Topology-Aware Representation Alignment for Semi-Supervised Vision-Language Learning

Junwon You[1] **Mihyun Jang**[2] **Sangwoo Mo**[2,†] **Jae-Hun Jung**[2,†]
[1]KAIST [2]POSTECH

Abstract

Vision-language models have shown strong performance, but they often generalize poorly to specialized domains. While semi-supervised vision-language learning mitigates this limitation by leveraging a small set of labeled image-text pairs together with abundant unlabeled images, existing methods remain fundamentally pairwise and fail to model the global structure of multimodal representation manifolds. Most existing topology-based alignment methods rely on persistence diagram matching, which neither guarantees geometric alignment nor uses the image-text pairing information central to vision-language learning. We propose Topology-Aware Multimodal Representation Alignment (ToMA), a framework that uses persistent homology to decompose pairwise relations into H_0-death and H_1-birth relations and aligns them across modalities through available cross-modal correspondences. This lightweight decomposition captures both connectivity and cycle-closing structure without constructing 2-simplices. Experiments show clear improvements on remote sensing, while fashion results reveal strong gains in classification and particularly strong retrieval improvements for difficult queries. Additional analysis shows that ToMA provides more consistent gains than alternative topology-based objectives and that lightweight H_1-birth edges provide useful higher-order structural signals.

1 Introduction

Vision-language models (VLMs) such as CLIP [1] have shown strong performance in learning joint image-text representations. However, they often generalize poorly to specialized domains, including remote sensing [2], medical imaging [3], and fashion, where the visual distribution deviates from that of the pretraining data. Adapting VLMs to these domains typically requires domain-specific image-text pairs, which are expensive to collect, while unlabeled images are readily available. This motivates semi-supervised vision-language learning, where adaptation is performed using a small set of labeled image-text pairs and abundant unlabeled images [4, 5].

Several recent works have extended CLIP to semi-supervised learning. S-CLIP [4] introduces a semi-supervised contrastive objective with caption-level and keyword-level pseudo-labeling. Semi-CLIP [5] builds on this idea by mining semantic concepts from labeled data and generating pseudo-labels for unlabeled samples. It further enforces cross-modal geometric consistency via trapezoidal consistency regularization over image-text embedding pairs. Despite these advances, existing objectives remain fundamentally pairwise. They align matched image-text pairs and impose geometric constraints only between pairs of embeddings (Figure 1, Pair-wise alignment).

Deep neural networks often learn representation spaces with manifold structure [6, 7]. These spaces encode semantic relationships through geometric organization, including clusters, connectivity, and higher-order structure. Preserving these geometric organizations is important for maintaining semantic consistency in learned representations [8–13]. In multimodal settings, however, semantically

[†]Equal advising. Code: https://github.com/junwon0/ToMA

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

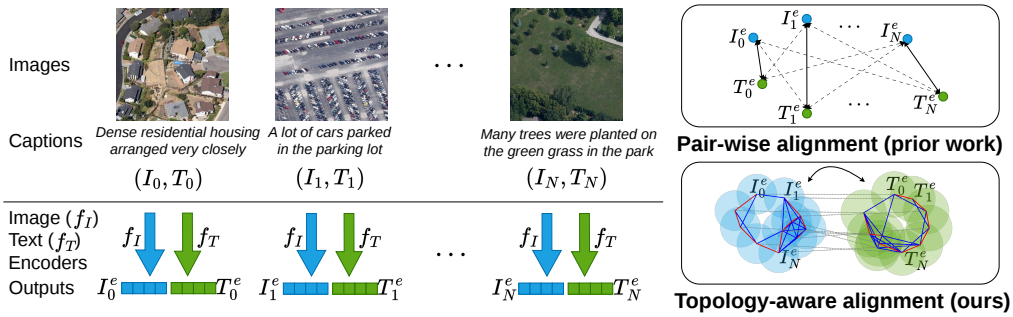


Figure 1: **Comparison between pair-wise and topology-aware alignment.** Given image-text pairs and their encoder outputs, prior approaches align corresponding embeddings independently. In contrast, our method aligns the global structural organization of image and text representation spaces by preserving topologically meaningful relationships among embeddings. This topology-aware perspective promotes cross-modal consistency at the level of the representation manifold.

related samples are often separated by the modality gap, which induces modality-specific geometric biases [14, 15]. The Platonic Representation Hypothesis suggests that, despite these differences, representations from distinct modalities may converge toward a shared abstract organization of the underlying manifold [16–19]. Under this view, cross-modal alignment should go beyond instance-level matching and instead target the shared global structure of the underlying manifold.

Topological data analysis (TDA) [20] provides a framework for characterizing the global structure of data manifolds. Rather than focusing only on pairwise distances, TDA captures structural properties that are preserved under continuous geometric transformations. In particular, it summarizes topological features of representation spaces, such as connected components and loop structures.

Most existing studies align representation spaces by matching their topological structures via persistence diagrams (PDs), which summarize the birth and death time of topological features [9–11, 13]. However, PD matching alone has two limitations in multimodal learning. First, PDs are not injective summaries of geometry: distinct representation spaces can yield identical or highly similar PDs. Thus, even if PD distances are minimized, this does not provide an inverse guarantee that the underlying representation geometries are aligned. Second, PD-based alignment does not incorporate the image-text pairing information that is central to VLMs. Moreover, most previous studies [9–13] focus on 0-dimensional persistent homology (PH), H_0. In particular, HC [12] replaces PD matching with persistence track coincidence defined over birth and death features, but their approach remains limited to H_0 because higher-dimensional PH significantly increases computational cost but brings almost no additional performance benefits.

Contribution. We propose Topology-Aware Multimodal Representation Alignment (ToMA) for semi-supervised vision-language learning. ToMA uses topology to identify topology-induced relations and aligns them using the available image-text pairing information (Figure 1, Topology-aware alignment). This design addresses the two main limitations of PD-based alignment: it avoids using PD as the alignment target, and it explicitly incorporates cross-modal pairing information.

To identify such edges, we use PH to capture topological structure. In particular, H_0 characterizes the merging of connected components and thus reflects semantic cluster structure. Beyond H_0, ToMA also leverages 1-dimensional PH (H_1) to capture higher-order relations. To keep computation lightweight, we use a filtration defined only over vertices and edges, without constructing 2-simplices (Figure 2a). Under this construction, the birth times of H_1 classes can be obtained directly from the filtration defined by Eq. 1, allowing us to exploit informative loop-level signals. Empirically, even this birth-only H_1 information provides consistent performance gains (Figure 3b).

Notably, the death simplices of H_0 recover the minimum spanning tree (MST) of the embeddings [13], which induces a unique path between any pair of points. By emphasizing these topologically significant edges and aligning their directions rather than their Euclidean distances, ToMA preserves meaningful structural relations while remaining well matched to CLIP, whose geometry is governed by cosine similarity. Together, topology-induced edge decomposition and directional alignment promote global consistency while respecting modality-specific geometry.

Across remote sensing [21–23] and fashion datasets [24–26], ToMA generally improves over the baseline. The gains are most pronounced on remote sensing, while fashion experiments reveal a complementary pattern: ToMA generally improves zero-shot classification and preferentially benefits difficult retrieval queries, accompanied by stronger cross-modal structural agreement and more semantically coherent retrieved neighborhoods. Additional analysis further shows that ToMA is more consistent than alternative topology-based objectives, and that lightweight H_1-birth edges provide effective higher-order structural signals beyond conventional H_0 alignment.

The contributions of this work are as follows:

- We propose ToMA, a topology-aware alignment framework for semi-supervised vision-language learning that aligns image and text representations at the level of global structure.
- We design a topology-aware objective $\mathcal{L}_{\text{ToMA}}$ tailored to semi-supervised vision-language learning, which leverages available image-text correspondences and respects the cosine-based geometry of CLIP representations.
- We show that topology can be effectively leveraged by decomposing pairwise relations into H_0-death and lightweight H_1-birth relations and aligning their directions, going beyond PD matching.
- Experiments on remote sensing and fashion datasets show that ToMA provides strong overall gains in zero-shot classification and provides task-dependent retrieval benefits.

2 Related Work

Vision-language pre-training (VLP) has become a core paradigm for diverse downstream vision-language tasks [1, 27–29], with many approaches proposed to learn aligned visual and textual representations [30–36]. Among them, CLIP [1] is a representative contrastive framework that learns a joint image-text embedding space and performs strongly on zero-shot classification, image-text retrieval, object detection, and few-shot learning [37–41].

Cross-modal alignment with semi-supervised learning. In semi-supervised vision-language learning, cross-modal alignment has become an effective strategy for exploiting unlabeled data [42, 43, 33]. Representative methods such as S-CLIP [4] and SemiCLIP [5] improve downstream performance by aligning image and text representations. Relatedly, recent studies have explored unimodal alignment based on the Platonic representation hypothesis [16], suggesting a broader alignment perspective beyond direct cross-modal matching [17–19]. Our method is instantiated in the former setting, while extending it to the latter direction is left for future work.

Persistent homology (PH) for preserving representation space. PH has been increasingly used to characterize and preserve the geometric structure [44–52]. Early work applies PH-based regularization: Topological Autoencoders [9] preserve the PH of the input during dimensionality reduction, while [53] enforce topology preservation through Representation Topology Divergence [54]. PH has also been adopted for knowledge transfer, including knowledge distillation [10] and semi-supervised continual learning [11]. More recently, topology-aware constraints have been introduced in multimodal learning, where HC [12] and ToMCLIP [13] regularize representations by encouraging topological consistency across modalities. Despite these advances, most prior work focuses on matching topological summaries [9–11, 13, 53]. Our method leverages the structural information carried by birth and death simplices to align topologically meaningful edges.

3 ToMA: Topology-Aware Multimodal Representation Alignment

3.1 Preliminaries

Contrastive language-image pre-training (CLIP) learns a shared embedding space for images and texts by contrasting matched and mismatched image-text pairs [1]. Given a batch $\mathcal{D}^l = \{(I_i, T_i)\}_{i=1}^N$, image and text encoders, $f_I(\cdot)$ and $f_T(\cdot)$, produce normalized embeddings I_i^e and T_i^e. The model is trained with a symmetric contrastive loss that increases the cosine similarity of matched pairs and decreases that of unmatched pairs:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{\exp(\langle I_i^e, T_i^e\rangle/\tau)}{\sum_{t=1}^{N}\exp(\langle I_i^e, T_t^e\rangle/\tau)} + \log\frac{\exp(\langle I_i^e, T_i^e\rangle/\tau)}{\sum_{t=1}^{N}\exp(\langle I_t^e, T_i^e\rangle/\tau)}\right)$$

Here, $\langle \cdot, \cdot \rangle$ denotes cosine similarity and τ is a temperature. While the original CLIP learns τ during pre-training, we fix $\tau = 1$ in our experiments following [5].

Semi-supervised vision-language alignment addresses the setting where paired image-text data $\mathcal{D}^l = \{(I_i, T_i)\}_{i=1}^N$ are scarce, while unlabeled images $\mathcal{D}^u = \{I_j\}_{j=1}^M$ are abundant with $M \gg N$. Its goal is to leverage unlabeled images to improve the alignment of the shared embedding space. SemiCLIP [5] approaches this problem with a two-stage framework. In Stage 1, it mines semantic concepts from labeled captions and trains a concept classifier together with the CLIP objective. In Stage 2, it assigns pseudo-labels to unlabeled images and fine-tunes the model using concept-level semantic consistency and caption-level trapezoidal consistency. We denote the corresponding objectives as $\mathcal{L}_{\text{Semi}}^{(1)}$ and $\mathcal{L}_{\text{Semi}}^{(2)}$, respectively, and build ToMA on top of these two objectives. A detailed explanation is provided in Appendix D.

Persistent homology (PH) studies the global structure of data beyond local pairwise relations. Given a finite set $X = \{x_i\}_{i=1}^N$ in a metric space $(\mathcal{X}, d)$, PH characterizes topology by constructing a nested family of simplicial complexes over increasing scales α. It records the appearance and disappearance of homological features, such as connected components and loops [55].

A simplicial complex K over a vertex set X is a collection of subsets of X closed under inclusion, i.e., if $\sigma \in K$ and $\tau \subseteq \sigma$, then $\tau \in K$. Each $\sigma \in K$ is a simplex with dimension $\dim(\sigma) = |\sigma| - 1$. We define a function $f : K \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ as

$$f(\sigma) = \begin{cases} 0, & |\sigma| = 1, \\ d(x, y), & \sigma = \{x, y\}, \\ \infty, & |\sigma| > 2 \end{cases} \tag{1}$$

which induces the filtration $\{K_{\alpha_i} \hookrightarrow K_{\alpha_j}\}_{\alpha_i \leq \alpha_j}$, with $K_\alpha = \{\sigma \in K \mid f(\sigma) \leq \alpha\}$. As α grows, edges are added to form a nested sequence of complexes. This is a Vietoris-Rips filtration truncated at the 1-skeleton, retaining only vertices and edges. Under this construction, PH can be computed using a minimum spanning tree, while capturing the birth events of H_1 [13]. Moreover, these H_1 birth events can be incorporated at no extra cost relative to H_0 alone. Despite this, most prior work [9–13] considers only H_0. In contrast, our method benefits from H_1 information (Figure 3b).

Fix a homology dimension $p \in \{0, 1\}$. The filtration induces homomorphisms $H_p(K_{\alpha_1}) \to H_p(K_{\alpha_2})$ for $\alpha_1 \leq \alpha_2$, enabling the tracking of p-dimensional topological features. Each PH class μ is characterized by a birth scale b_μ and a death scale d_μ, together with its corresponding birth p-simplex σ_μ^b and death $(p+1)$-simplex σ_μ^d. In particular, for $p = 0$, births occur at vertices and deaths at edges, whereas for $p = 1$, births occur at edges and deaths at 2-simplices. Because our filtration is restricted to the 1-skeleton, no 2-simplices enter at finite filtration values. Consequently, the resulting H_1 classes have no finite death times, and ToMA uses their birth edges rather than finite-persistence intervals of a full Vietoris-Rips filtration. Thus, our construction yields a homologically defined decomposition of pairwise relations into H_0-death edges and H_1-birth edges.

3.2 Our ToMA Framework

Our ToMA framework integrates two stages of SemiCLIP [5] and aligns image and text representation spaces through a topology-induced decomposition of pairwise relations rather than through persistence diagram (PD)-level matching. As illustrated in Figure 2, ToMA consists of two steps. First, under the 1-skeleton filtration by Eq. 1, the edge set is decomposed into two topologically distinct types, shown in Figure 2a. Second, these edges are aligned across image and text using cross-modal correspondences, as summarized in Figure 2b. Because the filtration is truncated at the 1-skeleton, ToMA can exploit not only the death edges of H_0 but also the birth edges of H_1 without constructing 2-simplices.

As shown in Figure 2a, the death simplices of H_0 form the edges of the MST, whereas the birth simplices of H_1 consist of all non-MST edges, each of which closes a cycle. This decomposition gives two complementary classes of topologically salient edges: H_0-death edges, which capture connectivity structure through the MST, and H_1-birth edges, which capture cycle-level relational structure through cycle-closing edges. In the context of vision-language representations, the former primarily reflects local neighborhood relations, while the latter captures higher-order organization

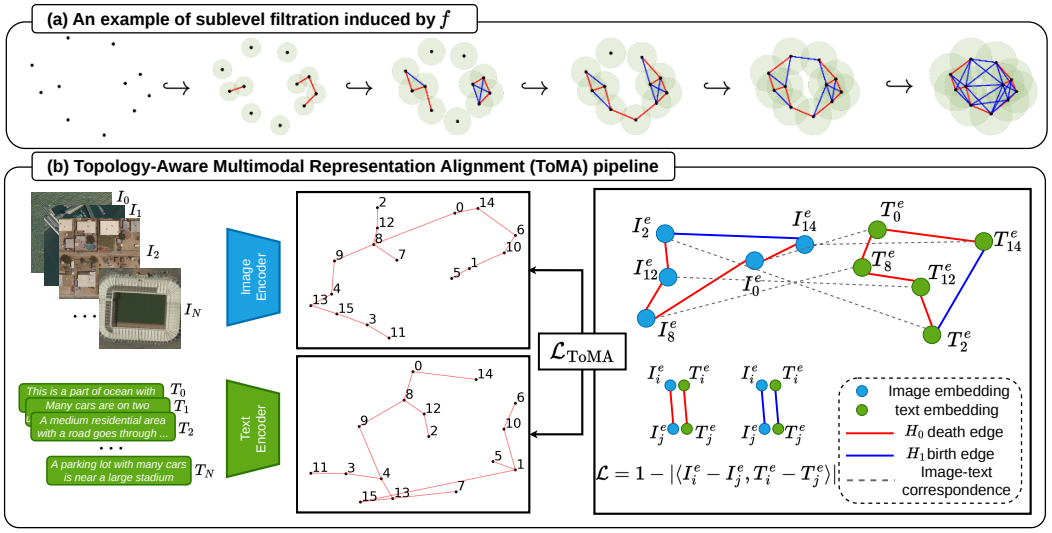


Figure 2: **Topology-Aware Multimodal Representation Alignment (ToMA). (a)** Topological decomposition of edges under the filtration induced by f. The death edges of H_0 form the minimum spanning tree (MST; red), capturing connectivity structure, while the birth edges of H_1 are cycle-closing non-MST edges (blue), capturing higher-order structure. **(b)** ToMA aligns the directions of topology-induced H_0-death and H_1-birth relations across image and text representations using cross-modal correspondences. For visual clarity, the two intermediate embedding plots in panel (b) show only the H_0-death edges.

among multiple nearby samples. This structural distinction, summarized schematically in Figure 2, is further illustrated with a concrete example based on real image-text embeddings in Appendix A.2.

We now formalize the alignment process illustrated in Figure 2b. Let $X \subset \mathbb{R}^d$ be a point cloud, and let Γ_p^X denote the set of p-dimensional PH classes of X. For each $\mu \in \Gamma_p^X$, let σ_μ^b and σ_μ^d denote its birth and death simplices, respectively. Let M_i and M_j denote the point clouds of two modalities. In our setting, M_i and M_j correspond to the image and text embeddings, respectively. We define a cross-modal pairing map $\pi_{i\to j} : M_i \to M_j$ according to the supervision available at each training stage: in stage 1, $\pi_{i\to j}$ is induced by ground-truth image-text pairs, whereas in stage 2 it is induced by ground-truth pairs for labeled samples and surrogate text embeddings for unlabeled samples.

