SIMFLOW: LEARNING SIMILARITY FLOWS FOR NOISY RELATIONS IN CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-modal contrastive learning on web-collected data suffers from noisy relations: mismatched image-text pairs are often treated as positives, whereas unpaired but semantically related pairs are treated as negatives. Existing methods reduce the influence of noisy relations, potentially discarding useful information about image-text relationships. We introduce SimFlow, which provides the required similarity for noisy relations by reformulating noisy correspondence correction as a trajectory learning problem. We first identify clean and noisy relations and train an auxiliary model on the clean ones to obtain matching trajectories from clean samples resembling those in the noisy relations. To derive target similarities for noisy relations, we introduce a velocity field supervised by these trajectories to predict similarity changes based on image-text representations and detected noise types. As the standard contrastive objective attracts false positives and repels false negatives, we reduce their contributions to avoid conflicts with similarity correction. We further preserve the similarity structure of semantically matched relations across modalities through structural regularization. Extensive experiments show that SimFlow outperforms existing methods on image-text retrieval tasks under synthetic and real-world noise on Flickr30K, MS-COCO, and CC140K, and achieves strong cross-domain transferability across diverse classification datasets.

1 INTRODUCTION

Cross-modal contrastive learning aligns images and texts in a shared embedding space by increasing the similarity of matched image-text pairs and decreasing that of unpaired image-text combinations. This paradigm has been widely adopted for image-text retrieval (Xiao et al., 2025), zero-shot recognition (Jose et al., 2025), and other vision-language tasks (Tschannen et al., 2025). However, web-collected image-text pairs are often obtained with limited manual verification, and some annotated pairs contain mismatched images and texts (Huang et al., 2021). These pairs are treated as positives during training, forming false-positive relations. Conversely, unpaired but semantically related images and texts are treated as negatives, resulting in false-negative relations (Byun et al., 2024; Ma et al., 2024). Consequently, the contrastive objective can increase the similarity of mismatched pairs and decrease that of semantically related pairs, which degrade cross-modal representation learning.

To mitigate the influence of noisy relations, existing methods identify unreliable image-text relations and reduce their harmful effect on contrastive learning. For false-positive relations, some methods estimate the correspondence of annotated pairs and reduce the influence of suspected mismatches (Huang et al., 2021; Yang et al., 2023), while others filter out pairs with inconsistent cross-modal semantics (Yang et al., 2024). For false-negative relations, recent methods recover semantically related unpaired pairs as positives (Byun et al., 2024; Ma et al., 2024) or reduce repulsive force between semantically similar negatives (Pan et al., 2025).

These methods reduce incorrect attraction or repulsion caused by noisy relations. However, reducing their influence or modifying their labels may discard useful semantic information contained in noisy relations. Moreover, as noisy-relation similarities evolve throughout adaptation, determining semantically appropriate targets is challenging. As illustrated in Figure 1, matching trajectories[1] generally decrease for false-positive relations and increase for false-negative relations, but the amount of change varies considerably even within the same noise type. This observation motivates us to derive the required similarity change for each noisy relation as adaptation proceeds.

[1] We define a *matching trajectory* as the sequence of image-text similarities for a given relation over training.

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

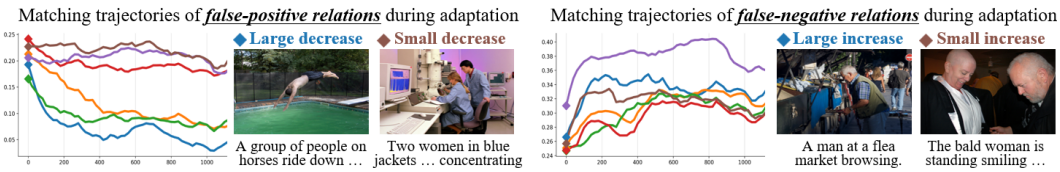


Figure 1: Image-text matching trajectories of false-positive (left) and false-negative (right) relations on Flickr30K when trained on clean relations. Each colored curve represents a different relation. False-positive relations with weak image-text semantic correspondence exhibit large similarity decreases, whereas false negatives with strong correspondence show large increases.

To this end, we propose a novel framework, SimFlow, that reformulates noisy relation correction as a matching-trajectory learning problem. Our intuition is that matching trajectories induced by training on clean relations reveal the similarity changes required to correct noisy relations, as shown in Figure 1. To obtain these trajectories, we first divide image-text relations into clean and noisy sets based on their contrastive losses and train an auxiliary model on the clean set. We then collect the matching trajectories of image-text combinations that resemble the noisy relations which serve as semantically relevant references. To guide the correction of noisy relations by leveraging the collected trajectories, we introduce a velocity field inspired by flow matching (Lipman et al., 2023). As the required similarity change differs across noisy relations, we condition the field on image-text representations and detected noise types. This enables the field to provide a target similarity that better reflects the cross-modal semantic correspondence of each noisy relation. However, the standard contrastive objective can conflict with flow-guided correction by incorrectly increasing false-positive similarities and decreasing false-negative similarities; we therefore introduce a noise-aware contrastive loss that reduces these conflicting updates. Additionally, we introduce a regularization loss to prevent flow-guided correction from distorting the intra-modal similarity structure of semantically matched relations. This framework allows SimFlow to correct noisy relations during adaptation by providing reliable supervision that reflects their underlying semantic correspondence, thereby leading to stronger cross-modal representation learning.

