021; Yang et al., 2023; Wu et al., 2023). These studies demonstrate the effectiveness of knowledge distillation across a broad range of settings. However, the trade-off that arises when transferring knowledge across models with different similarity notions and the factors governing its strength remain underexplored. Our work instead provides a unified perspective by interpreting representative distillation objectives as transferring neighborhood relations and identifies constraint density as a key factor governing the strength of the trade-off between cross-modal semantic alignment and instance-level visual discrimination.

Instance-level Visual Discrimination. Instance-level visual discrimination has long been studied in supervised image retrieval (Gordo et al., 2016) and instance-level representation learning (Liu et al., 2021). Recent advances in large-scale image–entity supervision, exemplified by MOFI (Wu et al., 2024b), have substantially improved instance-level image retrieval. Our work investigates how such instance-level knowledge can be transferred into vision-language models while preserving cross-modal semantic alignment.

5 Conclusion

In this paper, we investigated the trade-off between cross-modal semantic alignment and instance-level visual discrimination when transferring visual discrimination knowledge from an instance-oriented teacher to a vision-language model. We showed that different tasks favor different notions of visual similarity, leading to distinct neighborhood relations. These relations are not necessarily compatible, giving rise to the trade-off between cross-modal semantic alignment and instance-level visual discrimination. We then identified the density of transferred neighborhood relations, which we refer to as constraint density, as a key factor governing the strength of this trade-off. Beyond explaining the trade-off, this view provides a unified perspective on representative feature-, similarity-, and distribution-based distillation objectives: they can all be interpreted as transferring neighborhood relations, while constraint density provides a common design axis for characterizing their differences.

Finally, we proposed TNCD as a simple realization of the lowest-constraint-density regime. By transferring only nearest-neighbor relations, TNCD naturally supports heterogeneous teachers while achieving a favorable balance between cross-modal semantic alignment and instance-level visual discrimination. We hope constraint density provides a useful perspective for understanding knowledge distillation objectives and a practical design principle for developing future distillation methods.

AI USE STATEMENT

In this work, we used generative AI tools to help develop conceptual frameworks, refine hypotheses, provide feedback on research methodology and experiments, assist with translation, and interpret results. We did not use generative AI tools to implement methods. Generating synthetic datasets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, cleaning or reformatting datasets, and supporting qualitative or thematic data analysis were not applicable to this work. Additionally, we used generative AI tools for drafting parts of the manuscript, brainstorming, sourcing and searching for information, editing the manuscript to improve readability, suggesting the structure of the paper, and proposing the paper title or keywords. We reviewed all AI-assisted work and critically evaluated AI-generated suggestions before incorporating them into the research or manuscript. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We facilitate reproducibility by providing complete descriptions of the proposed method, experimental settings, and evaluation protocols. Section 3 describes the experimental setup and evaluation methodology, while the appendix provides complete training hyperparameters for all experiments together with detailed benchmark-wise evaluation results.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 446–461, 2014.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T Toshev, and Vaishaal Shankar. Data filtering networks. In *The Twelfth International Conference on Learning Representations*, 2024a.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024b.

Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. {SEED}: Self-supervised distillation for visual representation. In *International Conference on Learning Representations*, 2021.

Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 178–178, 2004.

Yuheng Feng, Changsong Wen, Zelin Peng, Li jiaye, and Siyu Zhu. Retaining knowledge and enhancing long-text representations in clip through dual-teacher distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24895–24904, 2025.

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 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 *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Albert Gordo, Jon Almazán, Jérôme Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 241–257, 2016.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 4904–4916, 2021.

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, Oriane Siméoni, Huy V. Vo, Patrick Labatut, and Piotr Bojanowski. Dinov2 meets text: A unified framework for image- and pixel-level vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24905–24916, 2025.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. In *Second Workshop on Fine-Grained Visual Categorization*, 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 12888–12900, 2022.

Yuqi Li, Chuanguang Yang, Junhao Dong, Zhengtao Yao, Haoyan Xu, Zeyu Dong, Hansheng Zeng, Zhulin An, and Yingli Tian. Ammkd: Adaptive multimodal multi-teacher distillation for lightweight vision-language models. *arXiv preprint arXiv:2509.00039*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Yu Liu, Lianghua Huang, Pan Pan, Bin Wang, Yinghui Xu, and Rong Jin. Train a one-million-way instance classifier for unsupervised visual representation learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8706–8714, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126:144–157, 2018.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115:211–252, 2015.

Konstantin Schall, Kai Uwe Barthel, Nico Hezel, and Klaus Jung. Gpr1200: A benchmark for general-purpose content-based image retrieval. *arXiv preprint arXiv:2111.13122*, 2021.

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. *arXiv preprint arXiv:2111.02114*, 2021.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Vishaal Udandarao, Nikhil Parthasarathy, Muhammad Ferjad Naeem, Talfan Evans, Samuel Albanie, Federico Tombari, Yongqin Xian, Alessio Tonioni, and Olivier J. Henaff. Active data curation effectively distills large-scale multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14422–14437, 2025.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. Mobileclip: Fast image-text models through multi-modal reinforced training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15963–15974, 2024.

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

Shuang Wang and Shuqiang Jiang. Instre: A new benchmark for instance-level object retrieval and recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11:1–21, 2015.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 - a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi (Stephen) Chen, Xinggang Wang, Hongyang Chao, and Han Hu. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21970–21980, 2023.