For $\mu \in \Gamma_0^{M_i}$, let $\sigma_\mu^d = \{x_0^{(\mu_d)}, x_1^{(\mu_d)}\}$ be its death simplex in modality M_i. We denote the corresponding points in modality M_j by $\{\pi_{i\to j}(x_0^{(\mu_d)}), \pi_{i\to j}(x_1^{(\mu_d)})\}$. We define the directional consistency of this H_0-death edge as

$$T_\mu^{\text{0-death}}(M_i \to M_j) = \left\langle x_0^{(\mu_d)} - x_1^{(\mu_d)},\ \pi_{i\to j}(x_0^{(\mu_d)}) - \pi_{i\to j}(x_1^{(\mu_d)}) \right\rangle, \tag{2}$$

where $\langle u, v \rangle := u^\top v / (\|u\|_2 \|v\|_2)$ denotes the cosine similarity between the two edge-direction vectors. Similarly, for a 1-dimensional PH class $\mu \in \Gamma_1^{M_i}$, let $\sigma_\mu^b = \{x_0^{(\mu_b)}, x_1^{(\mu_b)}\}$ be its birth simplex. The directional consistency of an H_1-birth edge is defined analogously as

$$T_\mu^{\text{1-birth}}(M_i \to M_j) = \left\langle x_0^{(\mu_b)} - x_1^{(\mu_b)}, \pi_{i\to j}(x_0^{(\mu_b)}) - \pi_{i\to j}(x_1^{(\mu_b)}) \right\rangle. \tag{3}$$

We align *edge directions* rather than raw Euclidean distances because CLIP operates on normalized features and measures similarity through cosine similarity. A direction-based topological alignment therefore better matches the angular geometry. Moreover, the PH-selected edge sets in the two modalities need not coincide exactly. For this reason, we align both directions, $M_i \to M_j$ and $M_j \to M_i$. For $* \in \{\text{0-death}, \text{1-birth}\}$, we define the bidirectional ToMA loss as

$$\mathcal{L}_{\text{ToMA}}^*(M_i, M_j) = \frac{1}{\left|\Gamma_*^{M_i}\right|} \sum_{\mu \in \Gamma_*^{M_i}} \left(1 - \left|T_\mu^*(M_i \to M_j)\right|\right) + \frac{1}{\left|\Gamma_*^{M_j}\right|} \sum_{\mu \in \Gamma_*^{M_j}} \left(1 - \left|T_\mu^*(M_j \to M_i)\right|\right), \tag{4}$$

where, by a slight abuse of notation, $\Gamma_*^{M_i}$ denotes $\Gamma_0^{M_i}$ for $* = 0$-death and $\Gamma_1^{M_i}$ for $* = 1$-birth. The absolute value makes the objective invariant to whether the mapped edge direction is parallel or anti-parallel to the source edge, thereby aligning the underlying one-dimensional relational subspace rather than enforcing a fixed global orientation (Appendix F.14). For each stage $s \in \{1, 2\}$, ToMA is applied to the corresponding SemiCLIP objective: $\mathcal{L}^{(s)} = \mathcal{L}^{(s)}_{\text{Semi}} + \frac{c}{2}\left(\mathcal{L}^{\text{0-death}}_{\text{ToMA}} + \mathcal{L}^{\text{1-birth}}_{\text{ToMA}}\right)$, where c controls the strength of ToMA.

4 Experiments

4.1 Experiment Setup

Datasets. Following S-CLIP [4] and SemiCLIP [5], we evaluate our method on specialized-domain datasets, including remote sensing datasets (Section 4.2) and fashion datasets (Section 4.3). We additionally report retrieval results on science-figure and comics datasets in Appendix F.3 to assess ToMA on further specialist domains. Dataset details are provided in Appendix C.1. In the semi-supervised setting, we randomly sample 10% of the training image-text pairs as labeled data and treat the remaining 90% as unlabeled data.

Implementation details. We adopt CLIP [1] as the pretrained backbone for experiments, using ViT-B/16 by default, with additional vision encoder results in Appendix F.5. The coefficient c in $\mathcal{L}^{(s)}$ is set to 0.5 for remote sensing and 0.1 for fashion based on the hyperparameter search in Appendix F.6. We consider two variants: ToMA, which applies $\mathcal{L}_{\text{ToMA}}$ to each batch without domain information, and ToMA-domain, which applies it separately to each domain. To isolate the effect of domain partitioning from that of topology-aware alignment, we additionally evaluate a domain-partitioned SemiCLIP control without the ToMA objective in Appendix F.12. Further details are in Appendix C.2.

Baselines. We compare ToMA with CLIP (original/fine-tuned/oracle), Hard-PL [56], Soft-PL [57], S-CLIP [4], and SemiCLIP [5]. The scores of CLIP, Hard-PL, Soft-PL, and S-CLIP are imported from [5], while we retrain SemiCLIP with the same random seeds as our method for fair comparison and report originally published scores in Appendix E. We additionally conduct controlled objective-level comparisons corresponding to ToMCLIP [13] and HC [12] in Figure 3. We evaluate zero-shot classification and image-text retrieval using Top-1 accuracy (%) and Recall@K (R@K), respectively, and report the mean and standard deviation over three runs with fixed random seeds $\{0, 1, 2\}$; further implementation details are in Appendix C.2.

4.2 Remote Sensing Datasets

For remote sensing, we follow the fine-tuning setup of [2] and use RSICD [21], UCM [22], and Sydney [23], collectively denoted as RS-ALL. We randomly use 10% of RS-ALL as labeled data L and the remaining 90% as unlabeled data U. When the unlabeled data are drawn from RS-ALL, we denote the setting as L=U; when they are drawn from RESISC45 [58], we denote it as L≠U.

Zero-shot classification. We evaluate zero-shot classification on the validation splits of the RSICD and UCM classification benchmarks, denoted as RSICD-CLS and UCM-CLS, respectively, and further assess generalization on WHU-RS19 [59], RSSCN7 [60], and AID [61]. Table 1 summarizes the results. Compared with SemiCLIP, both ToMA and ToMA-domain improve zero-shot performance on most datasets under both L=U and L≠U settings. Under L=U, ToMA achieves the best results on RSICD-CLS, WHU-RS19, and AID, while ToMA-domain performs best on UCM-CLS; both methods underperform SemiCLIP only on RSSCN7. Under L≠U, ToMA consistently improves SemiCLIP on all datasets, and ToMA-domain achieves the best results on UCM-CLS and AID, with the largest gain of 4.7 p.p. on UCM-CLS over SemiCLIP. These results suggest that topology-aware alignment provides a complementary training signal that improves transferability.

Image-text retrieval. We evaluate image-text retrieval on the validation sets of RSICD, UCM, and Sydney, with R@5 results reported in Table 2. Overall, both ToMA and ToMA-domain outperform SemiCLIP on most settings. Under L=U, ToMA performs best on RSICD and UCM for image-to-text retrieval and on RSICD for text-to-image retrieval, while ToMA-domain performs best on Sydney in both directions and on UCM for text-to-image retrieval. Under L≠U, ToMA-domain improves SemiCLIP on all six retrieval metrics and achieves the best results on RSICD and Sydney

Table 1: Zero-shot classification results on remote sensing datasets. We compare original CLIP, supervised CLIP fine-tuned on labeled data (L), and semi-supervised methods using unlabeled data from the same (L=U) or different (L≠U) distribution. Parentheses denote performance gains over supervised CLIP, and green values indicate gains larger than one. Bold denotes the best semi-supervised result in each setting.

Method	Data	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID
CLIP (original)	-	59.2	60.2	81.2	69.0	59.6
CLIP (fine-tuned)	L	75.7±2.2	80.1±6.8	92.2±1.5	70.7±5.5	79.8±3.8
CLIP (oracle)	L + U	87.5±1.8	67.9±2.3	94.2±1.5	75.7±2.9	89.3±1.2
Hard-PL [56]	L=U	78.0±2.0 (+2.3)	72.4±5.2 (-7.7)	92.3±2.6 (+0.1)	71.7±4.1 (+1.0)	81.4±2.2 (+1.6)
Soft-PL [57]	L=U	81.7±0.6 (+6.0)	78.9±5.2 (-1.2)	95.3±0.8 (+3.1)	72.0±2.5 (+1.3)	85.8±1.6 (+6.0)
S-CLIP [4]	L=U	81.4±1.8 (+5.7)	81.3±3.4 (+1.2)	95.9±1.8 (+3.7)	75.1±2.0 (+4.4)	86.4±1.7 (+6.6)
SemiCLIP [5]	L=U	81.4±1.1 (+5.7)	84.3±3.2 (+4.2)	95.6±0.3 (+3.4)	**76.2**±2.7 (+5.5)	84.0±1.2 (+4.2)
ToMA (ours)	L=U	**84.8**±1.3 (+9.1)	86.1±3.4 (+6.0)	**97.9**±0.4 (+5.7)	74.2±1.7 (+3.5)	**87.3**±2.2 (+7.5)
ToMA-domain (ours)	L=U	83.8±3.2 (+8.1)	**87.5**±4.4 (+7.4)	97.8±0.6 (+5.6)	73.9±1.0 (+3.2)	86.8±2.8 (+7.0)
Hard-PL [56]	L≠U	79.9±2.5 (+4.2)	76.2±1.9 (-3.9)	91.0±2.3 (-1.2)	71.4±2.2 (+0.7)	82.4±2.0 (+2.6)
Soft-PL [57]	L≠U	78.3±3.8 (+2.6)	78.3±3.4 (-1.8)	95.3±1.3 (+3.1)	73.8±2.7 (+3.1)	82.7±3.7 (+2.9)
S-CLIP [4]	L≠U	78.3±3.2 (+2.6)	79.5±4.9 (-0.6)	93.8±1.5 (+1.6)	73.9±5.0 (+3.2)	84.9±3.3 (+5.1)
SemiCLIP [5]	L≠U	80.1±1.4 (+4.4)	82.8±1.0 (+2.7)	95.1±0.1 (+2.9)	75.0±0.8 (+4.3)	82.5±1.9 (+2.7)
ToMA (ours)	L≠U	**83.1**±1.4 (+7.4)	85.7±2.5 (+5.6)	**96.7**±0.5 (+4.5)	**75.1**±3.0 (+4.4)	85.6±3.0 (+5.8)
ToMA-domain (ours)	L≠U	82.8±1.3 (+7.0)	**87.5**±1.9 (+7.4)	96.4±0.3 (+4.2)	74.5±2.0 (+3.8)	**86.5**±1.8 (+6.7)

Table 2: Image-text retrieval results on remote sensing datasets using the same setup as in Table 1.

		Image→Text R@5			Text→Image R@5		
Method	Data	RSICD	UCM	Sydney	RSICD	UCM	Sydney
CLIP (original)	-	12.6	46.7	44.8	13.9	39.5	44.8
CLIP (fine-tuned)	L	25.7±2.3	55.2±0.8	46.6±3.0	24.9±0.6	56.3±1.6	49.4±2.0
CLIP (oracle)	L + U	34.3±1.3	63.5±1.3	63.6±2.4	32.0±1.3	66.6±1.1	65.4±2.5
Hard-PL [56]	L=U	27.3±1.1	54.8±1.6	52.3±3.6	24.2±1.3	54.6±1.0	55.2±6.9
Soft-PL [57]	L=U	27.3±0.9	54.8±2.2	52.3±2.0	26.1±1.7	55.6±4.0	56.3±5.3
S-CLIP [4]	L=U	27.5±1.1	57.0±1.5	51.1±4.3	25.6±0.6	57.3±4.1	50.6±2.6
SemiCLIP [5]	L=U	27.9±0.7	57.3±2.6	52.3±4.3	25.8±0.2	59.4±1.2	58.6±2.4
ToMA (ours)	L=U	**29.0**±1.5	**61.4**±4.6	48.9±2.1	**27.6**±1.2	60.8±3.0	56.3±3.5
ToMA-domain (ours)	L=U	28.7±1.2	59.2±3.9	**54.0**±2.9	27.0±0.9	**63.0**±2.5	**62.6**±2.9
Hard-PL [56]	L≠U	25.9±1.6	52.7±2.2	50.6±7.0	22.4±1.5	54.3±1.4	51.7±5.2
Soft-PL [57]	L≠U	27.4±1.6	54.1±2.4	51.1±1.0	24.4±0.7	54.4±2.4	52.3±5.3
S-CLIP [4]	L≠U	27.2±0.6	55.7±3.0	50.0±4.6	26.1±0.7	55.7±2.9	52.9±5.5
SemiCLIP [5]	L≠U	26.1±0.9	57.9±1.0	54.6±0.8	25.1±0.2	59.4±1.5	59.2±1.6
ToMA (ours)	L≠U	28.0±0.3	**60.6**±1.9	54.6±0.8	**26.8**±0.3	60.5±1.7	56.9±1.4
ToMA-domain (ours)	L≠U	**28.2**±0.7	58.7±1.0	**55.2**±0.0	26.4±1.0	**61.3**±1.2	**60.9**±0.8

for image-to-text retrieval and on UCM and Sydney for text-to-image retrieval, whereas ToMA performs best on UCM for image-to-text retrieval and RSICD for text-to-image retrieval. This shows that $\mathcal{L}_{\text{ToMA}}$ also benefits cross-modal retrieval.

4.3 Fashion Datasets

We conduct experiments on three fashion datasets: Fashion200k [24], FashionGen [25], and Polyvore Outfits [26]. We train the models on the union of these datasets and evaluate them separately on each dataset.

Zero-shot classification. Table 3 reports the zero-shot classification results. Compared with SemiCLIP, ToMA improves performance in most settings. The gains are particularly notable on FashionGen super-class classification, 5.7 p.p. over SemiCLIP. The domain-wise variant improves the results on FashionGen, achieving 49.1% and 46.5% on super-class and sub-class classification, respectively, and surpassing the CLIP oracle on sub-class. In contrast, the batch-wise variant performs best on Fashion200k and Polyvore, obtaining 87.3%, 45.0%, and 74.0%, respectively.

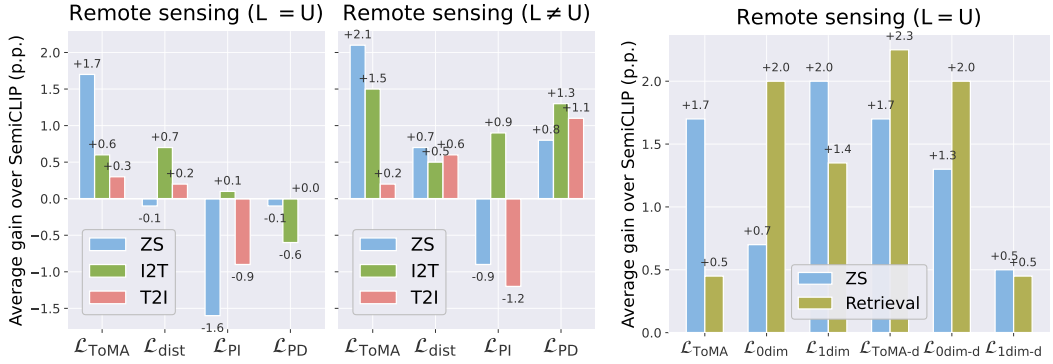
Image-text retrieval. On the standard fashion retrieval benchmarks, ToMA-domain improves SemiCLIP on five of six R@5 metrics, including all three image-to-text settings, although the aggregate gains remain modest (Appendix F.1). To better understand this behavior, we conduct a controlled

Table 3: Zero-shot classification results on fashion datasets. Parentheses indicate the performance gap from the supervised CLIP, where values highlighted in green indicate gaps larger than one.

	Fashion200k		FashionGen		Polyvore
Method	Super-class	Sub-class	Super-class	Sub-class	Class
CLIP (original)	73.2	27.7	34.8	26.4	70.2
CLIP (fine-tuned)	79.0$_{\pm 3.5}$	35.1$_{\pm 0.7}$	35.4$_{\pm 8.1}$	24.5$_{\pm 2.4}$	60.4$_{\pm 2.3}$
CLIP (oracle)	88.6$_{\pm 1.3}$	45.3$_{\pm 4.8}$	53.3$_{\pm 2.5}$	45.5$_{\pm 7.3}$	76.3$_{\pm 8.8}$
Hard-PL [56]	54.9$_{\pm 6.4}$ (-24.1)	23.9$_{\pm 2.3}$ (-11.2)	24.2$_{\pm 3.9}$ (-11.2)	18.3$_{\pm 2.4}$ (-6.2)	34.3$_{\pm 6.8}$ (-26.1)
Soft-PL [57]	82.5$_{\pm 2.8}$ (+3.5)	36.6$_{\pm 1.4}$ (+1.5)	44.8$_{\pm 3.5}$ (+9.4)	33.6$_{\pm 1.4}$ (+9.1)	73.6$_{\pm 1.7}$ (+13.2)
S-CLIP [4]	85.1$_{\pm 0.9}$ (+6.1)	38.4$_{\pm 0.7}$ (+3.3)	44.0$_{\pm 4.6}$ (+8.6)	29.6$_{\pm 4.0}$ (+5.1)	73.9$_{\pm 2.4}$ (+13.5)
SemiCLIP [5]	84.5$_{\pm 1.8}$ (+5.5)	43.7$_{\pm 1.2}$ (+8.6)	43.4$_{\pm 3.9}$ (+8.0)	43.0$_{\pm 2.6}$ (+18.5)	73.0$_{\pm 2.1}$ (+12.6)
ToMA (ours)	**87.3**$_{\pm 0.1}$ (+8.3)	**45.0**$_{\pm 0.6}$ (+9.9)	48.0$_{\pm 1.9}$ (+12.6)	40.9$_{\pm 3.9}$ (+16.4)	**74.0**$_{\pm 2.8}$ (+13.6)
ToMA-domain (ours)	86.2$_{\pm 1.4}$ (+7.2)	44.7$_{\pm 0.4}$ (+9.6)	**49.1**$_{\pm 1.8}$ (+13.7)	**46.5**$_{\pm 2.5}$ (+22.0)	72.0$_{\pm 2.9}$ (+11.6)

one-pair-per-product analysis on 7,519 FashionGen validation products (Appendix F.2). We divide queries into equally sized groups according to the bidirectional positive-negative retrieval margin of SemiCLIP. For the lowest-margin third, ToMA improves R@5 from 8.14 to 19.94 for I→T and from 7.66 to 20.10 for T→I, and improves the ground-truth rank for approximately 61% of queries. In contrast, performance decreases for some high-margin queries whose matched pairs are already well separated, partially offsetting these gains in aggregate Recall. Meanwhile, ToMA improves H_0 cross-modal neighbor overlap from 22.57% to 24.45% and edge-direction agreement from 33.21% to 42.77%, together with consistent improvements in semantic neighborhood coherence. These results indicate that topology-aware alignment improves cross-modal structural agreement and poorly resolved queries, while changes in the fine-grained ordering of already well-separated pairs limit the aggregate exact-pair retrieval gain.