We evaluate SimFlow on image-text retrieval tasks under varying levels of synthetic noise on Flickr30K (Plummer et al., 2015) and MS-COCO (Lin et al., 2014), and under real-world noise on CC140K (Huang et al., 2021). To assess cross-domain transferability, we further conduct zero-shot classification on ten datasets from diverse visual domains. Experimental results show that the proposed method consistently outperforms existing methods in image-text retrieval even under severe noise. It also achieves superior zero-shot classification performance, with average accuracy gains ranging from 4.8 to 14.4 over competing methods. Our key contributions are as follows:

- We reformulate noisy correspondence correction as matching-trajectory learning to provide semantically appropriate target similarities for noisy relations during adaptation.
- We introduce a velocity field that predicts the required similarity change for noisy relations, conditioned on image-text representations and detected noise types.
- Extensive experiments show that SimFlow achieves superior performance under both synthetic and real-world noise by accurately capturing cross-modal semantic correspondence, with effective transfer to unseen domains and categories.

2 Related Work

Image-Text Matching aims to estimate a similarity score that reflects the semantic correspondence between a given image and text, enabling cross-modal retrieval and zero-shot recognition. Existing methods align the two modalities at different granularities. Local alignment methods match fine-grained fragments across modalities: they represent images and texts with regions and words, and aggregate their correspondences through cross-modal attention (Lee et al., 2018), iterative matching (Chen et al., 2020), or graph-based reasoning (Diao et al., 2021). Global alignment methods encode the entire image and sentence into a shared embedding space to measure their similarity. These methods use hard negative mining (Faghri et al., 2018) or generalized pooling (Chen et al., 2021) to improve cross-modal alignment. Large-scale contrastive learning has extended the global paradigm to web-collected data (Radford et al., 2021; Jia et al., 2021). These methods, however, treat each

paired image and text as a positive during training. Consequently, mismatched pairs from web-collected data introduce false-positive supervision and degrade cross-modal representation learning.

Noisy Correspondence Learning addresses incorrect positive and negative relations in image-text training data. Many methods exploit the memorization effect (Arpit et al., 2017; Han et al., 2018; Li et al., 2020) to distinguish clean and noisy pairs based on contrastive losses or cross-modal semantic consistency (Yang et al., 2024). Some methods use estimated correspondence as a soft margin to weaken the alignment of mismatched pairs (Huang et al., 2021; Yang et al., 2023). Others re-pair mismatched images and texts, either by retrieving counterparts from reliable pairs (Xie et al., 2025) or via optimal transport (Han et al., 2024), while some down-weight pairs according to their estimated negative impact (Zhang et al., 2024). Another line of work relabels semantically matched negatives as positives to address false-negative relations (Byun et al., 2024; Ma et al., 2024). However, reducing the influence of noisy relations or modifying their labels may discard useful information about the relation between their images and texts. In contrast, the proposed method leverages semantic information of noisy relations to derive target similarities that reflect their underlying correspondence.

Flow Matching learns a time-dependent velocity field that transports a source distribution to a target distribution, and samples are generated by integrating the learned field (Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2025). Conditional flow matching makes this training tractable by regressing the field onto per-sample target velocities computed analytically along predefined paths between source and target samples, avoiding the need to simulate the flow during training (Lipman et al., 2023; Tong et al., 2024). Flow matching has been widely applied to image synthesis (Esser et al., 2024), video generation (Polyak et al., 2024), speech generation (Le et al., 2023), and robot action generation (Black et al., 2025). Unlike these generative applications, which transport samples between distributions, the proposed method adopts the conditional flow matching objective to similarity paths constructed from clean reference trajectories, learning a velocity field that corrects the similarity of each noisy relation during adaptation.

3 METHOD

3.1 PROBLEM FORMULATION

Cross-modal contrastive learning (Radford et al., 2021) aligns images and texts in a shared embedding space by attracting paired image-text samples and repelling unpaired combinations. Given a set of image-text pairs $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$, the image encoder $\mathcal{F}_v$ and the text encoder $\mathcal{F}_t$ encode images and texts into the shared embedding space, respectively. The similarity score between X_i and Y_j is computed using cosine similarity. For an image X_i and a text Y_j, we define the symmetric contrastive loss of their relation as

$$\ell_{ij} = -\frac{1}{2}\left[\log \frac{\exp(s_{ij}/\tau)}{\sum_{k=1}^{B}\exp(s_{ik}/\tau)} + \log \frac{\exp(s_{ij}/\tau)}{\sum_{k=1}^{B}\exp(s_{kj}/\tau)}\right], \tag{1}$$

where s_{ij} denotes cosine similarity between X_i and Y_j. B and τ denote the batch size and temperature parameter, respectively. The standard contrastive objective treats each annotated pair (X_i, Y_i) as positive and the other combinations as negatives.

We consider adaptation of $\mathcal{F}_v$ and $\mathcal{F}_t$ on image-text data containing noisy cross-modal relations. The contrastive objective can incorrectly increase the similarity of mismatched annotated pairs and decrease that of semantically matched unpaired relations, causing the learned similarities to misrepresent their semantic correspondence. Correcting these similarities is challenging because the required adjustment differs across noisy relations and changes during adaptation. Our goal is to derive target similarities for noisy relations that reflect their underlying semantic correspondence, rectifying the misrepresentation induced by the contrastive objective.

3.2 SIMFLOW: SIMILARITY-FLOW MATCHING FOR NOISY RELATIONS

The proposed method, SimFlow, addresses false-positive and false-negative relations by learning how their similarities should change during adaptation. SimFlow first identifies noisy relations from their contrastive losses and trains auxiliary encoders on detected clean relations. While training the