Sitong Wu, Haoru Tan, Zhuotao Tian, Yukang Chen, Xiaojuan Qi, and Jiaya Jia. Saco loss: Sample-wise affinity consistency for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27358–27369, 2024a.

Wentao Wu, Aleksei Timofeev, Chen Chen, Bowen Zhang, Kun Duan, Shuangning Liu, Yantao Zheng, Jonathon Shlens, Xianzhi Du, and Yinfei Yang. MOFI: Learning image representations from noisy entity annotated images. In *The Twelfth International Conference on Learning Representations*, 2024b.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

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. *arXiv preprint arXiv:2309.16671*, 2023.

Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, and Daiki Chijiwa. Post-pre-training for modality alignment in vision-language foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4256–4266, 2025.

Chuanguang Yang, Zhulin An, Helong Zhou, Fuzhen Zhuang, Yongjun Xu, and Qian Zhang. Online knowledge distillation via mutual contrastive learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10212–10227, 2023.

Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15952–15962, 2024.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022.

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

Yu-Chu Yu, Chi-Pin Huang, Jr-Jen Chen, Kai-Po Chang, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Select and distill: Selective dual-teacher knowledge transfer for continual learning on vision-language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 219–236, 2024.

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

Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19125–19136, 2023.

A Experimental Details

Fine-tuning. All fine-tuning experiments use OpenCLIP ViT-B/16 pretrained on LAION-400M as the student model. Depending on the experiment, we employ a cross-modal teacher (OpenCLIP ViT-L/14 or SigLIP2-Base), an instance-oriented teacher (MOFI-B/16), or both. The student is optimized on LAION-400M for 50k iterations using AdamW with cosine learning rate decay and linear warmup. Detailed hyperparameters are summarized in Table 2.

Throughout the appendix, we use the unified training objective

$$\mathcal{L} = \lambda_{\text{CLIP}}\mathcal{L}_{\text{CLIP}} + \lambda_{\text{cross}}\mathcal{L}^{\text{cross}}_{\text{TNCD}} + \lambda_{\text{inst}}\mathcal{L}^{\text{inst}}_{\text{TNCD}}, \tag{5}$$

where $\mathcal{L}_{\text{CLIP}}$ denotes the standard CLIP loss, $\mathcal{L}^{\text{cross}}_{\text{TNCD}}$ and $\mathcal{L}^{\text{inst}}_{\text{TNCD}}$ denote TNCD losses using the cross-modal and instance-oriented teachers, respectively, and λ_{CLIP}, λ_{cross}, and λ_{inst} are the corresponding loss weights. For each experiment, unused objectives are disabled by setting the corresponding loss weight to zero.

Training from scratch. All training-from-scratch experiments optimize the student using both a cross-modal teacher and an instance-oriented teacher with the unified training objective, setting $\lambda_{\text{CLIP}} = 0$, $\lambda_{\text{cross}} = 1$, and $\lambda_{\text{inst}} = 0.1$. The validity of omitting the CLIP objective ($\lambda_{\text{CLIP}} = 0$) when using a cross-modal teacher is verified in Appendix B.2. To evaluate the generality of TNCD across model architectures and training data, we consider two student backbones, ViT-B/16 and EVA02-Base, and use different training datasets for each setting. Detailed training hyperparameters are summarized in Tables 3 and 4, respectively.

B Additional Analyses

This section provides additional analyses supporting the proposed perspective. Section B.1 examines the relationship between constraint density and the trade-off across retrieval benchmarks and analyzes perplexity as a measure of constraint density. Section B.2 further investigates the properties of TNCD.

B.1 Additional analyses of constraint density

Generalization across retrieval benchmarks. Figure 6 presents the trade-off between Flickr30K retrieval and GPR1200 under the standard 1k-image retrieval pool and an enlarged 10k-image retrieval pool. Under the standard 1k-image evaluation, the separation between low- and high-constraint-density objectives is less pronounced than on COCO. However, enlarging the retrieval pool to 10k images reveals the same qualitative trend observed on COCO: low-constraint-density objectives (TNCD and CRD with $\hat{\tau} = 0.01$) achieve more favorable trade-offs than high-constraint-density objectives. These results further support the proposed perspective that increasing constraint density strengthens the trade-off between cross-modal semantic alignment and instance-level visual discrimination. Moreover, the clearer separation under the larger retrieval pool is consistent with our observation on COCO.

The Flickr30K curves exhibit greater variability than those on COCO, possibly due to the substantially smaller evaluation set. Nevertheless, the same qualitative trends are consistently observed across both retrieval benchmarks.

Median Perplexity as a Measure of Constraint Density. Throughout the paper, we summarize constraint density using the median perplexity of the teacher distribution. Figure 7 analyzes the sample-wise perplexity distributions for different teacher temperatures. At low teacher temperatures ($\hat{\tau} = 0.001$ and 0.01), the sample-wise perplexities are concentrated near one, resulting in nearly identical mean and median perplexities. As the teacher temperature increases, the distribution becomes increasingly right-skewed, with a small fraction of samples exhibiting substantially larger perplexities than the majority of samples. Consequently, the mean perplexity becomes increasingly influenced by these high-perplexity samples, while the median continues to represent the typical sample-wise constraint density.

Although the mean and median differ substantially at higher teacher temperatures, both exhibit the same monotonic relationship with teacher temperature. We therefore report the median perplexity

Table 2: Hyperparameters used for fine-tuning.