4.4 Ablation Study and Analysis



(a) Comparison with alternative topology-aware objectives

(b) Homology dimension analysis

Figure 3: Comparison with prior topology-aware objectives and ablation on homology dimensions. In (a) and (b), bars denote average gains over SemiCLIP for zero-shot classification (ZS), image-to-text retrieval (I2T), text-to-image retrieval (T2I) and average of I2T and T2I (Retrieval). Here, $\mathcal{L}_{\text{PD}}$ corresponds to the objective used in ToMCLIP, while $\mathcal{L}_{\text{0dim}}$ is equivalent to HC.

Comparison with alternative topology-aware objectives. We compare $\mathcal{L}_{\text{ToMA}}$ with several alternative topology-aware objectives while keeping the underlying SemiCLIP framework and training protocol unchanged. These objectives differ in what topological information is used as the alignment target. Topological Autoencoders [9] preserve persistent homology (PH)-selected pairwise distances. Following this idea, we define $\mathcal{L}_{\text{dist}}$ using the same PH-induced relations as ToMA, but replace the directional term $1 - \left|T^*_\mu(M_i \to M_j)\right|$ in Eq. 4 with $\left|\left\|x_0^{(\mu_*)} - x_1^{(\mu_*)}\right\| - \left\|\pi_{i\to j}(x_0^{(\mu_*)}) - \pi_{i\to j}(x_1^{(\mu_*)})\right\|\right|$. Thus, $\mathcal{L}_{\text{dist}}$ aligns the lengths of PH-induced relations, whereas $\mathcal{L}_{\text{ToMA}}$ aligns their directions.

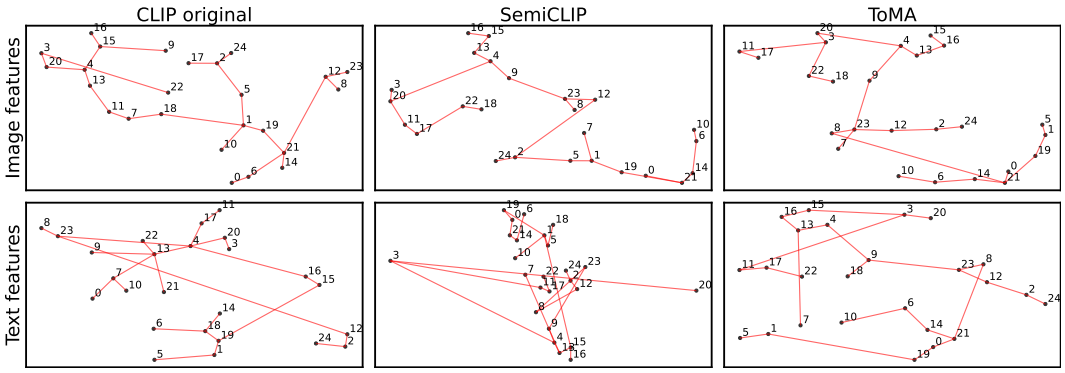


Figure 4: Topological structure analysis of image and text embeddings across methods.

We also compare against objectives that align topological summaries. TopKD [10] represents persistence diagrams (PDs) as persistence images and matches these vectorized summaries, which we denote by $\mathcal{L}_{\text{PI}}$. PsHD [11] and ToMCLIP [13] instead align PDs using the Wasserstein distance, denoted by $\mathcal{L}_{\text{PD}}$. Unlike these summary-level objectives, ToMA uses the available cross-modal correspondence map $\pi_{i\to j}$ to directly align the relations associated with H_0-death and H_1-birth events.

For a controlled comparison, we replace only $\mathcal{L}_{\text{ToMA}}$ with each alternative objective while keeping the remaining training components fixed. As shown in Fig. 3a, $\mathcal{L}_{\text{ToMA}}$ provides the most consistent improvement, particularly for zero-shot classification, whereas $\mathcal{L}_{\text{dist}}$, $\mathcal{L}_{\text{PI}}$, and $\mathcal{L}_{\text{PD}}$ show more variable behavior across classification and retrieval settings.

Contribution of H_1. We compare $\mathcal{L}_{\text{ToMA}}$ with H_0-only ($\mathcal{L}_{\text{0dim}}$, equivalent to HC [12]) and H_1-only ($\mathcal{L}_{\text{1dim}}$) variants, together with their domain-wise counterparts. As shown in Fig. 3b, the H_1-only variant already improves over SemiCLIP, demonstrating that H_1 provides useful structural information beyond H_0. Combining both dimensions yields the strongest overall results, particularly in the domain-wise setting, suggesting that cycle-closing H_1 relations provide complementary structural signals for cross-modal alignment.

Cross-modal structural alignment. To quantify whether corresponding samples induce consistent connectivity structure across modalities, we construct MSTs separately from the image and text embeddings and measure their normalized edge overlap. Table 4 shows that ToMA consistently increases this cross-modal MST agreement over SemiCLIP in both settings and across seeds (Appendix A.3). Figure 4 provides a complementary qualitative view: while CLIP and SemiCLIP exhibit substantial cross-modal connectivity mismatch, ToMA produces a more consistently organized connectivity backbone. In particular, the chain $15 \to 16 \to 13 \to 4 \to 9 \to 23 \to 12 \to 2 \to 24$ is more clearly preserved across both modalities, linking stadium-related and residential/green-building branches. Together, these results provide evidence that ToMA reduces cross-modal structural mismatch (Appendices A.1 and A.4).

Table 4: Cross-modal MST overlap under the L=U setting.

Method	Overlap
CLIP	22.11%
SemiCLIP	31.74%
ToMA	**34.26%**
ToMA-domain	**33.57%**

Comparison with topology-free graph baselines. To separate the effect of the additional alignment objective from the effect of using topology to construct the aligned relations, we replace the PH-based edge construction with random and k-NN graphs while retaining the same encoders, directional alignment objective, and training protocol. The random baseline uniformly samples pairwise relations, whereas the k-NN baseline uses nearest-neighbor relations in the current representation space; their edge counts are approximately matched to the H_0-death relations. The suffix "-d" denotes the corresponding variants in which graph construction is performed separately within each source-domain partition. Further implementation details are provided in Appendix F.8.

Table 5 summarizes the average gains over SemiCLIP. Under L=U, ToMA yields a larger zero-shot gain (+1.8 p.p.) than the random (+0.7) and k-NN (-0.3) controls, while the standard k-NN baseline is more competitive for retrieval (+0.8 vs. +0.1 for ToMA). In the domain-partitioned setting, ToMA-d achieves the largest retrieval gain (+1.9). Under L≠U, ToMA and ToMA

-d again provide larger zero-shot gains (+2.1 and +2.4) than the topology-free controls (+1.3), whereas random and k-NN remain competitive with standard ToMA for retrieval. Overall, these results show that generic graph regularization can itself be beneficial, but the topology-induced construction provides additional advantages, particularly for zero-shot classification and domain-partitioned retrieval.

Table 5: **Comparison with topology-free graph baselines.** Average gains over SemiCLIP (p.p.). Ret. denotes the average I2T/T2I R@5.

	L=U		L≠U	
Method	ZS	Ret.	ZS	Ret.
ToMA	**+1.8**	+0.1	+2.1	+0.8
Random	+0.7	-0.3	+1.3	+1.1
k-NN	-0.3	+0.8	+1.3	+1.1
ToMA-d	+1.7	**+1.9**	**+2.4**	**+1.4**
Random-d	-0.3	+0.6	+1.3	+1.1
k-NN-d	-0.3	+0.6	+1.3	+1.1

Effect of domain partitioning. We further examine whether the performance of ToMA-domain can be explained by domain partitioning itself. To this end, we evaluate SemiCLIP-domain, which applies the SemiCLIP consistency objectives separately within each source-domain subset without using the ToMA objective. As reported in Appendix F.12, domain partitioning alone provides only a marginal change in averaged zero-shot accuracy, from 84.30 for SemiCLIP to 84.48 for SemiCLIP-domain. In comparison, ToMA and ToMA-domain achieve 86.06 and 85.96, respectively. These results suggest that the main improvement is associated with the ToMA objective rather than domain partitioning itself, while domain-wise topology construction serves as an optional, dataset-dependent extension.

Robustness. Additional experiments show that ToMA remains effective across batch sizes of 32-256 and that the extracted connectivity structure is stable under moderate batch perturbations (Appendix F.9, F.10, and F.11). It is also robust to moderate choices of the loss coefficient and relative H_0/H_1 weights (Appendix F.6). Across labeled-pair ratios of 10%, 20%, and 30%, ToMA improves the five-dataset average over SemiCLIP by approximately 1.8, 2.3, and 0.6 p.p., respectively, outperforming SemiCLIP on four of the five datasets at each ratio; the gain is therefore maintained across label regimes but is not monotonic with the labeled ratio (Appendix F.13). Moreover, ToMA-domain matches or improves upon SemiCLIP in 25 of 27 seed-level aggregate comparisons (Appendix F.4).

Computational efficiency. Under a controlled comparison on a single NVIDIA A100 GPU with a batch size of 256, ToMA increases the total training time of SemiCLIP only from 799.5 to 802.0 seconds, corresponding to a 0.3% overhead. The Stage-1 run time increases from 19.2 to 19.3 seconds per epoch, while the Stage-2 run time remains unchanged at 21.3 seconds per epoch. This small overhead results from our lightweight PH computation, which extracts H_0-death and H_1-birth relations from the 1-skeleton without constructing 2-simplices (Appendix G).

5 Conclusion

We presented ToMA, a topology-aware alignment framework for semi-supervised vision-language learning that uses persistent homology to identify H_0-death and lightweight H_1-birth relations and align their directions across modalities. Across specialized-domain benchmarks, ToMA yields strong zero-shot classification gains and task-dependent retrieval benefits, with clear improvements on remote sensing and stronger effects on low-margin fashion queries and semantic neighborhood organization. These results show that topology-aware structural alignment provides a complementary signal beyond pairwise cross-modal alignment.

Limitations. Our evaluation is limited to specialized-domain semi-supervised vision-language adaptation, focusing primarily on remote sensing and fashion datasets. The additional retrieval experiments on science figures and comics provide only limited evidence beyond these domains and do not establish broader generalization to other multimodal or unimodal alignment settings. Further limitations are presented in Appendix B.1.

Broader impact. ToMA may help improve domain-specific retrieval and classification when labeled image-text pairs are scarce. At the same time, stronger multimodal alignment may be used in sensitive applications, especially in remote sensing, and may inherit biases from pretrained models and datasets, requiring careful evaluation before deployment in high-stakes settings. Further discussion of broader impacts is presented in Appendix B.2.

Acknowledgments and Disclosure of Funding

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (No. RS-2025-25397599); the Basic Science Research Institute Fund through the NRF (No. RS-2021-NR060139); the NRF grant funded by the Korea government (MSIT) (No. RS-2023-00219980); and the National Institute for Mathematical Sciences (NIMS) as a subproject of Project No. S26120000. This research was also supported in part by grants from the Institute of Information & Communications Technology Planning & Evaluation (IITP), funded by the Korea government (MSIT): No. RS-2026-25524173 (UltraLong-Term Hierarchical Memory and Reasoning Architecture for Next-Generation Omnimodal Agents) and the Global Data-X Leader HRD Program (No. IITP-2026-RS-2024-00441244).

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[2] Artashes Arutiunian, Dev Vidhani, Goutham Venkatesh, Mayank Bhaskar, Ritobrata Ghosh, and Sujit Pal. Fine tuning clip with remote sensing (satellite) images and captions. *HuggingFace Blog*, 2021.

[3] Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Bias-to-text: Debiasing unknown visual biases through language interpretation. *arXiv preprint arXiv:2301.11104*, 2:1, 2023.

[4] Sangwoo Mo, Minkyu Kim, Kyungmin Lee, and Jinwoo Shin. S-clip: Semi-supervised vision-language learning using few specialist captions. *Advances in Neural Information Processing Systems*, 36:61187–61212, 2023.

[5] Kai Gan, Bo Ye, Min-Ling Zhang, and Tong Wei. Semi-supervised clip adaptation by enforcing semantic and trapezoidal consistency. In *The Thirteenth International Conference on Learning Representations*, 2025.

[6] Xin Chen, Jian Weng, Wei Lu, Jiaming Xu, and Jiasi Weng. Deep manifold learning combined with convolutional neural networks for action recognition. *IEEE transactions on neural networks and learning systems*, 29(9):3938–3952, 2017.

[7] Uri Cohen, SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature communications*, 11(1):746, 2020.

[8] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34 (4):18–42, 2017.

[9] Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders, 2020. URL https://openreview.net/forum?id=HkgtJRVFPS.

[10] Jungeun Kim, Junwon You, Dongjin Lee, Ha Young Kim, and Jae-Hun Jung. Do topological characteristics help in knowledge distillation? In *Forty-first International Conference on Machine Learning*, 2024.

[11] Yan Fan, Yu Wang, Pengfei Zhu, Dongyue Chen, and Qinghua Hu. Persistence homology distillation for semi-supervised continual learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=qInb7EUmxz.

[12] Huatian Zhang, Lei Zhang, Yongdong Zhang, and Zhendong Mao. Homology consistency constrained efficient tuning for vision-language models. *Advances in Neural Information Processing Systems*, 37:93011–93032, 2024.

[13] Junwon You, Kang Dasol, and Jae-Hun Jung. Topological alignment of shared vision-language embedding space. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL https://openreview.net/forum?id=ecd8cgWZr6.

[14] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[15] Yonatan Gideoni, Yoav Gelberg, Tim G. J. Rudner, and Yarin Gal. Misalignment between vision-language representations in vision-language models. In *UniReps: 3rd Edition of the Workshop on Unifying Representations in Neural Models*, 2025. URL https://openreview.net/forum?id=jo2zpLRKMk.

[16] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=BH8TYy0r6u.

[17] Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Mohamed El Amine Seddik, Sanath Narayan, Karttikeya Mangalam, and Noel E O'Connor. Do vision and language encoders represent the world similarly? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14334–14343, 2024.

[18] Fabian Gröger, Shuo Wen, Huyen Le, and Maria Brbic. With limited data for multimodal alignment, let the STRUCTURE guide you. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=IkvQqD7hk3.

[19] Simon Roschmann, Paul Krzakala, Sonia Mazelet, Quentin Bouniot, and Zeynep Akata. Sotalign: Semi-supervised alignment of unimodal vision and language models via optimal transport. *arXiv preprint arXiv:2602.23353*, 2026.

[20] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2): 255–308, 2009.

[21] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017.

[22] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010.

[23] Fan Zhang, Bo Du, and Liangpei Zhang. Saliency-guided unsupervised feature learning for scene classification. *IEEE transactions on Geoscience and Remote Sensing*, 53(4):2175–2184, 2014.

[24] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pages 1463–1471, 2017.

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

[26] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European conference on computer vision (ECCV)*, pages 390–405, 2018.

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

[28] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2023.

[29] Runhui Huang, Yanxin Long, Jianhua Han, Hang Xu, Xiwen Liang, Chunjing Xu, and Xiaodan Liang. Nlip: Noise-robust language-image pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 926–934, 2023.

[30] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[31] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[32] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.

[33] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=GUrhfTuf_3.

[34] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023.

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

[36] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23390–23400, 2023.

[37] Xiang Li, Congcong Wen, Yuan Hu, and Nan Zhou. Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision. *International Journal of Applied Earth Observation and Geoinformation*, 124:103497, 2023.

[38] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.

[39] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=lL3lnMbR4WU.

[40] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European conference on computer vision*, pages 736–753. Springer, 2022.

[41] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

[42] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, 2022.

[43] Tzu-Jui Julius Wang, Jorma Laaksonen, Tomas Langer, Heikki Arponen, and Tom E Bishop. Learning by hallucinating: Vision-language pre-training with weak supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1073–1083, 2023.

[44] Hiren Madhu and Sundeep Prabhakar Chepuri. Toposrl: topology preserving self-supervised simplicial representation learning. *Advances in Neural Information Processing Systems*, 36: 64306–64317, 2023.

[45] Yuzhou Chen, Jose Frias, and Yulia R Gel. Topogcl: Topological graph contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 11453–11461, 2024.

[46] Ben Kizaric and Daniel Pimentel-Alarcón. Principle component trees and their persistent homology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13220–13229, 2024.

[47] Shihao Zhang, Kenji Kawaguchi, and Angela Yao. Deep regression representation learning with topology. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=HbdeEGVfEN.

[48] Theodore Papamarkou, Tolga Birdal, Michael Bronstein, Gunnar Carlsson, Justin Curry, Yue Gao, Mustafa Hajij, Roland Kwitt, Pietro Lio, Paolo Di Lorenzo, et al. Position: Topological deep learning is the new frontier for relational learning. *Proceedings of machine learning research*, 235:39529, 2024.

[49] Yam Eitan, Yoav Gelberg, Guy Bar-Shalom, Fabrizio Frasca, Michael M. Bronstein, and Haggai Maron. Topological blindspots: Understanding and extending topological deep learning through the lens of expressivity. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EzjsoomYEb.

[50] Wong Hiu Tung, Darrick Lee, and Hong Yan. Towards scalable topological regularizers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FjZcwQJX8D.

[51] Junwon You, Eunwoo Heo, and Jae-Hun Jung. Phlp: Sole persistent homology for link prediction - interpretable feature extraction. *Neurocomputing*, 665:132147, 2026. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2025.132147. URL https://www.sciencedirect.com/science/article/pii/S092523122502819X.

[52] Mathieu Carriere, Yuichi Ike, Théo Lacombe, and Naoki Nishikawa. Persistence-based topological optimization: a survey. *arXiv preprint arXiv:2603.24613*, 2026.

[53] Ilya Trofimov, Daniil Cherniavskii, Eduard Tulchinskii, Nikita Balabin, Evgeny Burnaev, and Serguei Barannikov. Learning topology-preserving data representations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lIu-ixf-Tzf.

[54] Serguei Barannikov, Ilya Trofimov, Nikita Balabin, and Evgeny Burnaev. Representation topology divergence: A method for comparing neural network representations. *arXiv preprint arXiv:2201.00058*, 2021.

[55] Tamal Krishna Dey and Yusu Wang. *Computational topology for data analysis*. Cambridge University Press, 2022.

[56] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013.

[57] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8443–8452, 2021.

[58] 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.