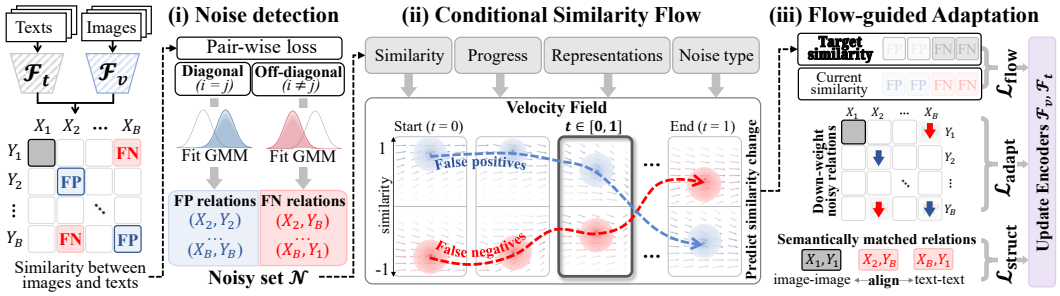


Figure 2: Overview of the proposed method. **(i)** We detect false-positive (FP) and false-negative (FN) relations by fitting GMMs to the contrastive losses. **(ii)** A velocity field predicts the required similarity change of each noisy relation based on its similarity, adaptation progress, image-text representations, and noise type. **(iii)** During adaptation, the learned flow provides target similarities for the noisy relations. We further introduce a noise-aware contrastive loss to down-weight noisy relations and a regularization loss to preserve the similarity structure of semantically matched relations.

auxiliary encoders, we record matching trajectories of image-text combinations resembling the noisy relations (Section 3.2.1). We then introduce a velocity field that captures the similarity changes in these trajectories and predicts the required change conditioned on image-text representations and detected noise types (Section 3.2.2). During adaptation, the learned field provides target similarities that guide noisy-relation correction. To support this correction, a noise-aware contrastive objective reduces conflicts between standard contrastive updates and flow-guided correction, while regularization preserves the intra-modal similarity structure of semantically matched relations (Section 3.2.3). An overview of the proposed framework is illustrated in Figure 2.

3.2.1 Noisy Relation Discovery

Before correcting noisy relations, we distinguish false-positive (FP) and false-negative (FN) relations from clean ones. We first detect false positives among the annotated image-text pairs. To do so, we fit a two-component Gaussian mixture model (GMM) to their contrastive losses ℓ_{ii}, computed from the embeddings produced by $\mathcal{F}_v$ and $\mathcal{F}_t$. Since mismatched annotated pairs tend to have larger losses (Huang et al., 2021), we associate the higher-mean component with false positives. The GMM gives p_{ii}, the probability that (X_i, Y_i) is a false positive. We then detect false negatives among unpaired image-text relations. Semantically matched unpaired relations tend to have losses similar to those of clean positive relations. We therefore fit another two-component GMM to the losses of the unpaired relations and the remaining paired relations. The lower-mean component is associated with false negatives, which gives p_{ij}, the probability that (X_i, Y_j) is a false negative. We identify a relation as a false positive or false negative if its corresponding probability exceeds a predefined threshold δ. The detected false-positive and false-negative relations form the noisy set $\mathcal{N} = \mathcal{N}^{\text{FP}} \cup \mathcal{N}^{\text{FN}} = \{(X_m, Y_m)\}_{m=1}^{M}$, with its corresponding noise probability p_m, where m indexes the detected noisy relations. The remaining relations constitute the clean set $\mathcal{C}$.

Reference Trajectory Acquisition. To obtain reliable reference trajectories that represent the sequence of similarities over adaptation, while reducing contamination from noisy relations, we train auxiliary image and text encoders, $\mathcal{F}_v^{\text{aux}}$ and $\mathcal{F}_t^{\text{aux}}$, on the detected clean relations in $\mathcal{C}$ using the standard cross-modal contrastive loss. Nevertheless, false negatives may arise because the contrastive objective treats image-text combinations from different pairs within a batch as negatives, even when some of them are semantically matched. To reduce such false negatives, we construct each batch by greedily selecting relations whose images and texts exhibit low similarity to those of the relations already selected.

To obtain a reference trajectory relevant to each noisy relation, we construct a reference relation from clean samples that resemble its image and text. Using the auxiliary encoders, we retrieve an image from $\mathcal{C}$ that resembles the image of the noisy relation and a text from $\mathcal{C}$ that resembles its text, and combine them to form the reference relation. We track the similarity of each reference relation while training the auxiliary encoders and record the similarity at K uniformly spaced points $t_k = k/(K-1)$ on the normalized training progress $t \in [0, 1]$, where $k \in \{0, \ldots, K-1\}$. For the m-th noisy relation, this yields a reference trajectory of similarities $\{s_m^{\text{ref}}(t_k)\}_{k=0}^{K-1}$.

3.2.2 CONDITIONAL SIMILARITY FLOW

The collected reference trajectories capture how the similarity of each reference relation changes while the auxiliary encoders are trained on the detected clean relations. To correct each detected noisy relation accordingly, we construct a correction path for the relation, which starts at its initial similarity and follows the reference trajectory of its associated reference relation. We then introduce a velocity field that predicts the required similarity change along these correction paths, conditioned on the context of each relation.

Similarity Path Construction. The collected trajectory reflects the similarity evolution of the reference relation, but the noisy relation may start from a different similarity. We therefore construct a correction path from the initial similarity of the noisy relation to the collected similarities. Let $s_m(0)$ denote the initial similarity of the m-th noisy relation. We compute the correction path at the collected progress points as

$$z_m(t_k) = (1 - t_k) \cdot s_m(0) + t_k \cdot s_m^{\text{ref}}(t_k). \tag{2}$$

This construction starts the path at $s_m(0)$ and uses the reference trajectory to guide the subsequent similarity evolution. Because $z_m(t_k)$ is available at the progress points where the reference similarities are collected, we linearly interpolate adjacent values to obtain $z_m(t)$ at intermediate progress. The interpolation yields a constant velocity $u_m(t)$, which represents the required rate of similarity change for the noisy relation.

To predict this velocity, we introduce a velocity field v_ϕ. For the m-th noisy relation, the field takes $z_m(t)$ and t and predicts the velocity $u_m(t)$. To reflect that the required similarity change differs across noisy relations, we additionally condition the field on the relation context c_m, which contains the concatenated image and text embeddings and the detected noise type. We train v_ϕ by minimizing the difference between the predicted velocity and $u_m(t)$:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{m,\, t \,\sim\, \mathcal{U}[0,1]} \Big[\big(v_\phi(z_m(t),\, t,\, c_m) - u_m(t) \big)^2 \Big], \tag{3}$$

where t is sampled uniformly from $[0, 1]$. This objective allows v_ϕ to predict the required similarity change of each noisy relation along its correction path conditioned on its relation context.