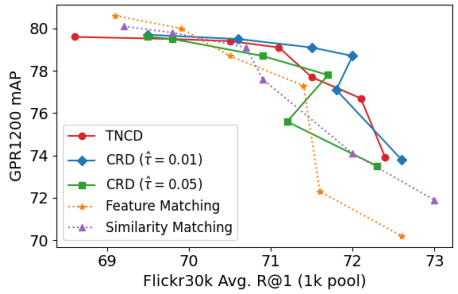
Item	Value
Student model	OpenCLIP ViT-B/16 (`laion400m_e32`)
Cross-modal teacher(s)	OpenCLIP ViT-L/14 (`laion400m_e32`) SigLIP2-Base (`patch16-224`)
Instance-oriented teacher	MOFI-B/16
Training dataset	LAION-400M
Batch size	16,384–131,072 (global)
Training iterations	50k
Warmup iterations	1k
Optimizer	AdamW
Optimizer hyperparameters	$\beta_1 = 0.9,\ \beta_2 = 0.98,\ \epsilon = 10^{-6}$
Learning rate	5×10^{-5}
Weight decay	0.2
Learning rate schedule	Cosine decay
Gradient clipping	5.0
Mixed precision	AMP (bfloat16)

Table 3: Hyperparameters used for training from scratch on ViT-B/16.

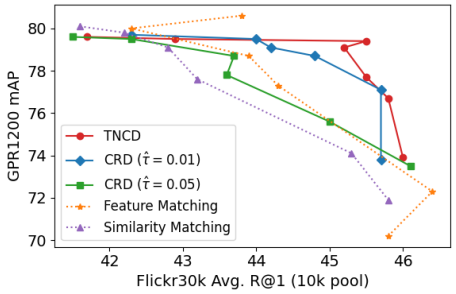
Item	Value
Student model	OpenCLIP ViT-B/16 (random initialization)
Cross-modal teacher	SigLIP2-Base (`patch16-224`)
Instance-oriented teacher	MOFI-B/16
Training dataset	LAION-400M
Batch size	131,072 (global)
Total training samples	13B
Warmup iterations	10k
Optimizer	AdamW
Optimizer hyperparameters	$\beta_1 = 0.9,\ \beta_2 = 0.98,\ \epsilon = 10^{-6}$
Learning rate	1×10^{-3}
Weight decay	0.2
Learning rate schedule	Cosine decay
Gradient clipping	5.0
Mixed precision	AMP (bfloat16)

Table 4: Hyperparameters used for training from scratch on EVA02-Base.

Item	Value
Student model	EVA02-Base (`patch14_224.mim_in22k`)
Cross-modal teacher	SigLIP2-SO400M (`patch14-224`)
Instance-oriented teacher	MOFI-L/14
Training dataset	LAION-400M COYO-700M (top 30% CLIP-score samples) CC12M
Batch size	131,072 (global)
Total training samples	8B
Warmup iterations	2k
Optimizer	AdamW
Optimizer hyperparameters	$\beta_1 = 0.9,\ \beta_2 = 0.98,\ \epsilon = 10^{-6}$
Learning rate	2×10^{-4}
Weight decay	0.05
Learning rate schedule	Cosine decay
Gradient clipping	5.0
Mixed precision	AMP (bfloat16)

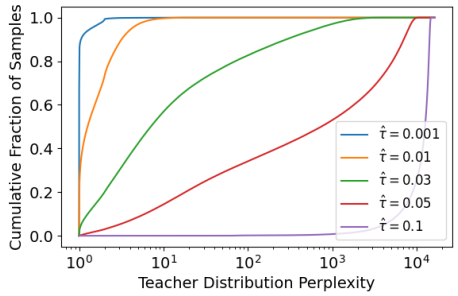


(a) Trade-off under the standard 1k-image retrieval pool.

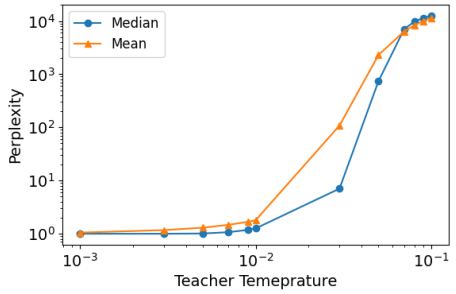


(b) Trade-off under the enlarged 10k-image retrieval pool.

Figure 6: Trade-off between Flickr30K Avg. R@1 and GPR1200 mAP under (a) the standard 1k-image retrieval pool and (b) an enlarged 10k-image retrieval pool. The separation between low- and high-constraint-density objectives becomes more pronounced with the larger retrieval pool, consistent with the observations on COCO. Low-constraint-density objectives (TNCD and CRD with $\hat{\tau} = 0.01$) exhibit more favorable trade-offs than high-constraint-density objectives.



(a) Empirical cumulative distribution functions of sample-wise teacher distribution perplexities. Higher teacher temperatures produce increasingly right-skewed distributions with a longer high-perplexity tail.



(b) Mean and median perplexities as functions of the teacher temperature. Although both increase monotonically, the discrepancy between them grows substantially at higher temperatures due to the increasingly right-skewed perplexity distribution.

Figure 7: Sample-wise perplexity distributions for different teacher temperatures. The two panels show that the median perplexity provides a more representative summary of the typical sample-wise constraint density than the mean, particularly at higher teacher temperatures.

throughout the paper, as it provides a more representative summary of the typical constraint density experienced by individual samples.