[59] Gui-Song Xia, Wen Yang, Julie Delon, Yann Gousseau, Hong Sun, and Henri Maître. Structural high-resolution satellite image indexing. In *ISPRS TC VII Symposium-100 Years ISPRS*, volume 38, pages 298–303, 2010.

[60] Qin Zou, Lihao Ni, Tong Zhang, and Qian Wang. Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and remote sensing letters*, 12(11):2321–2325, 2015.

[61] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

[62] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[63] Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. Scicap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264, 2021.

[64] Doron Adler. Simpsons blip captions. https://huggingface.co/datasets/Norod78/simpsons-blip-captions, 2023.

Appendix

A Qualitative and Quantitative Analysis **17**

A.1 Cross-Modal Structural Alignment: Qualitative Analysis 17

A.2 Illustrative Example of ToMA with the Embeddings in Figure 2b 17

A.3 Quantitative Evidence of Cross-Modal Topology Mismatch 18

A.4 Additional Examples of 0 and 1-Dimensional Edges 19

B Limitations and Broader Impacts **21**

B.1 Limitations and Future Work 21

B.2 Broader Impacts 21

C Experimental Details **21**

C.1 Dataset Details 21

C.2 Implementation Details 23

D Preliminaries of Semi-Supervised Vision-Language Alignment **24**

E Reported and Retrained SemiCLIP Results **24**

F Additional Experimental Results **26**

F.1 Image-text Retrieval Results on Fashion Datasets 26

F.2 Diagnostic Analysis of Fashion Retrieval 26

F.3 Additional Specialist-Domain Evaluation 27

F.4 Seed-wise Experimental Results 28

F.5 Different Neural Network Architectures 29

F.6 Loss Coefficient and Homology-Dimension Ablations 30

F.7 Results for R@1 31

F.8 Comparison with Topology-Free Graph Baselines 32

F.9 Results for Smaller Batch Size 33

F.10 Batch-Size Sensitivity 33

F.11 Stability of MST Edges under Batch Perturbations 34

F.12 Domain Partitioning as a Control 35

F.13 Different Labeled Image-Text Pair Ratios 36

F.14 Effect of Directional Invariance in Equation (4) 36

G Computational Time of ToMA **37**

H Compute Resources **38**

A Qualitative and Quantitative Analysis

A.1 Cross-Modal Structural Alignment: Qualitative Analysis

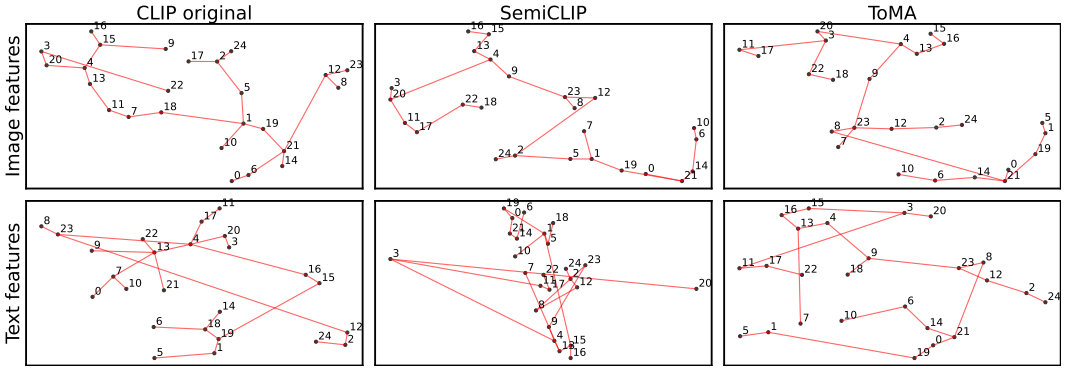


Figure 5: Topological structure analysis of image and text embeddings across methods.

Figure 5 compares the minimum spanning trees of the embeddings in Figure 6. While the original CLIP shows different connectivity patterns across image and text and SemiCLIP only partially recovers shared local structure, ToMA yields much stronger cross-modal agreement. In particular, ToMA preserves a larger connected backbone that is consistently organized across image and text, including the chain $15 \to 16 \to 13 \to 4 \to 9 \to 23 \to 12 \to 2 \to 24$. This connected structure links a stadium-related branch to a residential/green-building branch and is much less clearly aligned in SemiCLIP, especially in the text embedding space. These observations suggest that ToMA better preserves relational topology across modalities than SemiCLIP.

A.2 Illustrative Example of ToMA with the Embeddings in Figure 2b

To provide a more intuitive understanding of ToMA, we illustrate how the loss operates on the real image-text pairs shown in Figure 2b and Appendix Figure 6 (T_0–T_{15}). In this example, a mini-batch consists of 16 paired samples indexed from 0 to 15, where each image embedding I_i is matched with its corresponding text embedding T_i.

These samples exhibit several semantically related groups. For example, samples 0 and 14 both describe water scenes (ocean and beach), while samples 4, 13, and 15 are all related to stadium-like scenes. Other samples form additional local semantic neighborhoods, such as residential scenes $(2, 12)$ or vegetation-rich scenes $(8, 9)$. Thus, the batch contains multiple semantically meaningful local clusters together with higher-level relations between clusters.

Under the 1-skeleton filtration, the edge set of each modality decomposes into two types of topologically salient edges. The first type consists of the H_0-*death edges*, which form the minimum spanning tree (MST). These edges connect points that merge previously disconnected components, and therefore mainly capture direct neighborhood relations. In the example of Figure 2b, such edges connect samples within semantically coherent local groups, e.g., water-related samples or stadium-related samples. From a representation-learning perspective, they encode *local semantic proximity*.

The second type consists of the H_1-*birth edges*, which are non-MST edges that close cycles. Unlike H_0-death edges, an H_1-birth edge does not simply indicate direct proximity between two samples. Rather, it closes a loop over an existing path of H_0-death edges and thereby reflects how multiple local neighborhoods are integrated into a larger relational pattern. For example, in Figure 2b, the path $2 \to 12 \to 8 \to 0 \to 14$ is first established through H_0-death edges. This path begins with residential scenes $(2, 12)$, passes through an intermediate scene of a house-like structure surrounded by greenery (8), and then reaches water scenes $(0, 14)$. Accordingly, the path does not merely connect the closest samples, but organizes several semantically adjacent neighborhoods into a longer chain. An H_1-birth edge that closes a cycle over this chain therefore captures higher-order semantic organization beyond immediate pairwise similarity. In this sense, H_0-death edges preserve local semantic structure, whereas H_1-birth edges preserve higher-order semantic organization induced by paths formed by H_0-death edges.

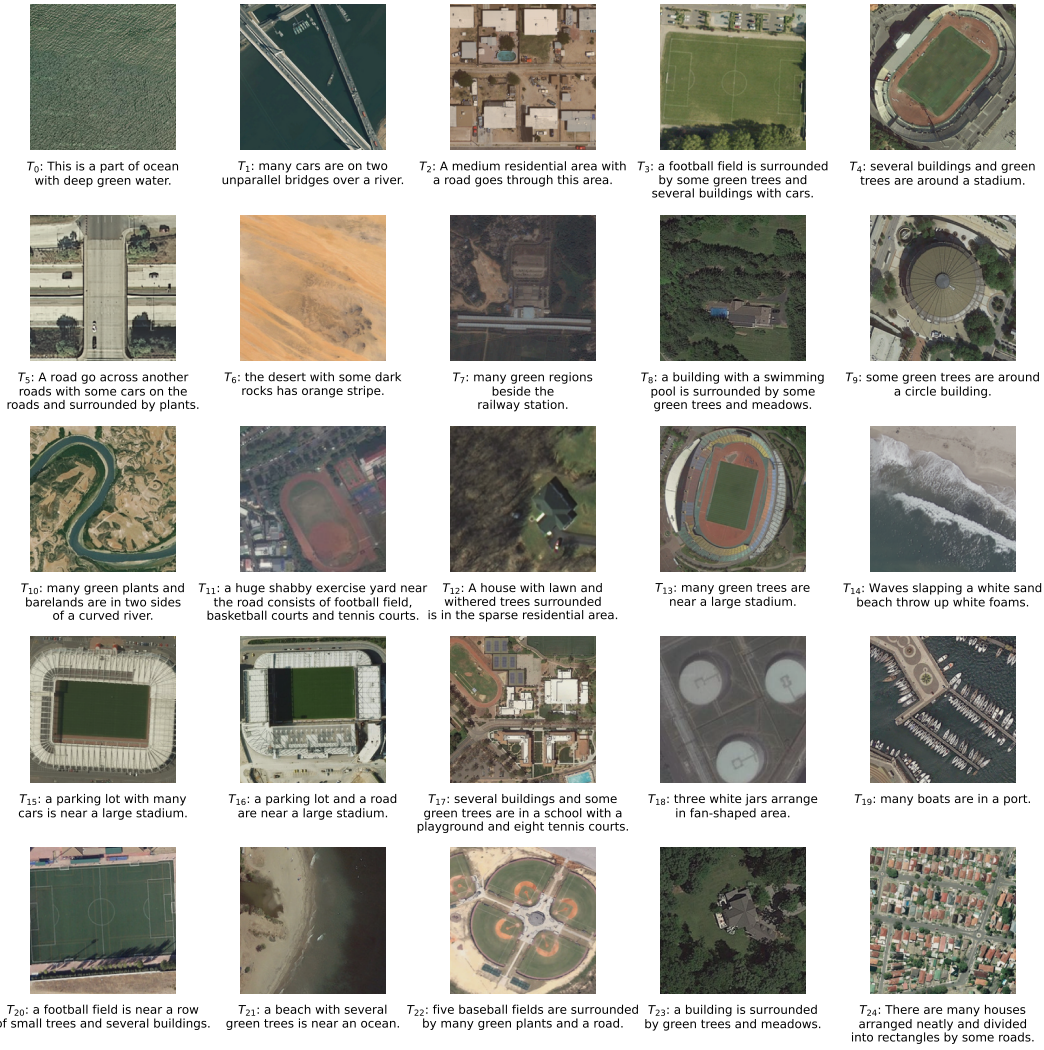


Figure 6: The real image-text data used to generate the embeddings in Figure 2b (0–15) and Figure 5.

ToMA aligns these topological relations across modalities by comparing corresponding edges in the image and text embedding spaces. For an edge (i, j) selected as a topological simplex in one modality, ToMA considers the paired relation between the corresponding samples (I_i, I_j) and (T_i, T_j). The loss encourages these two edge directions to be aligned, yielding the objective

$$\mathcal{L}(i,j) = 1 - |\,\langle I_i - I_j,\ T_i - T_j\rangle\,|,$$

where the inner product is computed after normalization.

This example clarifies the complementary roles of H_0 and H_1. The H_0-death edges constrain the connectivity backbone of the representation, while the H_1-birth edges provide additional cycle-closing relations among already connected regions. Their combination therefore captures complementary connectivity and higher-order relational structure across modalities. As a result, ToMA regularizes image and text encoders not only to match individual pairs, but also to preserve the topology of the underlying data manifolds at both local and higher-order levels.

A.3 Quantitative Evidence of Cross-Modal Topology Mismatch

Building on the illustrative example in Appendix A, Figure 5 qualitatively compares the image and text MSTs constructed from the same paired samples. The CLIP and SemiCLIP representations exhibit many modality-specific connectivity relations. This indicates that SemiCLIP improves pair-

wise cross-modal consistency without fully aligning the batch-level connectivity backbones of the two modalities. In contrast, ToMA produces visibly more consistent image-text MST structures.

To verify that this observation is systematic rather than specific to a single batch, we measure the normalized MST overlap

$$A_{\mathrm{MST}} = \frac{|M_I \cap M_T|}{N-1}, \tag{5}$$

where M_I and M_T denote the image and text MST edge sets constructed from the same batch of N paired samples. Since the H_0-death edges under our 1-skeleton filtration form the MST, this metric directly measures the agreement between the cross-modal H_0 connectivity backbones.

We evaluate 500 batches, each containing 128 image-text pairs sampled from RSICD, UCM, and Sydney according to their training-set proportions. We use the same batch composition and stochastic caption-selection procedure as during training.

Table 6: Normalized overlap between the image and text MST edge sets under the L=U setting. Higher values indicate stronger agreement between the cross-modal H_0 connectivity backbones.

Method	MST overlap (%)
CLIP (original)	22.11
SemiCLIP	31.74
ToMA	**34.26**
ToMA-domain	**33.57**

As shown in Table 6, SemiCLIP substantially increases MST agreement over the pretrained CLIP model, from 22.11% to 31.74%. Nevertheless, considerable cross-modal topology mismatch remains. ToMA and ToMA-domain further increases the overlap to 34.26% and 33.57%, corresponding to improvements of 2.51 and 1.83 percentage points over SemiCLIP, respectively. The improvement is observed for every random seed in both ToMA variants.

Together, Figure 5 and the quantitative analysis in Table 6 show that cross-modal topology mismatch is a systematic residual limitation of SemiCLIP rather than an artifact of a particular example. ToMA consistently reduces this mismatch, and the increased MST agreement is accompanied by improved average zero-shot classification and image-text retrieval performance. These results support the role of topology-aware alignment as a complementary structural objective beyond pairwise cross-modal consistency.

A.4 Additional Examples of 0 and 1-Dimensional Edges

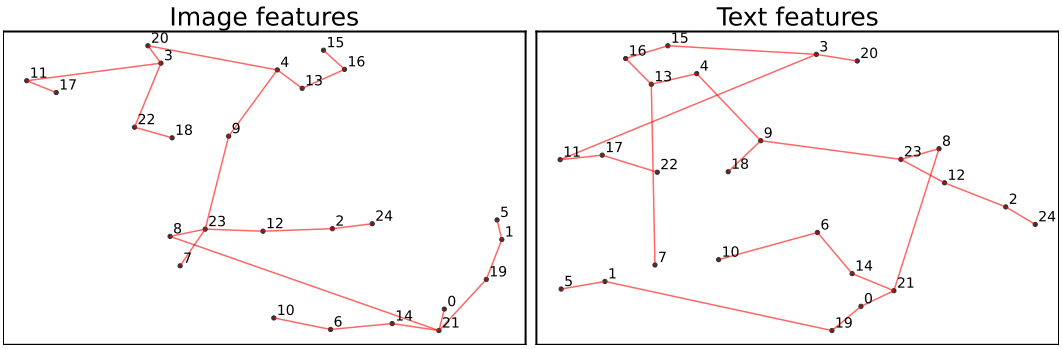


Figure 7: Minimum spanning trees of the image and text embeddings from the image-text pairs shown in Figure 6.

To further understand the structural differences highlighted in Figure 5, we provide a qualitative analysis of the H_0- and H_1-based edges extracted by ToMA on the remote sensing dataset. As shown in the main text, the original CLIP fails to align the overall MST skeletons across image and text, while SemiCLIP recovers some local structure but still leaves mismatched long-range connections and a less coherent global organization, especially in the text space. The examples

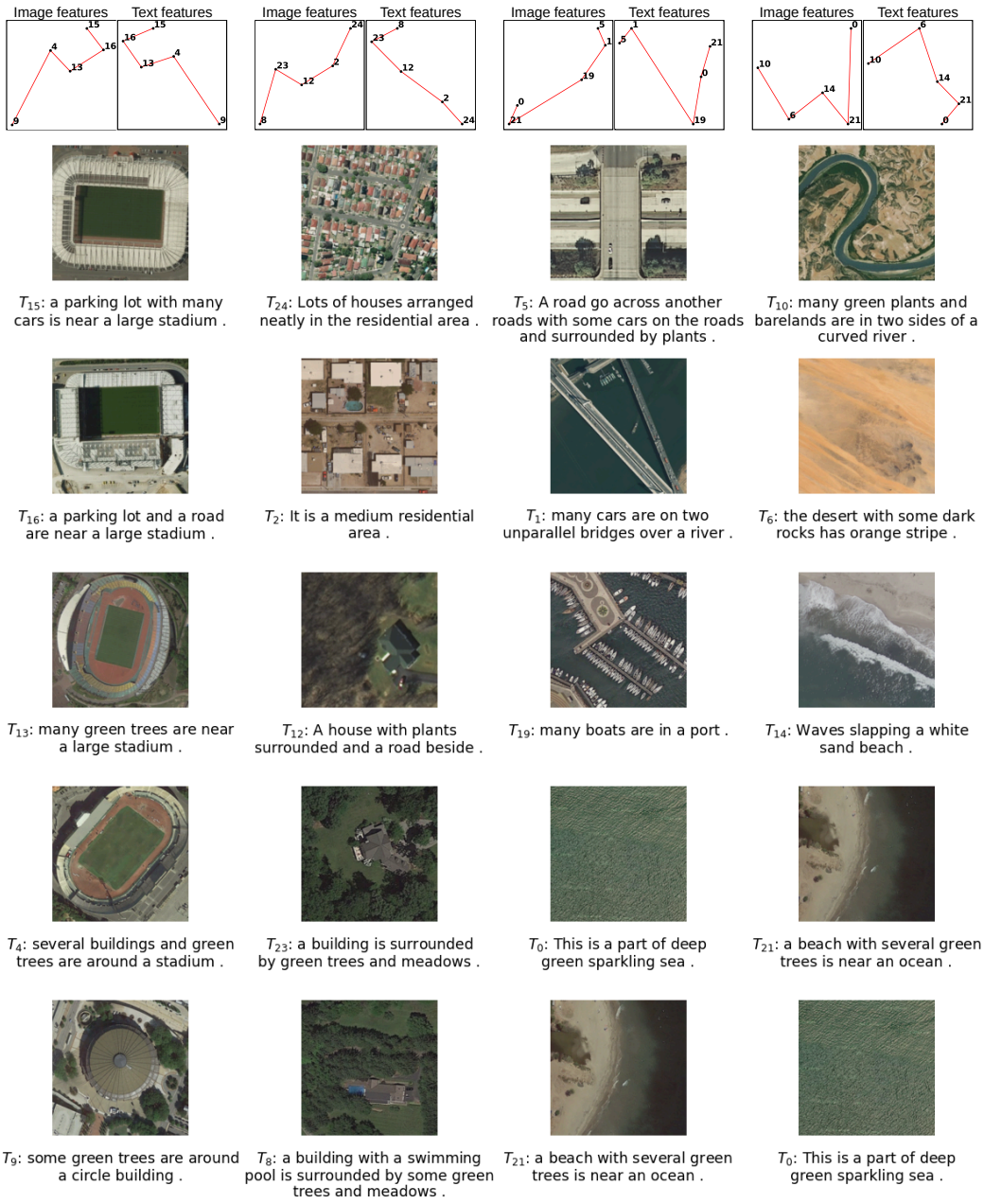


Figure 8: Visual and semantic examples of topologically salient relations identified by ToMA.

below illustrate how ToMA addresses these limitations by preserving both semantically meaningful local clusters and longer-range topological relations across modalities. Figure 7 visualizes the MST structure formed by 25 samples in Figure 6, while Figure 8 displays the actual images and captions for representative edges selected from this structure.