3.2.3 FLOW-GUIDED ADAPTATION

After training the velocity field, we adapt the image and text encoders, $\mathcal{F}_v$ and $\mathcal{F}_t$, on noisy image-text data using the learned similarity flow. At the current adaptation progress $t_n \in [0, 1]$, the encoders produce the similarity $s_m(t_n)$ for the m-th noisy relation. The velocity field predicts the similarity change over adaptation progress. To obtain the target similarity at t_n, we add the accumulated predicted change to the initial similarity of the correction path:

$$\hat{s}_m(t_n) \triangleq s_m(0) + \int_0^{t_n} v_\phi\big(\hat{s}_m(t), t, c_m\big)\, dt, \tag{4}$$

where $\hat{s}_m(t)$ denotes the similarity obtained by integrating the velocity field, with $\hat{s}_m(0) = s_m(0)$. The integral is approximated using Euler's method (Liu et al., 2023). The resulting similarity $\hat{s}_m(t_n)$ serves as the target similarity for the m-th noisy relation at t_n.

To guide the similarities of noisy relations toward their flow-derived targets, we train the image and text encoders by minimizing

$$\mathcal{L}_{\text{flow}} = \frac{1}{|\mathcal{N}_B|} \sum_{m \in \mathcal{N}_B} p_m \cdot \big(s_m(t_n) - \text{sg}[\hat{s}_m(t_n)] \big)^2, \tag{5}$$

where p_m denotes the estimated probability that the m-th relation is noisy. $\text{sg}[\cdot]$ denotes the stop-gradient operation to prevent the update of the velocity field during adaptation and $\mathcal{N}_B$ denotes the set of detected noisy relations in the current batch. This loss guides $s_m(t_n)$ toward the flow-derived target $\hat{s}_m(t_n)$, with stronger correction for relations with higher estimated noise probability p_m.

Training Objective. The flow-guided loss corrects detected noisy relations toward their target similarities. However, the standard contrastive objective still pulls every annotated pairs together and

pushes unpaired combinations apart, even when a relation is detected as a false positive or false negative. These updates can oppose the flow-guided correction. To reduce this conflict, we reduce the contribution of detected noisy relations according to their estimated probabilities:

$$\mathcal{L}_{\text{adapt}}^{\text{IT}} = -\frac{1}{B}\sum_{i=1}^{B}(1-p_{ii})\cdot\log\frac{\exp(s_{ii}/\tau)}{\exp(s_{ii}/\tau)+\sum_{k\neq i}(1-p_{ik})\cdot\exp(s_{ik}/\tau)}, \tag{6}$$

where p_{ii} and p_{ik} are set to 0 for relations not included in $\mathcal{N}$. The factors $1-p_{ii}$ and $1-p_{ik}$ reduce the attraction of detected false-positive relations and the repulsion of detected false-negative relations, respectively. We define the text-to-image counterpart symmetrically $\mathcal{L}_{\text{adapt}}^{\text{TI}}$ and combine the two directions as $\mathcal{L}_{\text{adapt}} = \frac{1}{2}(\mathcal{L}_{\text{adapt}}^{\text{IT}} + \mathcal{L}_{\text{adapt}}^{\text{TI}})$. This mitigates the effect of contrastive updates that interfere with the flow-derived correction while preserving the contribution of detected clean relations to the contrastive objective.

Additionally, flow-guided adaptation rectifies noisy cross-modal relations, but may inadvertently distort the intra-modal similarity structure of semantically matched relations. For example, for two semantically matched relations, the similarity between their images is expected to be consistent with that between their texts. However, an image or text from a noisy relation may also belong to one of these semantically matched relations, causing inconsistency between the intra-modal similarities during correction. To prevent this, we introduce a regularization loss that enforces the similarity between images to match the similarity between their corresponding texts:

$$\mathcal{L}_{\text{struct}} = \frac{1}{|\mathcal{S}_B|^2}\sum_{(i,j)\in\mathcal{S}_B}\sum_{(a,b)\in\mathcal{S}_B}\left(\mathcal{F}_v(X_i)^\top\mathcal{F}_v(X_a) - \mathcal{F}_t(Y_j)^\top\mathcal{F}_t(Y_b)\right)^2, \tag{7}$$

where $\mathcal{S}_B$ contains the detected positive and false-negative relations in the current batch. This loss aligns the relative similarities among clean images with those among their corresponding texts, maintaining consistent similarity structures across the modalities.

To summarize, the velocity field is trained with $\mathcal{L}_{\text{CFM}}$ to provide target similarities for noisy relations. Then, the image and text encoders, $\mathcal{F}_v$ and $\mathcal{F}_t$, are jointly optimized using the noise-aware contrastive objective, flow-guided correction, and structural regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{adapt}} + \lambda_{\text{flow}}\,\mathcal{L}_{\text{flow}} + \lambda_{\text{struct}}\,\mathcal{L}_{\text{struct}}, \tag{8}$$

where λ_{flow} and λ_{struct} control the contributions of the corresponding losses, respectively.

4 Experiments

4.1 Setup

Datasets. We evaluated the proposed method on image-text retrieval tasks using Flickr30K (Plummer et al., 2015), MS-COCO (Lin et al., 2014), and Conceptual Captions (CC) (Sharma et al., 2018). We injected synthetic noise into Flickr30K and MS-COCO with noise ratios of 20%, 40%, 60%, and 80%, following the previous practice (Huang et al., 2021). For MS-COCO, we followed the standard 1K test protocol and reported results averaged over five 1K test folds. For real-world noise, we used CC140K, which contains approximately 140K accessible image-text pairs from CC152K (Huang et al., 2021). To assess the generalization of the learned cross-modal representations, we further conducted zero-shot classification on 10 classification datasets across diverse visual domains: Aircraft (Maji et al., 2013), Caltech101 (Fei-Fei et al., 2004), CIFAR100 (Krizhevsky, 2009), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), Flowers (Nilsback & Zisserman, 2008), Food (Bossard et al., 2014), MNIST (LeCun et al., 1998), OxfordPet (Parkhi et al., 2012), and StanfordCars (Krause et al., 2013).