B.2 Additional Analyses of TNCD

Necessity of the CLIP objective. The scratch training experiments in Section 3.4 optimize the student using only TNCD objectives from heterogeneous teachers, without the standard CLIP objective. Here, we examine whether the CLIP objective is necessary when a cross-modal teacher is available.

Table 5 compares CLIP-only fine-tuning, joint training with the CLIP objective and cross-modal TNCD, and cross-modal TNCD alone using OpenCLIP ViT-L/14 as the cross-modal teacher. Cross-modal TNCD alone yields comparable or slightly better performance than joint training across all evaluation benchmarks, while also outperforming CLIP-only fine-tuning. These results suggest that the cross-modal teacher already provides sufficient semantic supervision and that the CLIP objective can be omitted in this setting. We therefore omit the CLIP objective in all subsequent experiments using a cross-modal teacher.

Table 5: Effect of the CLIP objective when using a cross-modal teacher. Cross-modal TNCD alone yields comparable or slightly better performance than joint training with the CLIP objective and cross-modal TNCD across all evaluation benchmarks, while also outperforming CLIP-only fine-tuning.

λ_{CLIP}	λ_{cross}	Avg.	ImageNet-1K	COCO	GPR1200
1.0	0.0	69.6	65.2	46.0	70.0
1.0	1.0	70.2	65.9	46.6	70.1
0.0	1.0	70.4	65.7	46.9	70.4

Table 6: Effect of the candidate pool size under different training settings. Settings: (A) standard CLIP; (B) cross-modal TNCD using OpenCLIP ViT-L/14 as the teacher; (C) instance-oriented TNCD using MOFI-B/16 as the teacher with $\lambda_{\text{inst}} = 1.0$; and (D) instance-oriented TNCD with $\lambda_{\text{inst}} = 0.1$.

Setting	λ_{CLIP}	λ_{cross}	λ_{inst}	batch size	Avg.	ImageNet-1K	COCO	GPR1200
A	1.0	0.0	0.0	16,384	69.6	65.2	46.0	70.0
				32,768	69.8	65.7	46.5	70.4
				65,536	69.9	66.0	45.9	70.7
				131,072	69.6	66.0	45.4	71.0
B	0.0	1.0	0.0	16,384	70.4	65.7	46.9	70.4
				32,768	71.1	66.8	48.1	70.9
				65,536	71.6	67.4	48.2	71.6
				131,072	72.1	68.1	48.9	72.1
C	1.0	0.0	1.0	16,384	70.1	66.9	42.4	79.5
				32,768	70.9	68.0	42.9	80.5
				65,536	71.3	68.6	43.1	81.3
				131,072	71.6	69.0	43.5	81.9
D			0.1	16,384	70.7	66.7	45.1	77.7
				131,072	71.3	67.9	45.9	79.2

Effect of Candidate Pool Size. TNCD transfers only the teacher's nearest-neighbor relation, making the quality of the transferred supervision dependent on the candidate pool from which the nearest neighbor is selected. Increasing the batch size enlarges this candidate pool and therefore has the potential to improve the quality of nearest-neighbor supervision.

Table 6 summarizes fine-tuning results under four training settings. For the standard CLIP objective (Setting A), enlarging the batch size provides only marginal improvements. In contrast, using OpenCLIP ViT-L/14 as the cross-modal teacher (Setting B), increasing the batch size consistently improves all evaluation benchmarks. Using MOFI-B/16 as the instance-oriented teacher (Setting C), increasing the batch size likewise improves all evaluation benchmarks. Notably, despite transferring only the teacher's nearest-neighbor relation, Setting C with the largest candidate pool achieves a GPR1200 mAP of 81.9, approaching the performance of the MOFI-B/16 teacher (83.4). The same trend is observed under the practical setting with a weaker instance-oriented distillation weight ($\lambda_{\text{inst}} = 0.1$, Setting D), where enlarging the candidate pool improves all evaluation benchmarks. Compared with the original CLIP model (Setting A, batch size 16,384), the resulting model substantially improves GPR1200 (70.0 $\rightarrow$ 79.2) while largely preserving the original COCO retrieval performance (46.0 $\rightarrow$ 45.9).

These results suggest that TNCD particularly benefits from large candidate pools, regardless of whether the teacher provides cross-modal semantic supervision or instance-oriented visual discrimination.

Table 7: Effect of heterogeneous teacher integration. The instance-oriented teacher substantially improves GPR1200, while adding cross-modal teachers progressively improves ImageNet-1K and COCO without sacrificing instance-level visual discrimination.

Teacher model		Performance			
Cross-modal	Instance-oriented	Avg.	ImageNet-1K	COCO	GPR1200
–	–	69.6	66.0	45.4	71.0
–	MOFI-B/16	71.3	67.9	45.9	79.2
OpenCLIP ViT-L/14	MOFI-B/16	73.2	69.5	48.2	79.7
SigLIP2-Base	MOFI-B/16	75.1	71.9	52.2	79.5

Integration of Heterogeneous Teachers. The scratch training experiments in Section 3.4 employ both a cross-modal teacher and an instance-oriented teacher. Here, we investigate the effect of integrating heterogeneous teachers through controlled fine-tuning experiments.