As shown in the individual edge configurations in Figure 8, the H_0-death edges do not merely connect points based on proximity; rather, they cluster samples by leveraging visual and semantic similarities. In the first column of Figure 8, the edges sequentially connect samples that share distinctive visual features of stadiums. In the second column, the edges group samples under the common theme of residential areas. Notably, the semantic flow from high-density residential areas (24) to vegetation-mixed low-density residential areas $(12, 23, 8)$ is well-preserved along the H_0-death edges.

Furthermore, H_1-birth edges complete the logical connections between different semantic clusters, thereby forming global cycles. The connection path $10 \to 6 \to 14 \to 21 \to 0$ in the last column of Figure 8 illustrates a significant geographical transition spanning from a river (10) and desert (6) to a beach (14) and the ocean $(21, 0)$. By aligning the H_1-birth edges that close these long-range paths into cycles, ToMA regularizes the model to maintain the topological characteristics of the global data manifold, consistently preserving transitions that start from a river, pass through the sea, and return to terrestrial structures across both modalities.

Comparing the image and text features in Figure 7, several corresponding connectivity patterns are preserved across the two modalities. In particular, the stadium- and residential-related branches observed in the image space are also reflected in the text space. This suggests that ToMA preserves batch-level relational organization beyond individual image-text pairs. Together, these examples help explain the increased cross-modal MST agreement observed quantitatively in Table 6.

B Limitations and Broader Impacts

B.1 Limitations and Future Work

Our work has several limitations that also point to promising directions for future research. First, we evaluate ToMA only in specialized-domain semi-supervised vision-language adaptation settings, specifically remote sensing, fashion, science figures and comics, and on two downstream tasks: zero-shot classification and image-text retrieval. Although the results are encouraging in these settings, they do not yet establish that the method generalizes to broader multimodal adaptation scenarios. Extending the evaluation to more diverse domains and tasks would help clarify the scope and robustness of the proposed approach. Second, our study is restricted to cross-modal semi-supervised alignment. While recent work has explored broader unimodal alignment perspectives, we do not examine whether the proposed topology-aware edge alignment transfers effectively to those settings. Investigating such extensions could reveal whether the structural bias introduced by ToMA is useful beyond the current paired multimodal formulation. Third, ToMA relies on a lightweight topological construction defined on the 1-skeleton, using H_0-death and H_1-birth edge-level signatures without constructing higher-order simplices. This design keeps the method computationally practical, but it may miss richer higher-order topological structure in the representation space. A natural direction for future work is therefore to explore whether incorporating higher-order simplicial constructions can provide additional benefits while maintaining reasonable computational cost. Finally, the benefits of combining H_0- and H_1-based signals are not uniform across tasks. While the joint formulation provides a favorable trade-off for zero-shot classification, its advantages are less consistent across retrieval settings. This suggests that more adaptive ways of weighting, selecting, or scheduling topological signals according to the task or data characteristics may further improve performance.

B.2 Broader Impacts

This work studies topology-aware representation alignment for semi-supervised vision-language learning in specialized domains, with experiments on remote sensing and fashion datasets. Potential positive impacts include improving adaptation of vision-language models when labeled image-text pairs are scarce, which may support applications in domain-specific retrieval and classification. At the same time, methods for stronger representation alignment may also be used in sensitive settings, especially in remote sensing, where improved modeling could contribute to surveillance-related or other privacy-sensitive downstream applications. More broadly, the method may inherit biases or limitations present in the underlying pretrained model and benchmark datasets. We therefore view this work as a methodological contribution and encourage careful evaluation and domain-specific oversight before deployment in high-stakes settings.

C Experimental Details

C.1 Dataset Details

Remote sensing. Our training pipeline utilizes a combined collection of RSICD [21], UCM [22], and Sydney [23], referred to as the RS-ALL suite. To implement our semi-supervised framework,

a random 10% subset of these image-text pairs is designated as labeled data (L), while the remaining 90% serves as the unlabeled pool (U). To evaluate the model's robustness against distribution shifts, we introduce the RESISC45 [58] dataset, which consists of 31,500 images across 45 scene categories, as an out-of-distribution unlabeled pool ($L \neq U$). The numerical distribution of RS-ALL is detailed in Table 7, and visual examples are illustrated in Figure 9. During training, if an image is associated with multiple textual descriptions, a single caption is selected stochastically at each epoch to ensure linguistic variety.

Table 7: Statistics of image-text paired train datasets in the remote sensing domain.

Dataset	RSICD	UCM	Sydney
# of Pairs	8,734	1,680	497



Figure 9: Examples of image-text pairs from the Remote sensing datasets.

For zero-shot classification, we employ the validation sets of RSICD-CLS and UCM-CLS, which are the classification-oriented versions of RSICD [21] and UCM [22], respectively. To further examine the cross-domain generalization capability of our model, we incorporate three additional external benchmarks: WHU-RS19 [59], RSSCN7 [60], and AID [61]. The numerical and categorical statistics for these evaluation sets are summarized in Table 8.

Table 8: Statistics of zero-shot classification benchmarks in the remote sensing domain.

Dataset	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID
# of Images	1,094	2,100	1,005	2,800	10,000
# of Classes	31	21	19	7	30

Fashion. We aggregate Fashion200k [24], FashionGen [25], and Polyvore Outfits [26] to form the primary training set for the fashion domain. To handle instances where a single item is shown from multiple visual perspectives, we stochastically select one image view per item during each training epoch. Regarding textual data, we merge product titles with their corresponding descriptions for the FashionGen and Polyvore subsets, while Fashion200k relies on its original sentence-based captions. Statistical details of these datasets are provided in Table 9, with visual examples showcased in Figure 10.

Table 9: Statistics of image-text paired datasets in the fashion domain.

Dataset	Fashion200k	FashionGen	Polyvore Outfits
# of Pairs	61,759	60,147	71,967

Zero-shot performance is assessed using the validation sets of the aforementioned fashion datasets, as outlined in Table 10. For Fashion200k and FashionGen, we perform evaluations at both the super-class (coarse) and sub-class (fine-grained) levels. Category labels in the Polyvore dataset are standardized to match the super-class level of other benchmarks, ensuring a consistent evaluation scale across all fashion datasets.



Figure 10: Examples of image-text pairs from the Fashion datasets.

Table 10: Statistics of zero-shot classification benchmarks in the fashion domain.

	Fashion200k		FashionGen		Polyvore Outfits
	Super-class	Sub-class	Super-class	Sub-class	Class
# of Images	29,789		32,528		15,288
# of Classes	5	31	48	121	11

SciCap. We incorporate the SciCap dataset, which comprises scientific figures such as charts, graphs, and mathematical notations. For our experiments, we utilize the SciCap-No-Subfig-Img subset, which excludes subfigures to ensure data consistency. This subset contains 106,834 image-text pairs. Sample images and their corresponding captions from this domain are presented in Figure 11.

Simpsons. For the Simpsons dataset, we employ 679 image-text pairs from the simpsons-blip-captions collection. Due to the limited scale of the labeled data, we integrate an additional image pool from the Simpsons Characters Data as an unlabeled reservoir. Samples from this domain are showcased in Figure 11.

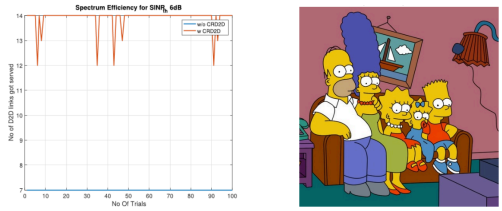


Figure 11: Examples of image-text pairs from the SciCap and Simpsons datasets.

C.2 Implementation Details

Setup. We adopt CLIP [1] as the pretrained vision-language backbone for all experiments. By default, ViT-B/16 is used as the vision encoder and the experiments with other vision encoders are in Appendix F.5. Unless otherwise noted, all experiments are performed on a single NVIDIA A100 GPU with a batch size of 128 for remote sensing and 256 for fashion datasets, where each mini-batch consists of an equal number of paired image-text samples and unpaired images. As ToMA is built on SemiCLIP [5], our training procedure follows the same two-stage paradigm and the model is trained for 25 epochs in stage 1 and 15 epochs in stage 2. We also follow SemiCLIP for the optimizer, cosine learning rate scheduler, learning rate, and the remaining training configurations. The top-k and top-P are set to 4 and 30, respectively, following SemiCLIP. The coefficient c in the loss function $\mathcal{L}^{(s)}$ is set to 0.5 for remote sensing and 0.1 for fashion datasets. The results of the hyperparameter search on the loss coefficients are shown in Appendix F.6. For the computation of PH, we follow the method proposed by ToMCLIP [13]. We conduct experiments with two settings: ToMA (ours),

which applies $\mathcal{L}_{\text{ToMA}}$ to each batch without considering domain information, and ToMA-domain (ours), which applies it separately to each domain. The H_0-death and H_1-birth edge sets are recomputed from the current embeddings at every mini-batch and treated as non-differentiable index selections. These indices determine which relations enter $\mathcal{L}_{\text{ToMA}}$, while the directional alignment loss is computed from the current embeddings, allowing gradients to propagate normally through the encoders.

Baselines. We compare ToMA against several competitive baselines. CLIP (original) refers to the pretrained model without further adaptation, and CLIP (fine-tuned) denotes the model fine-tuned using only labeled image-text pairs. We also consider CLIP (oracle), which is fine-tuned using both labeled and unlabeled image-text pairs with access to the ground-truth labels of the unlabeled data, serving as an upper-bound reference. In addition, we include Hard-PL [56], Soft-PL [57], and S-CLIP, all of which were introduced in S-CLIP [4]. SemiCLIP [5] is also adopted as a baseline, since ToMA is built on it. The scores of CLIP, Hard-PL, Soft-PL and S-CLIP are imported from [5].

The models are evaluated on zero-shot classification and image-text retrieval, where performance is measured by Top-1 classification accuracy (%) and Recall@K (R@K), respectively. Results are reported as the mean and standard deviation over three runs with fixed random seeds $\{0, 1, 2\}$ for reproducibility. For fair comparison, SemiCLIP [5] is retrained using the same set of random seeds.

D Preliminaries of Semi-Supervised Vision-Language Alignment

Semi-supervised vision-language alignment addresses the setting where paired image-text data $\mathcal{D}^l = \{(I_i, T_i)\}_{i=1}^N$ are scarce, while unlabeled images $\mathcal{D}^u = \{I_j\}_{j=1}^M$ are abundant with $M \gg N$. Its goal is to leverage unlabeled images to improve the alignment of the shared embedding space. SemiCLIP [5] approaches this problem with a two-stage framework.

Stage 1. Semantic Concepts Mining. During supervised pre-training, semantic concepts are extracted as nouns from the captions. Their CLIP text embeddings, obtained via $f_T(\cdot)$, initialize a linear classifier ψ, which is trained to predict image-level concepts using the soft cross-entropy loss $\mathcal{L}_{\text{SCM}} = -\sum_{i=1}^N \frac{y_{I_i}}{|y_{I_i}|} \log \psi(I_i^e)$ where $y_{I_i} \in \{0,1\}^V$ denotes the multi-hot concept label. The training objective is $\mathcal{L}_{\text{Semi}}^{(1)} = \mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{SCM}}$. The trained model then assigns pseudo-labels $\hat{y}_{I_j}$ to unlabeled images by selecting the top-k confident concepts [62].

Stage 2. Semi-supervised Fine-tuning. In the second stage, the model is fine-tuned to improve alignment between unlabeled images and pseudo-labels. This is achieved with two losses. The first, concept-level semantic consistency, trains the model to preserve the pseudo-labels of strongly augmented unlabeled images: $\mathcal{L}_{\text{SCM}}^u = -\sum_{j=1}^{|D^u|} \frac{\hat{y}_{I_j}}{|\hat{y}_{I_j}|} \log \psi(f_I(\Omega(I_j)))$. The second, caption-level trapezoidal consistency, constructs a surrogate caption $\hat{T}_j$ for each unlabeled image by concatenating its top-k concepts, and then enforces structural consistency over $\tilde{\mathcal{D}} = \mathcal{D}^l \cup \mathcal{D}^p = \{(\tilde{I}_i, \tilde{T}_i)\}_{i=1}^{N+M_P}$:

$$\mathcal{L}_{\text{Trap}} = \underbrace{\mathcal{L}_{\text{CLIP}}}_{\text{upper base}} + \frac{1}{|\tilde{D}|^2} \sum_{i=1}^{|\tilde{D}|} \sum_{j=1}^{|\tilde{D}|} \left(\underbrace{(\langle \tilde{I}_i^e, \tilde{T}_j^e \rangle - \langle \tilde{I}_j^e, \tilde{T}_i^e \rangle)^2}_{\text{diagonals}} + \underbrace{(\langle \tilde{I}_i^e, \tilde{I}_j^e \rangle - \langle \tilde{T}_j^e, \tilde{T}_i^e \rangle)^2}_{\text{legs}} \right),$$

where $\mathcal{D}^p \subset \mathcal{D}^u$. The overall objective is $\mathcal{L}_{\text{Semi}}^{(2)} = \mathcal{L}_{\text{SCM}}^u + \mathcal{L}_{\text{Trap}}$.

E Reported and Retrained SemiCLIP Results

As noted in the main paper, we retrain SemiCLIP under the same evaluation pipeline and random seeds as our method, and use these retrained results for the main comparison. Our intention is not to replace the originally reported SemiCLIP results in [5], but to provide a controlled baseline under the same experimental protocol. Because the reported results and our method come from different training runs, direct comparison to the originally reported numbers may reflect not only method differences but also run-level variation. We therefore report both the published and retrained SemiCLIP scores in this appendix for transparency.

Table 11: Comparison between the originally reported SemiCLIP results [5] and our retrained SemiCLIP results on zero-shot classification for remote sensing datasets. Parentheses denote performance gains over supervised CLIP, and green values indicate gains larger than one.

Method	Data	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID
CLIP (fine-tuned)	L	75.7±2.2	80.1±6.8	92.2±1.5	70.7±5.5	79.8±3.8
Reported SemiCLIP [5]	L=U	83.8±1.0 (+8.1)	85.6±3.4 (+5.5)	96.6±1.2 (+4.4)	75.3±1.9 (+4.6)	87.4±0.5 (+7.6)
Retrained SemiCLIP [5]	L=U	81.4±1.1 (+5.7)	84.3±3.2 (+4.2)	95.6±0.3 (+3.4)	76.2±2.7 (+5.5)	84.0±1.2 (+4.2)
Reported SemiCLIP [5]	L≠U	84.1±0.5 (+8.4)	85.4±2.8 (+5.3)	96.5±1.7 (+4.3)	72.9±3.5 (+2.2)	86.2±1.0 (+6.4)
Retrained SemiCLIP [5]	L≠U	80.1±1.4 (+4.4)	82.8±1.0 (+2.7)	95.1±0.1 (+2.9)	75.0±0.8 (+4.3)	82.5±1.9 (+2.7)

Table 12: Comparison between the originally reported SemiCLIP results [5] and our retrained SemiCLIP results on image-text retrieval (R@5) for remote sensing datasets.

		Image→Text R@5			Text→Image R@5		
Method	Data	RSICD	UCM	Sydney	RSICD	UCM	Sydney
CLIP (fine-tuned)	L	25.7±2.3	55.2±0.8	46.6±3.0	24.9±0.6	56.3±1.6	49.4±2.0
Reported SemiCLIP [5]	L=U	27.6±1.6	59.2±2.2	53.4±4.6	26.4±0.4	58.4±2.4	58.0±2.0
Retrained SemiCLIP [5]	L=U	27.9±0.7	57.3±2.6	52.3±4.3	25.8±0.2	59.4±1.2	58.6±2.4
Reported SemiCLIP [5]	L≠U	28.5±0.2	59.5±3.6	54.6±3.6	26.3±0.4	59.5±1.3	58.6±4.6
Retrained SemiCLIP [5]	L≠U	26.1±0.9	57.9±1.0	54.6±0.8	25.1±0.2	59.4±1.5	59.2±1.6

Table 13: Comparison between the originally reported SemiCLIP results [5] and our retrained SemiCLIP results on zero-shot classification for fashion datasets. Parentheses denote performance gains over supervised CLIP, and green values indicate gains larger than one.

	Fashion200k		FashionGen		Polyvore
Method	Super-class	Sub-class	Super-class	Sub-class	Class
CLIP (fine-tuned)	79.0±3.5	35.1±0.7	35.4±8.1	24.5±2.4	60.4±2.3
Reported SemiCLIP [5]	85.8±0.4 (+6.8)	44.7±1.0 (+9.6)	51.4±0.6 (+16.0)	47.6±1.7 (+23.1)	74.4±0.5 (+14.0)
Retrained SemiCLIP [5]	84.5±1.8 (+5.5)	43.7±1.2 (+8.6)	43.4±3.9 (+8.0)	43.0±2.6 (+18.5)	73.0±2.1 (+12.6)

Table 14: Comparison between the originally reported SemiCLIP results [5] and our retrained SemiCLIP results on image-text retrieval (R@5) for fashion datasets.

	Image→Text R@5			Text→Image R@5		
Method	Fashion200k	FashionGen	Polyvore	Fashion200k	FashionGen	Polyvore
CLIP (fine-tuned)	13.7±0.4	32.1±0.2	16.3±0.5	13.5±0.2	31.9±0.2	16.2±0.3
Reported SemiCLIP [5]	22.1±0.4	48.4±0.7	27.7±0.3	22.5±0.2	49.7±0.3	27.6±0.2
Retrained SemiCLIP [5]	23.3±0.5	48.2±0.8	28.2±0.1	23.4±0.2	49.5±0.5	28.6±0.1

Tables 11-14 compare the originally reported SemiCLIP results and our retrained results on the remote sensing and fashion benchmarks. Overall, the retrained model exhibits the same qualitative behavior as the reported one, while showing some differences in absolute scores across settings. On remote sensing zero-shot classification (Table 11), the retrained results are generally lower on RSICD-CLS, UCM-CLS, WHU-RS19, and AID, but slightly higher on RSSCN7 under both L=U and L≠U. On remote sensing retrieval (Table 12), the differences are smaller and more mixed, with several metrics remaining comparable and some text-to-image results on UCM and Sydney slightly improving after retraining.