Architecture. We used the image and text encoders in CLIP ViT-B/32 (Radford et al., 2021) as the backbone. The auxiliary encoders adopt the same architecture and initialization as the backbone. For the velocity field v_ϕ, we used a three-layer MLP with a hidden dimension of 128. The auxiliary encoders and velocity field were used during training and discarded for inference.

Implementation Details. We trained the backbone image and text encoders for ten epochs and the auxiliary encoders for three epochs. Both were optimized with Adam using a learning rate of

Table 1: Retrieval results on Flickr30K and MS-COCO under varying noise ratios.

Method	Flickr30K							MS-COCO						
	Image → Text			Text → Image			RSum	Image → Text			Text → Image			RSum
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
Noise ratio 20%														
NCR (Huang et al., 2021)	85.2	**97.9**	98.9	72.5	91.3	95.1	540.9	79.5	95.2	98.0	66.3	89.2	94.1	522.3
BiCro (Yang et al., 2023)	86.3	97.4	99.1	72.2	90.9	95.2	541.1	79.6	95.3	97.8	66.3	89.4	94.1	522.5
MSCN (Han et al., 2023)	86.6	96.9	99.0	73.1	92.0	95.4	543.1	80.2	96.1	98.7	67.3	91.3	96.0	529.6
GSC (Zhao et al., 2024)	82.1	96.3	98.0	70.0	90.5	94.6	531.5	75.2	94.4	97.6	63.6	89.1	94.5	514.4
ESC (Yang et al., 2024)	85.0	97.8	99.0	72.6	91.2	95.2	540.8	79.7	95.0	97.9	66.6	89.4	94.3	522.9
SPS (Xie et al., 2025)	86.3	97.1	99.3	73.4	92.2	95.9	544.3	79.7	95.8	98.3	67.2	91.1	96.2	528.3
AdaCL (Pan et al., 2025)	85.9	97.0	99.1	73.1	92.1	95.6	542.8	80.6	95.9	98.4	66.8	90.8	95.7	528.2
DCL (Zhong et al., 2026)	85.6	96.5	98.4	72.4	91.9	95.6	540.4	78.6	95.3	98.1	65.8	90.7	95.8	524.4
SimFlow (Ours)	**88.5**	**97.9**	**99.7**	**76.0**	**94.0**	**97.0**	**553.1**	**82.6**	**96.9**	**98.8**	**70.0**	**92.6**	**96.8**	**537.8**
Noise ratio 40%														
NCR (Huang et al., 2021)	86.1	96.6	98.5	71.0	89.8	93.8	535.8	75.5	91.1	94.9	62.7	85.4	90.4	499.9
BiCro (Yang et al., 2023)	86.7	96.6	98.8	71.0	89.9	93.7	536.7	74.7	90.8	95.2	60.9	83.6	88.8	494.0
MSCN (Han et al., 2023)	87.2	97.5	98.8	73.1	91.8	95.4	543.8	80.0	96.0	98.5	66.9	90.9	96.0	528.2
GSC (Zhao et al., 2024)	79.7	95.2	98.1	65.7	87.5	92.0	518.2	75.2	93.6	97.3	62.6	88.6	94.5	511.8
ESC (Yang et al., 2024)	86.6	96.8	98.4	71.1	89.7	93.5	536.1	75.9	91.0	95.3	62.8	85.4	90.5	501.0
SPS (Xie et al., 2025)	86.5	97.1	99.1	73.0	92.2	95.6	543.6	79.1	95.5	98.2	66.6	90.9	96.0	526.5
AdaCL (Pan et al., 2025)	86.5	97.3	98.8	72.3	91.0	94.6	540.6	80.1	95.4	97.9	66.4	90.5	95.4	525.7
DCL (Zhong et al., 2026)	85.1	97.0	99.0	72.9	92.6	95.7	542.2	79.1	95.6	98.3	65.8	90.4	95.7	525.0
SimFlow (Ours)	**88.7**	**97.8**	**99.3**	**75.7**	**93.5**	**96.7**	**551.7**	**81.6**	**96.6**	**98.7**	**69.5**	**92.5**	**96.9**	**535.8**
Noise ratio 60%														
NCR (Huang et al., 2021)	85.5	96.4	98.5	68.3	88.6	92.9	530.2	58.9	77.6	85.5	45.5	64.5	69.9	402.0
BiCro (Yang et al., 2023)	83.7	96.2	98.3	68.0	88.4	92.9	527.4	51.9	64.2	73.6	40.8	59.5	65.1	355.0
MSCN (Han et al., 2023)	86.7	97.0	99.0	71.7	91.5	95.8	541.6	80.2	95.7	98.3	66.9	90.9	95.9	527.8
GSC (Zhao et al., 2024)	76.9	93.0	97.0	62.3	84.9	90.1	504.2	73.2	92.8	97.1	60.8	87.0	93.4	504.3
ESC (Yang et al., 2024)	85.4	96.3	98.8	68.3	88.6	92.7	530.1	59.1	78.1	86.4	48.1	65.9	70.0	407.7
SPS (Xie et al., 2025)	85.2	96.4	98.6	72.4	92.0	95.5	540.0	77.9	95.1	98.0	65.7	90.2	95.8	522.8
AdaCL (Pan et al., 2025)	86.6	97.3	98.4	71.2	91.0	94.9	539.4	79.3	95.4	98.2	65.7	90.0	95.4	524.0
DCL (Zhong et al., 2026)	85.9	97.2	98.9	73.3	92.3	95.8	543.4	79.3	95.1	98.2	66.0	90.3	95.5	524.5
SimFlow (Ours)	**88.1**	**97.4**	**99.5**	**75.1**	**93.2**	**96.9**	**550.2**	**81.0**	**96.5**	**98.7**	**69.3**	**92.1**	**96.8**	**534.3**
Noise ratio 80%														
NCR (Huang et al., 2021)	60.3	82.2	89.9	44.4	65.5	72.2	414.6	50.9	81.0	89.8	42.5	73.2	83.4	420.7
BiCro (Yang et al., 2023)	44.6	72.9	82.4	23.0	44.1	54.0	321.0	45.7	77.1	86.5	35.9	66.3	77.4	388.9
MSCN (Han et al., 2023)	86.6	97.3	98.6	71.4	91.5	95.3	540.7	78.4	95.8	98.7	66.2	90.4	95.5	525.0
GSC (Zhao et al., 2024)	72.2	91.3	95.3	55.7	79.3	86.4	480.2	73.4	93.1	97.1	60.6	86.9	93.1	504.1
ESC (Yang et al., 2024)	62.6	84.2	91.2	44.8	64.5	71.9	419.2	50.0	80.5	89.1	40.5	71.7	82.5	414.4
SPS (Xie et al., 2025)	86.3	**97.5**	99.1	71.9	91.6	95.5	541.9	77.9	95.2	98.1	65.6	90.2	95.6	522.4
AdaCL (Pan et al., 2025)	84.7	95.8	98.0	69.6	89.5	94.0	531.6	78.4	94.7	97.8	64.0	89.2	94.8	519.0
DCL (Zhong et al., 2026)	85.6	97.0	98.8	72.3	92.1	95.6	541.4	79.0	95.4	98.1	65.4	89.9	95.4	523.2
SimFlow (Ours)	**88.2**	**97.5**	**99.3**	**74.8**	**93.4**	**96.6**	**549.8**	**81.1**	**96.4**	**98.8**	**68.8**	**92.0**	**96.7**	**533.8**