Table 7 summarizes fine-tuning results obtained with a batch size of 131,072 using $\lambda_{\text{CLIP}} = 0$, $\lambda_{\text{cross}} = 1.0$, and $\lambda_{\text{inst}} = 0.1$. Using only the instance-oriented teacher (MOFI-B/16) substantially improves GPR1200 while largely preserving the original COCO retrieval performance. Adding OpenCLIP ViT-L/14 as a cross-modal teacher further improves ImageNet-1K and COCO while maintaining the improved GPR1200 performance. Replacing OpenCLIP ViT-L/14 with the stronger SigLIP2-Base teacher yields additional improvements on the cross-modal tasks while preserving the gains in instance-level visual discrimination.

Unlike distribution-based distillation, TNCD transfers only nearest-neighbor relations and therefore does not depend on how pairwise similarities are converted into probability distributions. This formulation naturally enables knowledge distillation from heterogeneous teachers trained with different objectives, including CLIP-based and SigLIP-based models.

Generalization to large-scale scratch training. Section 3.4 demonstrated that TNCD extends to large-scale scratch training with heterogeneous teachers. Table 8 further evaluates TNCD across different student backbones and training datasets.

With OpenCLIP ViT-B/16, TNCD achieves overall performance comparable to OpenCLIP ViT-L/14 despite using a smaller backbone, while outperforming it on both COCO retrieval and GPR1200. TNCD also scales to EVA02-Base trained on a different dataset composition, achieving strong performance across all reported evaluation metrics. Detailed results on individual zero-shot classification benchmarks (Russakovsky et al., 2015; Krizhevsky, 2009; Fei-Fei et al., 2004; Krause et al., 2013; Nilsback & Zisserman, 2008; Bossard et al., 2014; Cimpoi et al., 2014; Parkhi et al., 2012; Cheng et al., 2017; Xiao et al., 2010), cross-modal retrieval benchmarks, and image retrieval benchmarks are provided in Tables 9, 10, and 11, respectively.

C Additional Retrieval Examples

Figure 1 presents representative examples illustrating that models trained with different forms of supervision exhibit different notions of visual similarity. This section provides an extended qualitative comparison on GPR1200 across diverse visual domains. Figures 8–12 present five representative examples: the first two extend the examples in Figure 1 to include our model and the top-five retrieval results, while the remaining three present additional queries. Across these examples, OpenCLIP and SigLIP2 tend to exhibit retrieval preferences driven by semantic context, dominant appearance, or presentation style, whereas MOFI favors retrievals of the same object instance. Our model largely follows the instance-level retrieval behavior of MOFI while also retrieving semantically or visually related images.

Table 8: Training-from-scratch results across different backbones and training datasets. LAION-400M+ denotes LAION-400M augmented with the top 30% CLIP-score samples from COYO-700M and the full CC12M dataset. IN-1K denotes ImageNet-1K. TNCD with OpenCLIP ViT-B/16 achieves overall performance comparable to OpenCLIP ViT-L/14 while outperforming it on both COCO retrieval and GPR1200, and further scales to EVA02-Base with a different training dataset composition.

Model	Dataset	Samples	Avg.	IN-1K	COCO	GPR1200
OpenCLIP ViT-B/16			71.2	67.1	46.9	71.3
OpenCLIP ViT-L/14	LAION-400M	13B	75.9	72.8	51.3	74.7
Ours ViT-B/16			75.6	72.7	53.1	79.7
Ours EVA02-Base	LAION-400M+	8B	78.1	76.9	53.8	82.2

Table 9: Detailed zero-shot classification results corresponding to Table 8.

Model	ImageNet-1K	CIFAR100	Caltech-101	Stanford Cars	Oxford Flowers 102	Food-101	Describable Textures	Oxford-IIIT Pets	RESISC45	SUN397
OpenCLIP ViT-B/16	67.1	70.7	83.0	83.7	69.3	86.1	51.2	89.3	64.6	67.8
OpenCLIP ViT-L/14	72.8	77.2	83.9	89.6	75.5	90.1	60.6	91.9	68.6	71.3
Ours ViT-B/16	72.7	73.9	83.7	89.4	72.6	88.4	59.4	92.2	67.1	70.2
Ours EVA02-Base	76.9	79.2	83.9	90.6	78.9	91.4	60.3	94.1	69.6	72.2

Table 10: Detailed cross-modal retrieval results corresponding to Table 8.

	COCO		Flickr30k	
Model	T2I R@1	I2T R@1	T2I R@1	I2T R@1
OpenCLIP ViT-B/16	38.3	55.4	65.6	83.5
OpenCLIP ViT-L/14	43.0	59.6	70.2	87.6
Ours ViT-B/16	44.7	61.6	72.6	87.6
Ours EVA02-Base	45.7	61.9	74.1	89.4

Table 11: Detailed image-to-image retrieval results corresponding to Table 8.