A similar pattern appears on the fashion benchmarks. In zero-shot classification (Table 13), the retrained SemiCLIP results are consistently below the reported scores, although the overall relative pattern across datasets is preserved. In image-text retrieval (Table 14), however, the retrained model remains highly competitive and slightly exceeds the reported results on several metrics, including Fashion200k and Polyvore.

Taken together, these comparisons suggest that SemiCLIP is reasonably sensitive to the specific training run, especially for zero-shot classification. For this reason, we use the retrained SemiCLIP results in the main paper, as they provide a more directly matched and controlled baseline for evaluating ToMA, while the originally reported results are included here to make the comparison fully transparent.

F Additional Experimental Results

F.1 Image-text Retrieval Results on Fashion Datasets

Table 15: Image-text retrieval results on fashion datasets using the same setup as in Table 3.

	Image→Text R@5			Text→Image R@5		
Method	Fashion200k	FashionGen	Polyvore	Fashion200k	FashionGen	Polyvore
CLIP (original)	12.5	23.5	24.3	12.3	27.7	27.4
CLIP (fine-tuned)	13.7±0.4	32.1±0.2	16.3±0.5	13.5±0.2	31.9±0.2	16.2±0.3
CLIP (oracle)	34.9±0.7	63.4±0.1	48.5±0.8	34.9±0.9	35.3±0.9	49.3±0.5
Hard-PL [56]	12.6±0.7	26.8±0.7	14.9±0.8	14.5±0.8	30.2±0.4	17.3±0.9
Soft-PL [57]	15.2±0.8	35.0±0.3	18.8±0.3	15.9±0.7	37.4±0.7	21.9±0.2
S-CLIP [4]	16.9±0.4	37.9±0.4	22.6±0.4	17.4±0.3	40.9±0.4	24.2±0.5
SemiCLIP [5]	23.3±0.5	48.2±0.8	28.2±0.1	**23.4**±0.2	49.5±0.5	28.6±0.1
ToMA (ours)	23.5±0.2	48.6±0.3	28.5±0.3	23.2±0.3	49.5±0.1	28.5±0.3
ToMA-domain (ours)	**23.8**±0.4	**48.7**±0.3	**28.6**±0.1	23.3±0.4	**49.8**±0.3	**28.7**±0.2

Table 15 reports the R@5 results on the fashion datasets. Compared with SemiCLIP, ToMA-domain improves five of the six retrieval metrics, including all three image-to-text settings, while remaining within 0.1 percentage points on Fashion200k text-to-image retrieval. The aggregate improvements are nevertheless modest, particularly compared with the classification gains in Table 3. We therefore conduct a query-level diagnostic analysis in Appendix F.2 to examine how topology-aware alignment affects different retrieval queries and semantic neighborhood structure.

F.2 Diagnostic Analysis of Fashion Retrieval

Although the aggregate changes in exact-pair retrieval are modest, they may mask heterogeneous effects across queries and semantic neighborhood structure. We therefore conduct a controlled query-level analysis on FashionGen to better understand how topology-aware alignment changes retrieval behavior. We use 7,519 products from the validation set, selecting one image-caption pair per product. This controlled one-pair-per-product evaluation is used only for query-level diagnostics and is therefore not directly comparable to the standard FashionGen retrieval scores in Table 15.

Cross-modal topology alignment. We first examine whether ToMA improves the agreement between the image and text representation structures. We construct the H_0-death edge graphs in batches of 256 over five shuffled passes. For each query, neighbor Jaccard measures the overlap between its image- and text-graph neighbors, while direction agreement measures the absolute cosine agreement between corresponding edge directions in the two modalities.

Table 16: Cross-modal topology alignment on the controlled FashionGen validation set. Improvements are reported in percentage points (p.p.).

Metric	SemiCLIP	ToMA	Improvements
H_0 neighbor Jaccard	22.57%	24.45%	+1.88 p.p.
H_0 direction agreement	33.21%	42.77%	+9.56 p.p.

As shown in Table 16, ToMA improves both neighborhood overlap and edge-direction agreement for the H_0-derived relations. Thus, the modest aggregate exact-pair retrieval gains occur despite clear improvements in the measured cross-modal topology alignment.

Behavior across query difficulty. We define query difficulty using the bidirectional retrieval margin of SemiCLIP. For each product, we compute the matched-pair similarity minus the highest non-matching similarity separately for image-to-text and text-to-image retrieval, and average the two margins. We then divide queries into equally sized low- (hard), medium-, and high-margin (easy) thirds. A smaller margin indicates that the ground-truth pair is less clearly separated from competing candidates. Table 17 reports the results for the hardest and easiest thirds.

Table 17: Retrieval performance for the hardest and easiest FashionGen queries. Query difficulty is determined using the SemiCLIP positive-negative retrieval margin.

Difficulty	Direction	SemiCLIP (R@1)	ToMA (R@1)	SemiCLIP (R@5)	ToMA (R@5)
Hard	I→T	0.12	**3.95**	8.14	**19.94**
Hard	T→I	0.12	**3.55**	7.66	**20.10**
Easy	I→T	**77.45**	64.57	**99.36**	92.58
Easy	T→I	**78.33**	63.17	**99.32**	91.06

ToMA substantially improves retrieval for the low-margin queries. In both retrieval directions, it improves the ground-truth rank for approximately 61% of these queries. In contrast, performance decreases for some high-margin queries whose exact pairs are already well separated by SemiCLIP. This pattern suggests that ToMA can improve cases that are poorly resolved by the baseline while perturbing the fine-grained ordering of already well-separated pairs. These opposing effects partially cancel in aggregate Recall.

Semantic coherence of retrieved neighborhoods. Exact-pair Recall evaluates whether the single paired item is retrieved but does not measure whether the remaining retrieved items are semantically related to the query. We therefore additionally evaluate subcategory and coarse-category P@5, excluding the exact paired item. We also report semantic nDCG@5, which assigns graded relevance to the exact pair, items from the same subcategory, and items from the same coarse category. For semantic nDCG@5, relevance scores of 3, 2, 1, and 0 are assigned to the exact pair, an item from the same subcategory, an item from the same coarse category, and all remaining items, respectively.

Table 18: Changes from SemiCLIP to ToMA in exact-pair retrieval and semantic neighborhood coherence on the controlled FashionGen validation set.

Metric	I→T change	T→I change
Exact R@5	+0.24 p.p.	−0.82 p.p.
Subcategory P@5	+1.29 p.p.	+1.27 p.p.
Coarse-category P@5	+1.06 p.p.	+1.28 p.p.
Semantic nDCG@5	+0.86 p.p.	+0.39 p.p.

Table 18 shows that ToMA consistently improves the semantic coherence of the retrieved neighborhoods, even when the change in exact-pair Recall is small. Together with the query-difficulty analysis, this result indicates that topology-aware alignment is particularly helpful for ambiguous queries and neighborhood-level semantic organization.

Fashion retrieval requires precise instance-level ordering among many closely related products. A topology-aware structural objective can improve cross-modal relational agreement and semantic neighborhood coherence without necessarily preserving the exact ordering of pairs that are already well separated. Taken together, these results indicate that the effect of ToMA on fashion retrieval is task dependent: it improves cross-modal structural agreement and the retrieval of low-margin queries, while its effect on aggregate exact-pair Recall remains modest.

F.3 Additional Specialist-Domain Evaluation

We further evaluate ToMA on two additional specialist domains: science figures and comics. Specifically, we use the SciCap [63] and Simpsons [64] datasets. Each model is trained separately on the corresponding dataset and evaluated for image-text retrieval on its validation split. Table 19 reports the results.

Table 19: Image-text retrieval results on SciCap (scientific figures) and Simpsons (comics) datasets. We train each model separately on each dataset and evaluate it on the corresponding validation set. Bold denotes the best result among semi-supervised methods for each metric.

	SciCap				Simpsons			
	Image→Text		Text→Image		Image→Text		Text→Image	
Method	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
CLIP (original)	8.3	13.8	8.8	14.2	13.2	35.5	10.5	32.9
CLIP (fine-tuned)	11.0±0.6	19.0±0.8	11.1±0.9	18.1±0.0	19.3±2.0	48.2±6.0	14.5±0.0	45.6±2.0
CLIP (oracle)	14.8±0.9	25.5±1.4	14.8±0.5	25.7±0.3	28.3±3.7	52.3±5.6	22.3±3.1	55.8±4.1
Hard-PL [56]	11.2±0.2	18.9±0.2	12.2±0.1	20.1±0.1	16.7±2.7	42.1±5.3	15.4±4.2	43.9±3.3
Soft-PL [57]	11.5±0.2	19.4±0.3	12.2±0.2	20.5±0.2	18.4±1.3	40.4±4.0	15.8±1.3	43.4±3.5
S-CLIP [4]	11.6±0.5	20.6±0.5	12.6±0.4	21.1±0.4	15.4±2.0	43.9±5.5	17.5±1.5	47.4±6.8
SemiCLIP [5]	13.1±0.0	21.5±0.1	13.6±0.3	21.9±0.1	23.7±0.0	54.4±2.0	**19.7**±1.3	54.4±3.3
ToMA (ours)	**13.2**±0.0	**21.7**±0.1	**13.7**±0.2	**22.0**±0.2	**27.2**±4.0	**55.7**±1.5	18.9±3.3	**55.7**±1.5

SciCap. On SciCap, ToMA performs comparably to SemiCLIP across all retrieval metrics. It slightly improves image-to-text retrieval from 13.1 to 13.2 in R@1 and from 21.5 to 21.7 in R@5, and text-to-image retrieval from 13.6 to 13.7 in R@1 and from 21.9 to 22.0 in R@5. Although the gains are marginal, these results show that the topology-aware alignment signal does not degrade performance on the science-figure domain and remains stable beyond the main benchmarks.

Simpsons. On Simpsons, ToMA yields more visible improvements for image-to-text retrieval, improving over SemiCLIP by 3.5 percentage points in R@1 and 1.3 percentage points in R@5. For text-to-image retrieval, ToMA improves R@5 by 1.3 percentage points, while its R@1 is slightly lower than SemiCLIP. Overall, these results suggest that ToMA can provide additional benefits on comics-style data, particularly for image-to-text retrieval, while maintaining comparable performance in the reverse retrieval direction.

F.4 Seed-wise Experimental Results

Table 20: Seed-wise averaged results over each dataset group. ZS denotes zero-shot classification accuracy, and I2T/T2I denote image-to-text and text-to-image retrieval performance, respectively.

Dataset	Metric	Seed 0			Seed 1			Seed 2		
		SemiCLIP	ToMA	ToMA-domain	SemiCLIP	ToMA	ToMA-domain	SemiCLIP	ToMA	ToMA-domain
RS (L=U)	ZS	82.5	85.9	**87.7**	84.4	**84.6**	83.3	85.9	**87.8**	86.9
	I2T	45.1	45.4	**47.2**	46.2	45.7	**46.9**	46.2	**48.2**	47.8
	T2I	46.9	46.0	**51.9**	47.2	49.0	**51.1**	49.7	**49.8**	49.7
RS (L≠U)	ZS	83.6	**85.1**	84.8	82.1	83.7	**84.8**	83.5	86.9	**87.0**
	I2T	45.7	46.9	**47.3**	45.8	**47.6**	47.1	47.2	**48.7**	47.7
	T2I	46.8	48.2	**48.9**	47.8	48.1	**50.0**	49.0	47.9	**49.7**
Fashion	ZS	58.8	58.7	**61.2**	58.5	**60.6**	60.2	55.3	**57.8**	**57.8**
	I2T	33.3	**33.8**	33.7	33.2	33.4	**33.7**	33.3	33.4	**33.7**
	T2I	33.8	**33.9**	33.6	33.7	33.5	**34.0**	34.0	33.8	**34.1**

Table 20 reports seed-wise averaged results over each dataset group. This analysis is intended to verify whether the improvements of our method are consistent across random seeds rather than driven by a single favorable run. Across the three dataset settings, three metrics, and three random seeds, ToMA-domain matches or improves upon SemiCLIP in 25 out of 27 seed-level aggregate comparisons, with strict improvements in 24 cases. The gains are particularly consistent in the remote sensing settings, including the distribution-shift case. On fashion datasets, the improvements are more modest, especially for retrieval, but the results remain comparable or slightly better in most seed-level comparisons. These results support the stability of the proposed topology-aware alignment signal.

F.5 Different Neural Network Architectures

Our main experiments use ViT-B/16 as the default vision encoder. To examine whether the effect of ToMA depends on this particular backbone choice, we additionally evaluate three different neural architectures on the remote sensing benchmarks: ResNet-50, ViT-B/32, and ViT-B/16. For each backbone, we follow the same semi-supervised training protocol as in Section 4.2 and report zero-shot classification and image-text retrieval results in Tables 21 and 22, respectively.

Table 21: Zero-shot classification results on remote sensing datasets using different neural architectures. Parentheses denote performance gains over supervised CLIP, and green values indicate gains larger than one. Bold denotes the best semi-supervised result in each setting.

Model	Method	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID
ResNet-50	CLIP (original)	45.3	50.5	65.5	58.9	47.8
	CLIP (fine-tuned)	58.3±0.3	63.5±3.4	76.5±3.2	61.9±1.2	63.1±1.3
	S-CLIP	66.9±1.7 (+8.6)	66.7±1.6 (+3.2)	86.9±2.0 (+10.4)	66.2±1.1 (+4.3)	73.0±0.3 (+9.9)
	SemiCLIP	69.2±1.8 (+10.9)	73.9±0.9 (+10.4)	83.8±1.9 (+7.3)	**68.2**±1.3 (+6.4)	75.4±2.1 (+12.3)
	ToMA (ours)	**72.0**±2.1 (+13.7)	**79.5**±0.8 (+16.0)	**89.7**±1.8 (+13.2)	67.5±0.9 (+5.6)	**78.2**±2.4 (+15.1)
ViT-B/32	CLIP (original)	56.2	58.5	76.3	62.3	55.6
	CLIP (fine-tuned)	73.6±3.1	79.0±4.4	91.1±1.3	70.4±3.1	79.6±2.5
	S-CLIP	78.5±0.5 (+4.9)	77.4±4.4 (-1.6)	94.9±0.5 (+3.8)	71.1±2.6 (+0.7)	84.2±1.6 (+4.6)
	SemiCLIP	80.7±1.4 (+7.1)	83.8±3.3 (+4.8)	94.6±0.8 (+3.5)	**75.0**±1.8 (+4.6)	83.0±2.0 (+3.4)
	ToMA (ours)	**82.0**±1.1 (+8.4)	**87.1**±3.3 (+8.1)	**97.0**±0.4 (+6.0)	72.1±1.7 (+1.7)	**84.9**±1.4 (+5.3)
ViT-B/16	CLIP (original)	59.2	60.2	81.2	69.0	59.6
	CLIP (fine-tuned)	75.7±2.2	80.1±6.8	92.2±1.5	70.7±5.5	79.8±3.8
	S-CLIP	81.4±1.8 (+5.7)	81.3±3.4 (+1.2)	95.9±1.8 (+3.7)	75.1±2.0 (+4.4)	86.4±1.7 (+6.6)
	SemiCLIP	81.4±1.1 (+5.7)	84.3±3.2 (+4.2)	95.6±0.3 (+3.4)	**76.2**±2.7 (+5.5)	84.0±1.2 (+4.2)
	ToMA (ours)	**84.8**±1.3 (+9.1)	**86.1**±3.4 (+6.0)	**97.9**±0.4 (+5.7)	74.2±1.7 (+3.5)	**87.3**±2.2 (+7.5)

Table 22: Image-text retrieval results on remote sensing datasets using different neural architectures. Bold denotes the best semi-supervised result in each setting.

		Image→Text R@5			Text→Image R@5		
Model	Method	RSICD	UCM	Sydney	RSICD	UCM	Sydney
ResNet-50	CLIP (original)	9.4	34.3	36.2	10.1	24.8	51.7
	CLIP (fine-tuned)	15.4±1.7	41.3±1.8	47.1±6.5	15.1±1.0	40.9±1.6	56.1±2.4
	S-CLIP	18.4±0.6	45.7±1.4	50.0±3.0	16.8±1.2	43.5±1.5	55.1±2.0
	SemiCLIP	20.0±0.9	45.2±3.1	**49.4**±3.2	17.4±0.5	45.4±3.5	54.6±3.5
	ToMA (ours)	**21.1**±1.1	**47.5**±1.1	47.1±2.1	**18.6**±0.5	**47.1**±1.0	**56.9**±4.2
ViT-B/32	CLIP (original)	12.1	37.6	41.4	12.9	35.2	37.9
	CLIP (fine-tuned)	24.5±0.6	56.3±1.5	47.1±5.0	24.6±1.0	57.0±1.1	50.0±4.6
	S-CLIP	25.0±0.8	55.7±2.5	43.1±3.0	23.9±0.5	59.2±1.8	51.7±1.7
	SemiCLIP	26.3±1.6	**60.5**±2.3	**52.3**±3.5	24.0±1.4	60.5±1.4	55.8±2.1
	ToMA (ours)	**27.8**±0.7	60.2±1.5	50.6±3.2	**24.8**±0.5	**61.3**±0.8	**56.3**±2.9
ViT-B/16	CLIP (original)	12.6	46.7	44.8	13.9	39.5	44.8
	CLIP (fine-tuned)	25.7±2.3	55.2±0.8	46.6±3.0	24.9±0.6	56.3±1.6	49.4±2.0
	S-CLIP	27.5±1.1	57.0±1.5	51.1±4.3	25.6±0.6	57.3±4.1	50.6±2.6
	SemiCLIP	27.9±0.7	57.3±2.6	**52.3**±4.3	25.8±0.2	59.4±1.2	**58.6**±2.4
	ToMA (ours)	**29.0**±1.5	**61.4**±4.6	48.9±2.1	**27.6**±1.2	**60.8**±3.0	56.3±3.5

Overall, the results show that ToMA generalizes well across different backbone architectures. In zero-shot classification (Table 21), ToMA achieves the best semi-supervised performance in most settings and consistently improves over supervised CLIP by clear margins. These gains are particularly pronounced on RSICD-CLS, UCM-CLS, WHU-RS19, and AID, indicating that the proposed topology-aware alignment is not tied to a specific encoder family. While the improvement on RSSCN7 is relatively smaller and the ViT-B/16 backbone remains slightly below SemiCLIP on this dataset, the overall trend remains consistently favorable to ToMA.