10^{-5} and a batch size of 512. For auxiliary training, we randomly sampled 5,000 relations from the detected clean set $\mathcal{C}$ and collected their reference similarities at $K = 10$ points during training. We trained the velocity field for 6,000 iterations using Adam with a learning rate of 10^{-3} and a batch size of 4,096. We set the noise detection threshold δ to 0.5 and the loss coefficients λ_{struct} and λ_{flow} to 25 and 5, respectively. We conducted all experiments on an NVIDIA A100 GPU. Please refer to the appendix for more implementation details.

Baselines. We compared SimFlow with noisy correspondence correction methods: NCR (Huang et al., 2021), BiCro (Yang et al., 2023), MSCN (Han et al., 2023), GSC (Zhao et al., 2024), ESC (Yang et al., 2024), and SPS (Xie et al., 2025). We also included AdaCL (Pan et al., 2025) and DCL (Zhong et al., 2026), which consider false-negative relations.

Evaluation Metrics. We evaluated image-to-text and text-to-image retrieval on Flickr30K, MS-COCO, and CC140K. For both retrieval tasks, we reported Recall@K (R@K, $K \in 1, 5, 10$), which measures the percentage of queries whose correct match appears within the top-K retrieved results. We also reported RSum, the sum of the six recall scores from image-to-text and text-to-image retrieval. For cross-domain transfer, we reported zero-shot classification performance on 10 classification datasets using top-1 accuracy.

4.2 Results

Results on Synthetic and Real-world Noise. We evaluated the proposed method and the baselines on Flickr30K and MS-COCO under synthetic noise ratios of 20%, 40%, 60%, and 80%. Table 1 shows the image-to-text and text-to-image retrieval results. Overall, SimFlow achieves the highest RSum across noise ratios on both datasets. As the noise ratio increases, NCR, BiCro, and ESC show

Table 2: Retrieval results on CC140K with real-world noisy correspondences.

Method	Image → Text			Text → Image			RSum
	R@1	R@5	R@10	R@1	R@5	R@10	
NCR (Huang et al., 2021)	68.6	87.8	92.2	67.2	87.9	92.6	496.3
BiCro (Yang et al., 2023)	68.0	88.2	91.6	67.4	88.0	92.6	495.7
MSCN (Han et al., 2023)	69.0	89.3	93.3	68.2	90.3	93.3	503.4
GSC (Zhao et al., 2024)	63.3	83.7	88.1	63.4	83.9	88.3	470.7
ESC (Yang et al., 2024)	68.3	88.1	91.6	67.0	87.7	92.6	495.2
SPS (Xie et al., 2025)	69.5	88.7	92.7	70.1	89.4	93.7	504.0
AdaCL (Pan et al., 2025)	69.4	88.5	92.5	70.2	88.4	93.6	502.5
DCL (Zhong et al., 2026)	68.5	88.1	92.7	67.8	88.9	93.2	499.1
SimFlow (Ours)	**72.2**	**90.6**	**95.2**	**70.5**	**90.5**	**94.6**	**513.7**

Table 3: Cross-domain transfer results on ten datasets using encoders trained on CC140K.