Model	mAP	Landmarks	IMSketch	iNat	Instre	SOP	faces
OpenCLIP ViT-B/16	71.3	90.0	56.0	36.0	78.1	93.1	74.6
OpenCLIP ViT-L/14	74.7	91.5	60.8	41.1	82.4	94.0	78.0
Ours: ViT-B/16	79.7	93.4	78.4	44.4	85.6	91.7	84.6
Ours: EVA02-Base	82.2	94.1	84.6	49.2	85.1	91.8	88.2

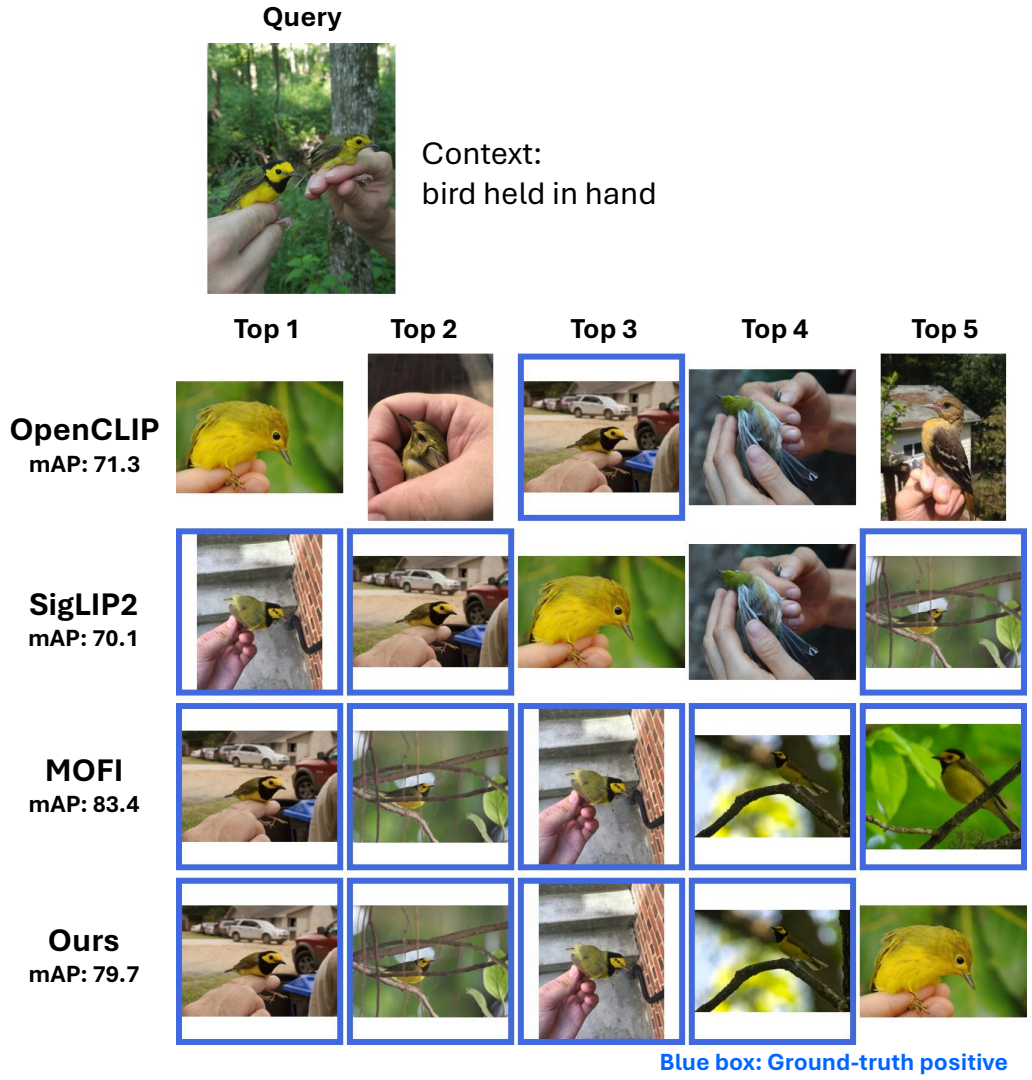


Figure 8: Detailed image-to-image retrieval results corresponding to the left example in Figure 1. The mAP reported for each model denotes its GPR1200 performance, and blue boxes indicate ground-truth positive images. OpenCLIP and SigLIP2 primarily retrieve semantically related images (e.g., birds held in hand), whereas MOFI consistently retrieves the same object instance despite context variations. Our model closely follows the instance-level retrieval behavior of MOFI, retrieving four positive instances among the top five, while its only incorrect retrieval is a semantically related image.