A similar pattern is observed for image-text retrieval (Table 22). Across different backbones, ToMA improves over SemiCLIP on most retrieval settings, with especially consistent gains on RSICD and UCM for both image-to-text and text-to-image retrieval. The results on Sydney are comparatively mixed, suggesting that the benefit of topology-aware alignment may vary depending on the dataset

characteristics and retrieval direction. Nevertheless, the overall results confirm that ToMA provides a complementary structural signal beyond the baseline semi-supervised objective.

These observations suggest that the effectiveness of ToMA does not rely on the default ViT-B/16 configuration. Instead, the proposed topology-aware regularization transfers across both convolutional and transformer-based encoders, supporting its usefulness as a backbone-agnostic approach for semi-supervised vision-language adaptation.

F.6 Loss Coefficient and Homology-Dimension Ablations

We conduct ablation studies to analyze the role of the topology-aware alignment in

$$\mathcal{L}^{(s)} = \mathcal{L}^{(s)}_{\text{Semi}} + \frac{c}{2}\left(\mathcal{L}^{\text{0-death}}_{\text{ToMA}} + \mathcal{L}^{\text{1-birth}}_{\text{ToMA}}\right). \tag{6}$$

Specifically, we vary the coefficient $c \in \{0.1, 0.5, 1, 2\}$ and compare three variants: using only the H_0-death term (+ $(c/2)\mathcal{L}^{\text{0-death}}_{\text{ToMA}}$), using only the H_1-birth term (+ $(c/2)\mathcal{L}^{\text{1-birth}}_{\text{ToMA}}$), and using both terms together (+ $c/2(\mathcal{L}^{\text{0-death}}_{\text{ToMA}} + \mathcal{L}^{\text{1-birth}}_{\text{ToMA}})$). This experiment is designed to examine two questions: how sensitive ToMA is to the weighting coefficient c, and whether the 1-dimensional topological signal provides useful information beyond the 0-dimensional one. Notably, our formulation uses only the birth edges of H_1 under the 1-skeleton filtration, which allows us to incorporate 1-dimensional information in a lightweight form rather than relying on full higher-dimensional simplicial computations.

Table 23: Ablation on the loss coefficient c and homology components on remote sensing dataset. We compare ToMA with only H_0-death, only H_1-birth, and both jointly under different values of c.

	Zero-Shot				Image→Text				Text→Image			
	$c=0.1$	$c=0.5$	$c=1$	$c=2$	$c=0.1$	$c=0.5$	$c=1$	$c=2$	$c=0.1$	$c=0.5$	$c=1$	$c=2$
CLIP (fine-tuned)		79.7				42.5				43.5		
SemiCLIP [5]		84.3				45.8				47.9		
+ $(c/2)\mathcal{L}^{\text{0-death}}_{\text{ToMA}}$	84.7	85.0	85.5	85.8	46.7	47.8	48.7	46.5	49.7	49.9	49.4	47.7
+ $(c/2)\mathcal{L}^{\text{1-birth}}_{\text{ToMA}}$	85.5	86.3	85.8	85.7	46.0	47.3	47.8	45.6	48.8	49.1	50.0	48.7
+ $c/2(\mathcal{L}^{\text{0-death}}_{\text{ToMA}} + \mathcal{L}^{\text{1-birth}}_{\text{ToMA}})$	84.6	86.0	85.9	85.4	46.8	46.2	46.9	46.1	49.6	48.7	50.1	47.9

As shown in Table 23, ToMA is reasonably robust to the choice of the loss coefficient c, especially in the range $c \in \{0.1, 0.5, 1\}$. Across all three variants, adding topology-aware alignment consistently improves the zero-shot average over SemiCLIP, confirming that both H_0- and H_1-based signals provide useful structural supervision. Among the single-component variants, the H_1-birth-only term yields the strongest zero-shot result, reaching 86.3 at $c = 0.5$, which suggests that even the lightweight 1-dimensional signal carries informative structure beyond the connectivity information captured by H_0. For retrieval, the pattern is slightly different: the H_0-death-only variant gives the best image-to-text score (48.7 at $c = 1$), while the joint variant achieves the best text-to-image score (50.1 at $c = 1$). These results indicate that the two homology components play complementary roles, with H_1 being particularly helpful for classification and H_0 remaining beneficial for retrieval-oriented local structure.

At the same time, the results also show that excessively large topology weights can be harmful. In particular, when $c = 2$, performance tends to drop for retrieval across several variants, suggesting that overly strong topological regularization may begin to interfere with the original semi-supervised alignment objective. This trend supports the use of a moderate coefficient in practice.

Table 24: Results for different weights for $\mathcal{L}^{\text{0-death}}_{\text{ToMA}}$ and $\mathcal{L}^{\text{1-birth}}_{\text{ToMA}}$.

	Zero-Shot				Image→Text				Text→Image			
c_2 \ c_1	0.1	0.5	1	2	0.1	0.5	1	2	0.1	0.5	1	2
0.5	84.9	85.7	85.8	85.5	46.5	45.5	46.4	47.3	49.6	49.5	49.7	48.3
1	84.6	86.0	85.9	85.4	46.8	46.2	46.9	46.1	49.6	48.7	50.1	47.9
2	84.7	85.8	85.4	84.0	47.8	45.6	45.7	44.3	49.2	49.1	47.3	46.4

Table 24 further analyzes the relative contribution of the two topological terms. In this experiment, we use

$$\mathcal{L}^{(s)} = \mathcal{L}^{(s)}_{\text{Semi}} + \frac{c_1}{2}\left(\mathcal{L}^{\text{0-death}}_{\text{ToMA}} + c_2\mathcal{L}^{\text{1-birth}}_{\text{ToMA}}\right).$$

Here, c_1 controls the overall magnitude of the topology-aware regularizer, while c_2 reweights the H_1-birth term relative to the H_0-death term. Overall, the results suggest that performance is more sensitive to the global scale c_1 than to the relative weighting c_2. Moderate values of c_1 yield the most stable results, whereas larger values often degrade retrieval performance. The best zero-shot classification result is obtained at $(c_1, c_2) = (0.5, 1)$, and the best text-to-image retrieval result at $(1, 1)$, indicating that comparable weighting of the two components already works well in practice. Although image-to-text retrieval shows a slight preference for a larger H_1 weight, no single asymmetric configuration consistently outperforms the balanced setting across all metrics. Taken together, these results suggest that the H_1-birth term contributes useful complementary information, but that the benefit of ToMA mainly comes from combining both terms under a moderate overall regularization strength.

F.7 Results for R@1

We additionally report image-text retrieval results in terms of Recall@1 (R@1), which provides a stricter evaluation than Recall@5 (R@5). Tables 25 and 26 summarize the results on remote sensing and fashion datasets, respectively.

Table 25: Image-text retrieval results of R@1 on remote sensing datasets using the same setup as in Table 1. Bold denotes the best results within the same setups.

		Image→Text R@1			Text→Image R@1		
Method	Data	RSICD	UCM	Sydney	RSICD	UCM	Sydney
CLIP (original)	-	4.5	17.1	10.3	3.6	9.0	12.1
CLIP (fine-tuned)	L	10.6±1.2	18.9±2.7	20.7±0.0	7.3±0.7	14.9±0.5	12.6±1.0
Hard-PL [56]	L=U	11.4±1.7	17.0±2.9	17.9±2.8	6.8±0.5	14.1±2.3	**20.7**±3.0
Soft-PL [57]		11.6±1.2	19.2±1.1	20.7±3.0	7.1±0.4	13.7±0.6	20.3±5.6
S-CLIP [4]		11.4±0.3	19.5±1.3	17.8±2.6	7.1±0.8	14.0±1.4	16.7±1.0
SemiCLIP [5]		11.0±0.7	19.4±0.6	18.4±4.5	6.8±0.5	15.2±0.8	18.4±0.8
ToMA (ours)		**12.4**±1.0	19.4±2.8	20.7±1.4	**8.1**±0.7	**17.1**±0.8	20.1±1.6
ToMA-domain (ours)		12.3±0.8	**20.0**±2.8	**21.3**±3.5	7.8±0.2	16.2±0.7	**20.7**±2.8
Hard-PL [56]	L≠U	11.1±1.5	17.1±2.4	19.0±3.4	6.1±0.2	16.0±1.5	16.7±5.3
Soft-PL [57]		11.2±0.4	16.0±1.5	16.7±3.6	6.5±0.2	15.1±2.2	18.4±1.0
S-CLIP [4]		10.9±0.6	18.9±1.5	**20.4**±2.2	7.2±0.6	**16.5**±1.2	14.9±2.6
SemiCLIP [5]		10.5±0.7	20.6±2.0	20.1±0.8	7.2±0.7	15.7±0.8	20.7±0.0
ToMA (ours)		13.2±0.9	**21.1**±2.1	18.4±3.5	8.0±0.6	15.4±1.8	21.8±2.9
ToMA-domain (ours)		**13.3**±0.5	19.7±1.6	19.5±1.6	**8.1**±0.2	14.8±2.2	**24.7**±3.2

Table 26: Image-text retrieval results of R@1 on fashion datasets using the same setup as in Table 3. Bolds denotes the best results within the same setups.

	Image→Text R@1			Text→Image R@1		
Method	Fashion200k	FashionGen	Polyvore	Fashion200k	FashionGen	Polyvore
CLIP (original)	4.7	10.9	11.4	4.5	12.5	13.0
CLIP (fine-tuned)	4.8±0.3	12.9±0.4	6.4±0.2	4.2±0.0	13.0±0.0	6.2±0.1
Hard-PL [56]	4.8±0.2	11.1±0.2	6.5±0.3	5.1±0.5	13.7±0.4	7.8±0.3
Soft-PL [57]	5.6±0.3	14.7±0.4	8.2±0.2	5.3±0.4	15.4±0.5	9.8±0.0
S-CLIP [4]	6.6±0.3	16.7±0.4	10.4±0.2	6.1±0.2	18.9±0.2	10.9±0.4
SemiCLIP [5]	**10.1**±0.2	23.8±0.2	14.0±0.1	**9.1**±0.1	24.8±0.2	13.6±0.2
ToMA (ours)	10.0±0.2	24.2±0.2	**14.2**±0.3	9.0±0.1	25.0±0.1	13.8±0.1
ToMA-domain (ours)	9.9±0.2	**24.5**±0.1	14.1±0.1	8.9±0.1	**25.1**±0.3	**13.8**±0.1

Overall, the R@1 results are consistent with the R@5 results. On remote sensing datasets, ToMA and ToMA-domain generally improve over SemiCLIP, with particularly clear gains on RSICD under both L=U and L≠U settings. The two variants also achieve the best results in several settings on UCM and Sydney, showing that topology-aware alignment improves not only coarse retrieval performance but also top-ranked retrieval precision.

On fashion datasets, the improvements are more selective. ToMA and ToMA-domain outperform SemiCLIP on FashionGen and Polyvore in several retrieval settings, while SemiCLIP remains slightly stronger on Fashion200k. Overall, these R@1 results support the same conclusion as the R@5 evaluation: ToMA provides a complementary structural signal that is particularly effective on remote sensing datasets and on more heterogeneous fashion datasets.

F.8 Comparison with Topology-Free Graph Baselines

To separate the effect of introducing an additional relational alignment objective from the effect of topology-informed edge construction, we compare ToMA with two topology-free graph baselines: a random graph and a k-nearest-neighbor (k-NN) graph. The random baseline connects randomly selected pairs, whereas the k-NN baseline constructs neighborhoods according to the current representation geometry. To control for the number of aligned relations, we approximately match the number of graph edges to that of the H_0-death edges: for a batch of $N = 128$ samples, the H_0 minimum spanning tree contains $N - 1 = 127$ edges, while the random baseline samples 128 edges and the k-NN baseline uses $k = 1$, yielding one nearest-neighbor relation per sample. These baselines replace the homology-based edge construction of ToMA while retaining the same encoders, directional alignment objective, training protocol, and evaluation settings. We evaluate both the standard and domain-partitioned variants, denoted by the suffix "-d."

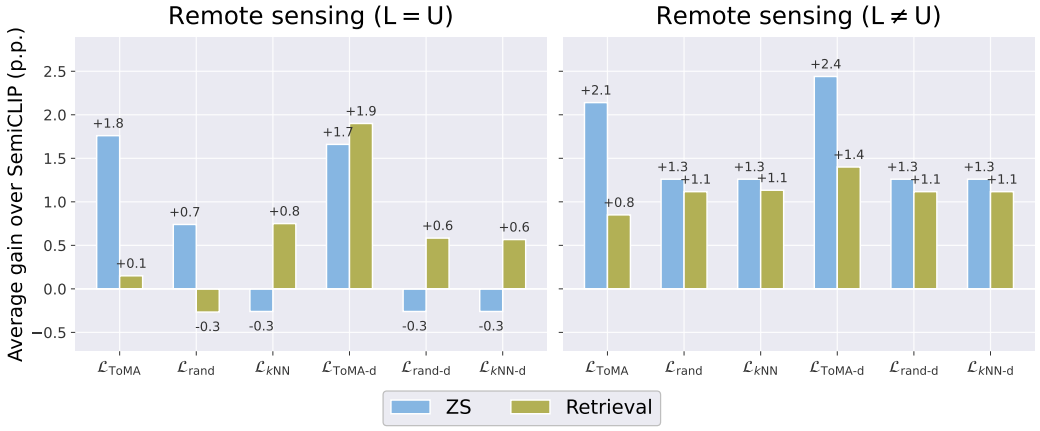


Figure 12: Comparison with topology-free graph baselines. Bars denote average gains over SemiCLIP for zero-shot classification (ZS) and image-text retrieval under L=U and L≠U. Random and k-NN replace the homology-based edge construction with randomly sampled and nearest-neighbor graphs, respectively. The suffix "-d" denotes the corresponding domain-partitioned variant.

Figure 12 reports the average gains over SemiCLIP for zero-shot classification and image-text retrieval. When the labeled and unlabeled data are drawn from the same domains (L=U), ToMA improves zero-shot classification by 1.8 percentage points, compared with 0.7 for the random graph and -0.3 for the k-NN graph. The domain-partitioned ToMA variant achieves a gain of 1.7 percentage points, whereas both domain-partitioned graph baselines yield a decrease of 0.3 percentage points. For retrieval, the standard k-NN baseline achieves a larger average gain than standard ToMA ($+0.8$ versus $+0.1$ percentage points). In contrast, ToMA-d provides the largest retrieval improvement, reaching $+1.9$ percentage points, compared with approximately $+0.6$ for the domain-partitioned random and k-NN baselines. For the L=U zero-shot comparison, we additionally perform a paired analysis between ToMA and random edge alignment over the three matched seeds. The macro-averaged ToMA−random difference across the five evaluation datasets is $+1.11$ percentage points (95% CI $[-2.01, 4.24]$, paired t-test $p = 0.265$). Thus, the observed advantage is positive for all three matched seeds but is not statistically significant under the current three-seed evaluation.

Under the more challenging L≠U setting, ToMA and ToMA-d improve zero-shot classification by 2.1 and 2.4 percentage points, respectively, exceeding the 1.3 point gains obtained by the corresponding random and k-NN baselines. Although their underlying scores differ slightly, the random and k-NN controls yield nearly identical aggregate gains at the reported precision. For retrieval, the topology-free baselines remain competitive with standard ToMA: the random and k-NN graphs improve the average score by approximately 1.1 percentage points, whereas ToMA yields a gain of 0.8 percentage points. Nevertheless, ToMA-d again obtains the strongest domain-partitioned result, with an improvement of 1.4 percentage points.

Overall, the results show that generic graph regularization can itself benefit representation alignment, particularly for retrieval, and that the topology-aware construction is not uniformly superior under every evaluation setting. However, ToMA provides consistently stronger zero-shot classification performance, while its domain-partitioned variant achieves the largest retrieval gains in both domain settings. These findings suggest that the gains of ToMA are not attributable solely to the use of an arbitrary graph regularizer, while domain-wise topology construction can provide additional benefits for retrieval.

Together with the homology-dimension analysis in Figure 3b, these results further suggest that the benefit of ToMA does not simply arise from aligning a larger number of pairwise relations. Under the 1-skeleton filtration, the H_1-birth term already contains the dense set of all non-MST edges, whereas the H_0-death term contains only the $N-1$ MST edges; nevertheless, the H_0-only domain variant outperforms the H_1-only variant, and their combination achieves the strongest performance. Since the two terms are normalized separately before being combined, this pattern highlights the importance of topology-induced edge decomposition and the separate treatment of connectivity and cycle-closing relations, rather than merely generic graph regularization or dense pairwise alignment.

F.9 Results for Smaller Batch Size

We additionally evaluate ToMA with a smaller batch size of 64 on the remote sensing benchmarks. Since topology-aware alignment is computed within each mini-batch, this experiment tests whether the benefit of ToMA remains effective even when fewer samples are available to estimate the batch-level structure.

Table 27: Zero-shot classification results on remote sensing datasets for batch size 64. Parentheses denote performance gains over supervised CLIP, and green values indicate gains larger than one. Bold denotes the best semi-supervised result in each setting.

Method	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID
CLIP (original)	59.2	60.2	81.2	69.0	59.6
CLIP (fine-tuned)	75.3±2.7	77.2±7.6	92.3±2.9	72.4±3.2	79.7±3.5
Hard-PL	78.5±3.3 (+3.2)	70.1±4.7 (-7.1)	91.6±0.4 (-0.7)	62.4±4.8 (-10.0)	82.2±2.9 (+2.5)
Soft-PL	77.8±3.4 (+2.5)	71.2±3.7 (-6.0)	94.3±1.0 (+2.0)	64.9±5.5 (-7.5)	82.6±1.6 (+2.9)
S-CLIP	74.2±1.6 (-1.1)	69.7±3.0 (-7.5)	93.3±1.5 (+1.0)	71.3±1.7 (-1.1)	79.8±1.6 (+0.1)
SemiCLIP	83.8±0.4 (+8.5)	**85.5**±5.0 (+8.3)	94.9±0.8 (+2.6)	71.6±5.0 (-0.8)	87.0±1.1 (+7.3)
ToMA (ours)	**85.3**±2.0 (+10.0)	84.3±5.0 (+7.1)	**97.5**±0.6 (+5.2)	**75.7**±2.9 (+3.3)	**88.5**±3.2 (+8.8)

As shown in Table 27, ToMA consistently achieves the best semi-supervised performance across four datasets. Compared with SemiCLIP, it improves zero-shot classification on RSICD-CLS, WHU-RS19, RSSCN7, and AID, while remaining only slightly below SemiCLIP on UCM-CLS. In particular, ToMA reaches 85.3 on RSICD-CLS, 97.5 on WHU-RS19, 75.7 on RSSCN7, and 88.5 on AID, showing clear gains over both supervised CLIP and the semi-supervised baselines.