Method	Average	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars
NCR (Huang et al., 2021)	52.3	13.9	80.0	60.0	40.1	36.4	51.4	74.7	44.9	78.3	43.1
BiCro (Yang et al., 2023)	51.7	12.9	79.3	58.9	39.9	36.4	50.2	74.5	45.8	78.5	41.0
MSCN (Han et al., 2023)	53.0	14.6	80.1	57.4	40.2	42.9	48.2	73.5	48.1	81.1	43.5
GSC (Zhao et al., 2024)	46.5	13.0	76.3	50.7	35.5	26.9	46.0	69.9	39.2	66.0	41.9
ESC (Yang et al., 2024)	51.6	13.9	80.0	59.9	39.7	34.8	51.0	74.8	41.7	78.2	42.2
SPS (Xie et al., 2025)	54.1	14.1	80.5	61.5	41.8	33.8	58.3	75.4	48.6	82.1	44.5
AdaCL (Pan et al., 2025)	56.1	15.1	81.2	64.3	40.8	48.3	57.0	77.7	43.8	83.5	49.4
DCL (Zhong et al., 2026)	53.1	14.4	80.7	64.6	42.1	**51.1**	57.7	76.7	25.4	76.3	42.4
SimFlow (Ours)	**60.9**	**17.4**	**82.4**	**67.6**	**46.3**	50.1	**65.4**	**81.2**	**54.9**	**86.2**	**57.2**

substantial performance degradation, particularly on MS-COCO. At higher noise ratios, MSCN, SPS, and DCL perform better than the other baselines, but SimFlow consistently achieves higher RSum than these methods on both datasets. At 80% noise, SimFlow outperforms the strongest competing methods, SPS on Flickr30K and MSCN on MS-COCO, with performance gaps of 7.9 and 8.8 in terms of RSum, respectively. These results show that the learned similarity flow provides appropriate target similarities for noisy relations, which enables effective correction even when noisy pairs constitute most of the training data.

We further evaluated SimFlow under real-world noise on CC140K. Table 2 reports the retrieval results. SimFlow achieves the highest RSum of 513.7, outperforming the strongest competing method, SPS, with a performance gap of 9.7. It also achieves the highest R@1 in both retrieval directions, with 72.2 for image-to-text retrieval and 70.5 for text-to-image retrieval. These results demonstrate the effectiveness of SimFlow on web-collected data with real-world correspondence noise.

Cross-domain Transferability. To evaluate cross-domain transferability, we evaluated the encoders trained on CC140K through zero-shot classification on ten classification datasets. Table 3 reports the zero-shot classification results. The proposed method achieves the highest average zero-shot accuracy of 60.9, outperforming SPS, DCL, and MSCN by 6.8, 7.8, and 7.9, respectively. These results show that SimFlow learns cross-modal representations that transfer effectively to unseen domains and classes despite noisy correspondence in the training data. Please refer to the appendix for additional results on cross-domain transfer evaluated through linear probing.

4.3 Analysis

Ablation Study. We conducted additional experiments on MS-COCO under 20% and 80% noise to evaluate the contribution of three components: flow-guided correction $\mathcal{L}_{\text{flow}}$, structural regularization $\mathcal{L}_{\text{struct}}$, and noise-aware contrastive loss $\mathcal{L}_{\text{adapt}}$. Table 4 reports the ablation results. Overall, the proposed method consistently achieves the best performance at both noise ratios. Compared to the proposed method, the method without $\mathcal{L}_{\text{flow}}$ shows performance drops of 16.0 and 18.1 in terms of RSum at 20% and 80% noise, respectively, showing the benefit of flow-guided correction with target similarities for noisy relations. Ours without $\mathcal{L}_{\text{struct}}$ shows performance drops of 11.4 and 13.0 in RSum at the two noise ratios. Since flow-guided correction updates the shared image and text encoders through noisy-relation targets, $\mathcal{L}_{\text{struct}}$ helps maintain consistent similarity structures for semantically matched relations across modalities. The method without $\mathcal{L}\text{adapt}$ shows performance

Table 4: Ablation results with different component combinations.

$\mathcal{L}_{\text{flow}}$	$\mathcal{L}_{\text{struct}}$	$\mathcal{L}_{\text{adapt}}$	MS-COCO noise ratio 20%							MS-COCO noise ratio 80%						
			Image → Text			Text → Image			RSum	Image → Text			Text → Image			RSum
			R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
✓	✓	✓	**82.6**	**96.9**	**98.8**	**70.0**	**92.6**	**96.8**	**537.8**	**81.1**	**96.4**	**98.8**	**68.8**	**92.0**	**96.7**	**533.8**
-	✓	✓	78.2	95.2	97.9	65.5	89.9	95.1	521.8	76.7	94.4	97.7	63.8	88.6	94.5	515.7
✓	-	✓	79.5	95.1	98.1	66.6	90.9	96.1	526.4	77.6	95.1	97.9	64.7	90.1	95.5	520.8
✓	✓	-	79.2	95.6	98.1	65.9	90.3	95.8	524.9	76.9	94.4	97.9	63.5	89.0	94.8	516.6



Figure 3: Matching trajectories of false-positive (top) and false-negative (bottom) relations, with examples showing similarities before and after training and their target similarities.

drops of 12.9 and 17.2 in RSum at noise ratios of 20% and 80%, respectively. This is because using the naïve contrastive loss continues to increase false-positive similarities and decrease false-negative similarities. These updates interfere with the correction toward the flow-derived targets, weakening the effect of the correction and leading to larger performance degradation as the noise ratio increases.

Visualization of Similarity Correction. To examine similarity correction by the proposed method, we visualize the matching trajectories of randomly selected relations throughout training on Flickr30K with 80% noise in Figure 3. For comparison, we included a baseline fine-tuned with the standard cross-modal contrastive loss (CLIP-FT) to examine similarity changes without correction. Compared with CLIP-FT, SimFlow lowers false-positive similarities and raises false-negative similarities toward their flow-derived targets. The randomly selected image-text examples further show that SimFlow provides different target similarities according to their semantic correspondence, even within the same noise type. The resulting similarities are effectively corrected toward these targets, reflecting the degree of semantic correspondence in individual relations. Please refer to the appendix for additional experiments, including hyperparameter sensitivity, evaluations with different backbones, computational overhead, and qualitative examples of noisy relations and their references.