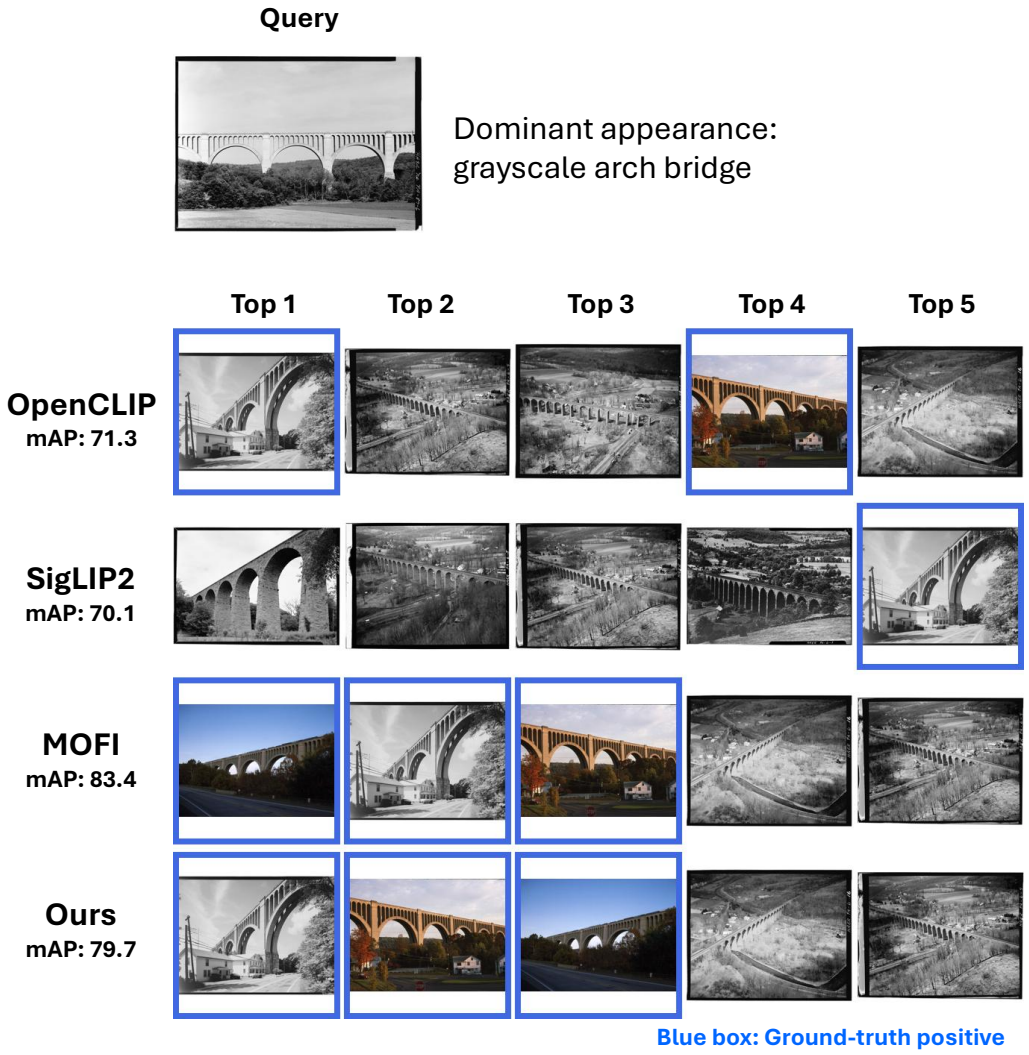


Figure 9: Detailed image-to-image retrieval results corresponding to the right example in Figure 1. The mAP reported for each model denotes its GPR1200 performance, and blue boxes indicate ground-truth positive images. OpenCLIP and SigLIP2 primarily retrieve grayscale bridge images sharing the query's dominant appearance. In contrast, both MOFI and our model retrieve the same object instance despite appearance variations. Among these positive instances, our model ranks the grayscale bridge, which most closely matches the query's appearance, first.