These results show that ToMA remains effective with a batch size of 64, indicating that it does not require the default batch size of 128 to provide a useful structural signal. We next examine a broader range of batch sizes to characterize the batch-size sensitivity of ToMA more systematically.

F.10 Batch-Size Sensitivity

To further characterize the effect of batch size, we evaluate SemiCLIP and ToMA using batch sizes of 8, 16, 32, 64, 128, and 256. Figure 13 reports macro-averaged performance on the remote sensing

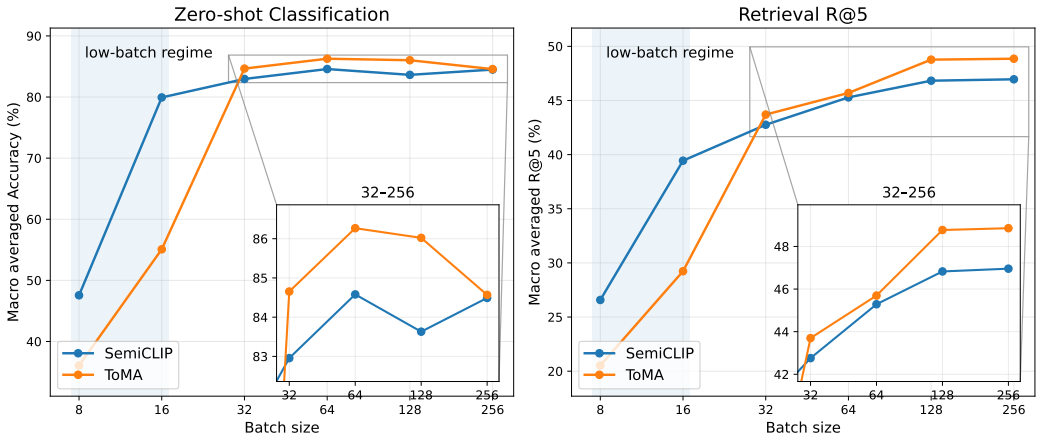


Figure 13: Effect of batch size on remote sensing datasets (L=U). We report macro-averaged zero-shot classification accuracy across five benchmarks and macro-averaged retrieval R@5 across three datasets and both retrieval directions. The main panels show the full range of batch sizes from 8 to 256, and the shaded regions indicate the low-batch regime. The insets enlarge the range from 32 to 256. Both methods degrade at very small batch sizes, whereas ToMA matches or outperforms SemiCLIP consistently once the batch size reaches 32.

datasets (L=U). For zero-shot classification, we average the accuracy across the five evaluation datasets. For retrieval, we average R@5 across the three datasets and both retrieval directions. The main panels show the full batch-size range, while the insets provide an enlarged view of the results between 32 and 256.

Both methods deteriorate when the batch size becomes very small, indicating that the semi-supervised alignment objectives generally benefit from a sufficient number of samples within each mini-batch. The degradation is more pronounced for ToMA at batch sizes of 8 and 16. Since ToMA estimates topologically salient relationships within each mini-batch, an extremely small batch provides only a limited sample of the representation manifold. The resulting connectivity and cycle structure can therefore become more sensitive to the particular composition of the batch, reducing the reliability of the topological alignment signal.

Once the batch size reaches 32, ToMA recovers sharply and matches or outperforms SemiCLIP across all evaluated batch sizes. Between batch sizes 32 and 256, ToMA improves the macro-averaged classification accuracy by approximately 0.1–2.4 percentage points and retrieval R@5 by approximately 0.4–1.9 percentage points. The classification performance remains broadly stable throughout this range, while retrieval performance continues to improve up to batch sizes of 128 and 256. These results indicate that ToMA does not require a very large batch, but it benefits from having enough samples to estimate meaningful batch-level topology. Notably, this transition at a batch size of 32 is consistent with the batch-size sensitivity reported by the TopKD paper [10], who found that topology-informed knowledge distillation became consistently beneficial once the batch size exceeded 16, as a sufficient number of points was required to characterize the underlying manifold topology reliably.

F.11 Stability of MST Edges under Batch Perturbations

Because ToMA extracts topologically salient edges within each mini-batch, we examine whether these edges capture reproducible representation structure or are dominated by the particular composition of a batch. We conduct a batch perturbation analysis on RS-ALL using a batch size of 128. For each original batch, we randomly replace 13 samples, corresponding to approximately 10% of the batch, while keeping the remaining image-text pairs fixed. We then recompute the image and text MSTs for the perturbed batch. The reported results are averaged over 500 batch perturbations for each of three training seeds.

Edges incident to a replaced sample necessarily disappear from the perturbed batch and therefore cannot be used to assess structural stability. We thus measure the conditional retention rate only over

original MST edges whose two endpoints remain in the perturbed batch. Let E and $\widetilde{E}$ denote the MST edge sets obtained from the original and perturbed batches, respectively, and let E_{surv} denote the edges whose end points (samples) are retained after perturbation. We define edge retention as

$$\mathrm{Ret}(E, \widetilde{E}) = \frac{\left|E \cap \widetilde{E} \cap E_{\mathrm{surv}}\right|}{|E \cap E_{\mathrm{surv}}|}. \tag{7}$$

We compute this quantity separately for the image MST, the text MST, and the edges shared by both modalities.

Table 28: Conditional retention rates of MST edges under batch perturbations on RS-ALL. For each batch of 128 samples, 13 samples are randomly replaced, and edge retention is evaluated only among original edges whose two endpoints remain in the perturbed batch. Results are averaged over 500 perturbations for each of three training seeds.

Method	Image MST	Text MST	Cross-modal common edges
CLIP	91.31%	91.15%	89.70%
SemiCLIP	91.42%	91.36%	88.78%
ToMA	91.53%	91.40%	88.74%
ToMA-domain	91.47%	91.42%	88.77%

As shown in Table 28, the unimodal MSTs exhibit high retention across all methods. Image and text MST retention remains approximately 91%, while the retention of cross-modal common edges remains approximately 89%. In particular, SemiCLIP and ToMA achieve nearly identical retention rates for cross-modal common edges, with 88.78% and 88.74%, respectively. These results indicate that ToMA does not artificially increase the generic stability of an MST relative to the underlying baseline.

The main effect of ToMA instead lies in the amount of cross-modal structure shared before perturbation. ToMA increases the average number of image-text common MST edges from 40.32 for SemiCLIP to 43.50, while preserving essentially the same retention rate under batch perturbation. ToMA produces a larger set of image-text connectivity relations that remains reproducible under nearby perturbations of the mini-batch.

F.12 Domain Partitioning as a Control

To disentangle the effect of domain partitioning from that of the ToMA objective, we introduce a domain-partitioned SemiCLIP control. ToMA-domain assumes that the source-domain identity of each unlabeled sample is available as metadata. This setting is applicable when samples are collected from known datasets, acquisition sites, sensors, or data processing pipelines; standard ToMA does not require such information.

In our original implementation, both ToMA and ToMA-domain use the same domain-agnostic SemiCLIP objective. The only difference lies in the construction of the ToMA loss. ToMA identifies and aligns topological relations over the full mini-batch, whereas ToMA-domain partitions the mini-batch by source domain before constructing and applying the ToMA loss. Therefore, the comparison between ToMA and ToMA-domain isolates the effect of domain-wise topology construction.

We additionally evaluate **SemiCLIP-domain**, which applies the consistency losses of SemiCLIP separately within each source-domain subset but does not use the ToMA objective. The remaining training components, including supervised contrastive learning and pseudo-label generation, remain unchanged.

As shown in Table 29, SemiCLIP-domain improves the average accuracy only marginally, from 84.30 to 84.48, and its effects vary across datasets. Likewise, ToMA-domain does not improve the aggregate performance over standard ToMA, achieving 85.96 compared with 86.06.

In contrast, both ToMA variants substantially outperform their corresponding SemiCLIP baselines on average. ToMA improves over SemiCLIP by 1.76 percentage points, while ToMA-domain improves over SemiCLIP-domain by 1.48 percentage points. These results indicate that the improvement of ToMA-domain cannot be explained by domain partitioning alone.

Table 29: Zero-shot classification results for the domain-partitioned controls. The average is the macro-average across the five evaluation datasets.

Method	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID	Average
SemiCLIP	81.4	84.3	95.6	**76.2**	84.0	84.30
SemiCLIP-domain	81.9	85.3	95.6	75.5	84.1	84.48
ToMA	**84.8**	86.1	**97.9**	74.2	**87.3**	**86.06**
ToMA-domain	83.8	**87.5**	97.8	73.9	86.8	85.96

We therefore regard domain-wise topology construction as an optional, dataset-dependent extension. It can redistribute performance across individual datasets, such as improving UCM-CLS, but does not provide a consistent aggregate advantage. Overall, the results are more consistent with the ToMA objective, rather than domain separation itself, accounting for the principal improvement.

F.13 Different Labeled Image-Text Pair Ratios

We further evaluate ToMA under different labeled image-text pair ratios on the remote sensing datasets. In addition to the default 10% setting, we consider 20% and 30% labeled pairs, while treating the remaining samples as unlabeled. Table 30 reports the zero-shot classification results.

Table 30: Zero-shot classification results on remote sensing datasets using different image-text pair ratios. Parentheses indicate the performance gap from the supervised CLIP, and bold denotes the best results within each pair ratio.

Method	Ratio	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID
CLIP (original)	0%	59.2	60.2	81.2	69.0	59.6
CLIP (fine-tuned)	10%	75.7±2.2	80.1±6.8	92.2±1.5	70.7±5.5	79.8±3.8
SemiCLIP	10%	81.4±1.1 (+5.7)	84.3±3.2 (+4.2)	95.6±0.3 (+3.4)	**76.2**±2.7 (+5.5)	84.0±1.2 (+4.2)
ToMA (ours)	10%	**84.8**±1.3 (+9.1)	**86.1**±3.4 (+6.0)	**97.9**±0.4 (+5.7)	74.2±1.7 (+3.5)	**87.3**±2.2 (+7.5)
CLIP (fine-tuned)	20%	78.8±1.4 (+3.1)	80.8±1.6 (+0.7)	91.9±0.4 (-0.3)	74.4±3.9 (+3.7)	82.7±0.5 (+2.9)
SemiCLIP	20%	85.5±1.8 (+9.8)	**89.1**±0.4 (+9.0)	95.5±1.1 (+3.3)	67.4±2.6 (-3.3)	88.6±1.4 (+8.8)
ToMA (ours)	20%	**87.0**±0.9 (+11.3)	84.9±1.0 (+4.8)	**97.2**±0.9 (+5.0)	**77.9**±1.9 (+7.2)	**90.4**±2.0 (+10.6)
CLIP (fine-tuned)	30%	81.4±1.9 (+5.7)	83.8±2.8 (+3.7)	93.6±1.2 (+1.4)	**72.9**±3.3 (+2.2)	85.2±1.6 (+5.4)
SemiCLIP	30%	85.9±0.5 (+10.2)	**88.5**±0.9 (+8.5)	96.3±0.5 (+4.1)	68.9±3.4 (-1.8)	89.7±0.2 (+9.9)
ToMA (ours)	30%	**87.4**±0.8 (+11.7)	84.5±3.3 (+4.4)	**97.8**±0.7 (+5.6)	72.4±4.2 (+1.7)	**90.2**±2.2 (+10.4)
CLIP (fine-tuned)	100%	87.5±1.8	67.9±2.3	94.2±1.5	75.7±2.9	89.3±1.2

Overall, ToMA remains effective across different labeled pair ratios. At the default 10% setting, ToMA achieves the best results on RSICD-CLS, UCM-CLS, WHU-RS19, and AID, while SemiCLIP performs better on RSSCN7. When the amount of labeled supervision increases to 20% and 30%, ToMA continues to outperform SemiCLIP on RSICD-CLS, WHU-RS19, RSSCN7, and AID, although SemiCLIP remains stronger on UCM-CLS. These results suggest that the benefit of topology-aware alignment is not limited to the low-label regime, but continues to provide a useful complementary structural signal even when more labeled pairs are available.

In particular, the gains of ToMA remain consistently strong on RSICD-CLS, WHU-RS19, and AID across all three ratios. This indicates that ToMA improves the transferability of the learned representation in a robust manner, rather than relying on a specific labeled-data regime. On the other hand, the trends on UCM-CLS and RSSCN7 are less uniform. Importantly, this behavior is not unique to ToMA: the supervised CLIP baseline itself does not improve monotonically as the labeled ratio increases on these datasets. This suggests that the effect of additional labeled supervision is dataset-dependent, while ToMA still provides a robust complementary structural signal across different label regimes.

F.14 Effect of Directional Invariance in Equation (4)

In Eq. 4, we use the absolute value of the directional consistency term. This design is not intended to resolve the arbitrary orientation of an unordered edge. Indeed, reversing the order of the two

endpoints reverses both the source-space and mapped target-space difference vectors, leaving their cosine similarity unchanged. Instead, the absolute value makes the objective invariant to whether the mapped edge direction is parallel or anti-parallel to the source edge. Thus, ToMA aligns the underlying one-dimensional relational subspace induced by a topologically selected edge, rather than enforcing a fixed global orientation across modalities.

Table 31: Ablation on the edge-direction term in Eq. 4. ToMA-direction removes the absolute value from the directional consistency term. Parentheses denote performance gains over supervised CLIP, and green values indicate gains larger than one. Bold denotes the best semi-supervised result in each setting.

Method	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID
CLIP (fine-tuned)	75.7±2.2	80.1±6.8	92.2±1.5	70.7±5.5	79.8±3.8
ToMA	**84.8**±1.3 (+9.1)	**86.1**±3.4 (+6.0)	**97.9**±0.4 (+5.7)	**74.2**±1.7 (+3.5)	**87.3**±2.2 (+7.5)
ToMA-direction	77.4±1.0 (+1.7)	72.3±4.7 (-7.8)	94.2±2.4 (+2.0)	57.1±3.1 (-13.6)	81.1±1.2 (+1.3)

To verify this design choice, we compare ToMA with a variant, denoted as *ToMA-direction*, that removes the absolute value and directly uses the raw cosine similarity of edge directions. This variant penalizes anti-parallel mapped directions even when they preserve the same one-dimensional relational structure. The results are reported in Table 31. Without the absolute value, performance drops substantially across most datasets. In particular, ToMA-direction falls far behind the full ToMA on UCM-CLS and RSSCN7, and also shows weaker gains on RSICD-CLS, WHU-RS19, and AID.

These results indicate that enforcing a fixed orientation is unnecessarily restrictive for topology-aware alignment. The absolute value in Eq. 4 is therefore not a minor implementation detail, but a design choice that allows ToMA to focus on structural consistency of corresponding topological relations while remaining invariant to parallel and anti-parallel realizations of the same edge-induced relational subspace.

G Computational Time of ToMA

We also compare the run time of CLIP fine-tuned, SemiCLIP, and ToMA in Table 32. Since CLIP fine-tuned is trained in a single stage for 25 epochs, whereas SemiCLIP and ToMA use a two-stage schedule with 25 epochs in stage 1 and 15 epochs in stage 2, we report both per-epoch and total training time. For this run-time comparison only, we unify the batch size to 256 for all methods, although the default batch size of ToMA in the main experiments is 128. Since both SemiCLIP and ToMA rely on batchwise pairwise computations, changing the batch size affects the run time of both methods. The reported values should therefore be interpreted as a controlled comparison under matched conditions.

Table 32: Run time comparison between CLIP fine-tuned, SemiCLIP, and ToMA. Run time is measured under the default remote sensing setting on a single NVIDIA A100 GPU. For fair comparison, the batch size is unified to 256 only for this run-time experiment, although the default batch size of ToMA in the main experiments is 128. CLIP fine-tuned is trained for 25 epochs in a single stage, whereas SemiCLIP and ToMA are trained for 25 epochs in Stage 1 and 15 epochs in Stage 2.

	CLIP fine-tuned	SemiCLIP	ToMA
Stage 1 run time per epoch (s)	7.4	19.2	19.3
Stage 2 run time per epoch (s)		21.3	21.3
Increase over SemiCLIP (stage 1)	–	–	+0.5%
Increase over SemiCLIP (stage 2)	–	–	+0.0%
Total training time (s)	185.0	799.5	802.0
Increase over SemiCLIP (total)	–	–	+0.3%

Under this controlled setting, ToMA introduces almost no additional computational overhead over SemiCLIP: the stage-1 run time increases only from 19.2s to 19.3s per epoch, while the stage-2 run time remains unchanged at 21.3s. The total training time also increases only marginally, from

799.5s to 802.0s. This small overhead results from the lightweight PH computation, which extracts H_0-death and H_1-birth relations from the 1-skeleton without constructing 2-simplices. Overall, these results suggest that the proposed topology-aware alignment is computationally lightweight in practice.

H Compute Resources

All experiments were conducted on a local server equipped with two AMD EPYC 7513 CPUs (2.5 GHz, 32 cores each; 64 cores / 128 threads in total), 512 GB RAM, and two NVIDIA A100 GPUs with 80 GB memory each. The server also provides 512 GB of working storage and 3.5 TB of data storage. Unless otherwise noted, each individual experimental run used a single NVIDIA A100 80GB GPU. The run-time comparison reported in Appendix G was also measured on a single A100 GPU.

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: The abstract and introduction accurately describe the proposed method, its scope, and the main empirical findings.

   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: In Appendix B.1, we address the limitations of our work.

   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: This paper is primarily an empirical method paper and does not present theorem-level results requiring formal proofs.

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: All experimental settings needed for reproduction are provided in Section 4.1, Appendix C.2 and the supplementary code submission.

   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: An anonymized supplementary file containing the code and implementation details is included for reproducibility.

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: The experimental settings, hyperparameters, and training configurations are specified in Section 4.1, Appendix C.2 and the supplementary code submission.

   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: [Yes]

   Justification: Results are reported as mean and standard deviation over three runs with fixed random seeds.

   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 report the compute resources used for our experiments in Appendix H. We report the run time in the Appendix G.

   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: To the best of our knowledge, this research conforms to 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: [Yes]

    Justification: We discuss both potential benefits and possible risks of the proposed method in Appendix B.2.

    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: The paper does not release a high-risk model or dataset requiring special safeguards.

    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: [No]

    Justification: Existing datasets and methods are cited, but their licenses are not explicitly documented in the manuscript.

    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: The current submission does not release a new dataset, model, or software asset.

    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 work does not involve crowdsourcing or human-subject research.

    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: IRB approval was not applicable because the work does not involve 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: LLMs are not a core methodological component of this work.

    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.