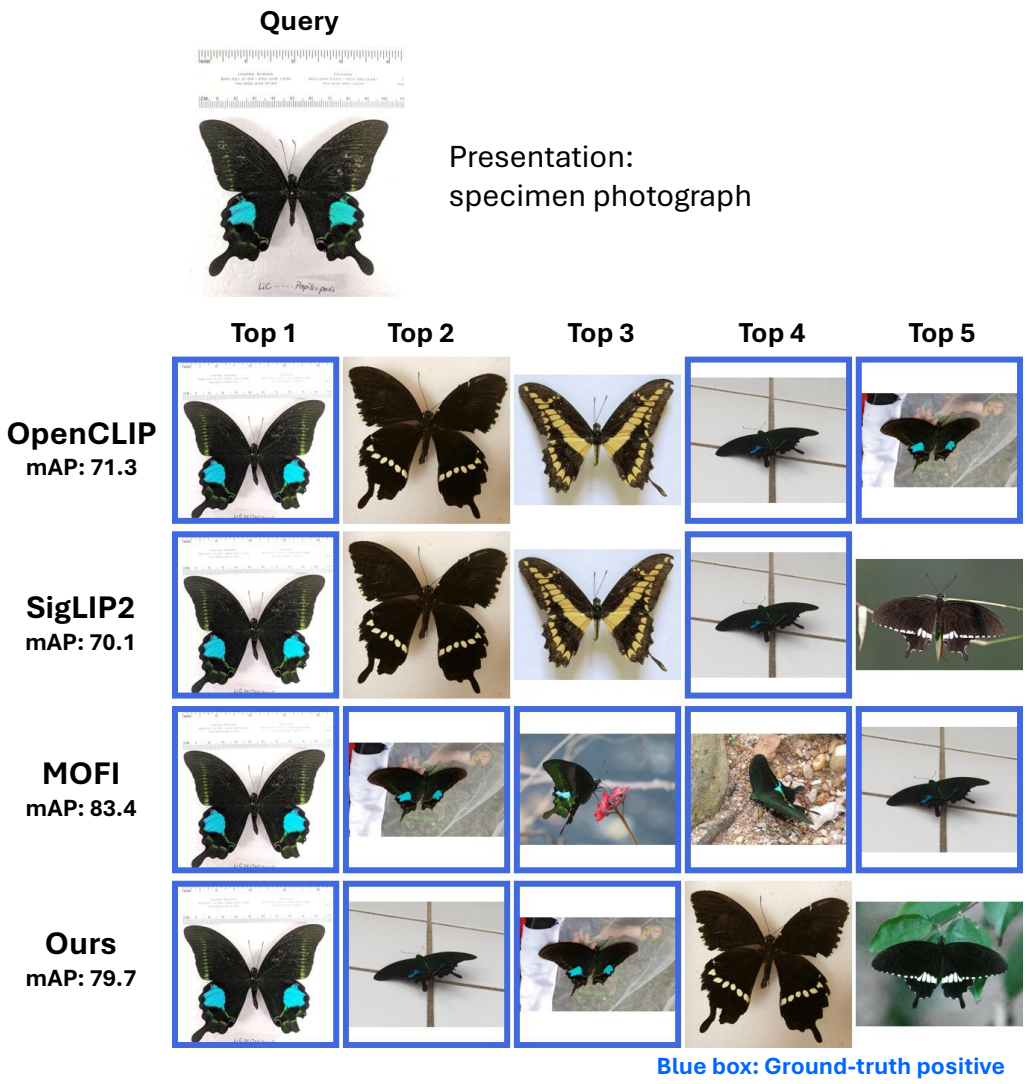


Figure 10: Additional image-to-image retrieval example from GPR1200. The mAP reported for each model denotes its GPR1200 performance, and blue boxes indicate ground-truth positive images. OpenCLIP and SigLIP2 primarily retrieve specimen-style butterfly images sharing the query's presentation and viewpoint, whereas our model retrieves the same object instance in the top three results. Its remaining incorrect retrievals are also visually similar butterflies, including a specimen-style image and a butterfly photographed in a similar spread-wing pose.

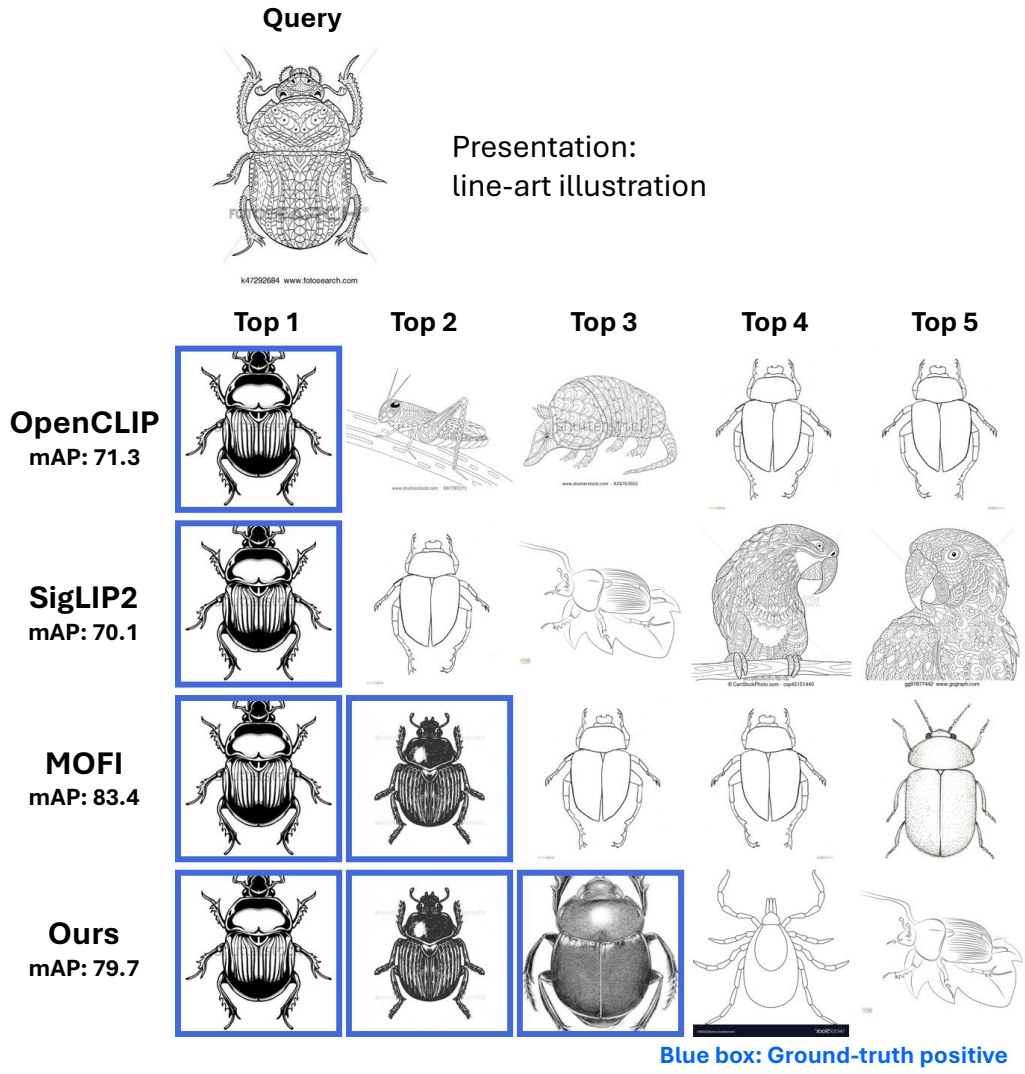


Figure 11: Additional image-to-image retrieval example from GPR1200. The mAP reported for each model denotes its GPR1200 performance, and blue boxes indicate ground-truth positive images. OpenCLIP and SigLIP2 tend to retrieve line-art illustrations sharing the query's presentation, whereas our model retrieves the same object instance despite substantial differences in drawing style and appearance.



Figure 12: Additional image-to-image retrieval example from GPR1200. The mAP reported for each model denotes its GPR1200 performance, and blue boxes indicate ground-truth positive images. OpenCLIP and SigLIP2 tend to retrieve portraits sharing the query's event context rather than the same person. In contrast, our model retrieves the same person across substantially different scenes. Its remaining incorrect retrievals are portraits sharing the query's event context.