5 Conclusion

In this paper, we proposed SimFlow, which learns the similarity changes required to correct noisy relations in cross-modal contrastive learning. The proposed method reformulates the correction problem as matching-trajectory learning to provide appropriate target similarities without discarding useful semantic information in noisy relations. To capture these changes, we learn a velocity field from matching trajectories obtained from detected clean relations. The learned field provides target similarities used to adjust noisy-relation similarities to better represent cross-modal semantic correspondence. To prevent the contrastive loss from interfering with these corrections, we introduce a noise-aware contrastive loss that adjusts the contribution of detected noisy relations to support their correction toward the flow-derived targets. We further preserve the similarity structure of semantically matched relations through structural regularization. Extensive experiments show that SimFlow consistently outperforms existing methods on image-text retrieval under synthetic and real-world noise, while learning cross-modal representations that transfer effectively to unseen domains.

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 233–242. PMLR, 2017.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. In *Robotics: Science and Systems*, 2025.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101: Mining discriminative components with random forests. In *European Conference on Computer Vision*, pp. 446–461. Springer, 2014.

Jaeseok Byun, Dohoon Kim, and Taesup Moon. Mafa: Managing false negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27314–27324, 2024.

Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12655–12663, 2020.

Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15789–15798, 2021.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1218–1226, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 12606–12633. PMLR, 2024.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference*, pp. 12, 2018.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pp. 178–178, 2004.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8527–8537, 2018.

Haochen Han, Kaiyao Miao, Qinghua Zheng, and Minnan Luo. Noisy correspondence learning with meta similarity correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7517–7526, 2023.

Haochen Han, Qinghua Zheng, Guang Dai, Minnan Luo, and Jingdong Wang. Learning to rematch mismatched pairs for robust cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26679–26688, 2024.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242.

Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021.

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

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, Oriane Siméoni, Huy V. Vo, Patrick Labatut, and Piotr Bojanowski. DINOv2 meets text: A unified framework for image- and pixel-level vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24905–24916, 2025.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

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

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*, pp. 201–216, 2018.

Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Xinran Ma, Mouxing Yang, Yunfan Li, Peng Hu, Jiancheng Lv, and Xi Peng. Cross-modal retrieval with noisy correspondence via consistency refining and mining. *IEEE Transactions on Image Processing*, 33:2587–2598, 2024. doi: 10.1109/TIP.2024.3374221.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729, 2008.

Renjie Pan, Jihao Dong, and Hua Yang. Discovering clone negatives via adaptive contrastive :earning for image-text matching. In *International Conference on Learning Representations*, 2025.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012.

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

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

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

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

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

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

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

Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. Flair: Vlm with fine-grained language-informed image representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24884–24894, 2025.

Yucheng Xie, Songyue Cai, Tao Tong, Ping Hu, and Xiaofeng Zhu. Seeking proxy point via stable feature space for noisy correspondence learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 2072–2080, 2025. doi: 10.24963/ijcai.2025/231.

Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19883–19892, June 2023.

Yuchen Yang, Likai Wang, Erkun Yang, and Cheng Deng. Robust noisy correspondence learning with equivariant similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17700–17709, June 2024.

Xu Zhang, Hao Li, and Mang Ye. Negative pre-aware for noisy cross-modal matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7341–7349, 2024.

Zihua Zhao, Mengxi Chen, Tianjie Dai, Jiangchao Yao, Bo Han, Ya Zhang, and Yanfeng Wang. Mitigating noisy correspondence by geometrical structure consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27381–27390, June 2024.

Zihua Zhao, Feng Hong, Mengxi Chen, Pengyi Chen, Benyuan Liu, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Differential-informed sample selection accelerates multimodal contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2930–2940, 2025.

Fangming Zhong, Xinyu He, Haiquan Yu, Xiu Liu, and Suhua Zhang. Negative can be positive: A stable and noise-resistant complementary contrastive learning for cross-modal matching. *Information Fusion*, 130:104156, 2026. doi: 10.1016/j.inffus.2026.104156.

A Appendix

A.1 Additional Implementation Details

Details of Training Procedure. To clarify the order of relation identification, auxiliary training, and encoder adaptation, we additionally describe the overall training procedure. We first trained the image and text encoders for two epochs and identified the clean and noisy relations. We then trained the auxiliary encoders on the detected clean relations, collected the reference trajectories, and trained the velocity field. The auxiliary encoders were discarded after trajectory acquisition. We then adapted the image and text encoders for eight additional epochs using $\mathcal{L}_{\text{adapt}}$, $\mathcal{L}_{\text{flow}}$, and $\mathcal{L}_{\text{struct}}$, while keeping the velocity field fixed. At the beginning of every adaptation epoch, we re-identified the clean and noisy relations using the image and text encoders. For a detected noisy relation, we constructed the relation context c_m from the image and text representations at that epoch and the detected noise type. The adaptation progress t_n was set to 0.125 in the first adaptation epoch and increased by 0.125 per epoch, reaching 1.00 in the final epoch. At every epoch, we obtained the target similarity for a detected noisy relation by integrating the learned velocity field from 0 to t_n using the Euler method with eight equal steps over the interval.

Details of Detecting Noisy Relations. In the cross-modal contrastive learning objective (Radford et al., 2021), diagonal image-text relations in a batch are treated as positives, whereas off-diagonal relations are treated as negatives. Most off-diagonal image-text relations in a batch are semantically unrelated, which makes false-negative relations difficult to identify among them. We first narrowed the search space for potential false negatives using mutual nearest neighbors. We treated a relation (X_i, Y_j) as a false-negative candidate when Y_j is the most similar text to X_i and X_i is also the most similar image to Y_j within the batch. This gives at most B candidate relations per batch. To fit the GMM, we used equal numbers of losses from detected clean pairs and randomly sampled off-diagonal relations, so that the much larger off-diagonal group does not dominate the fitting. Before fitting the GMM, we removed the lowest 1% of the sampled off-diagonal losses because they are more likely to be false negatives. We then applied the fitted GMM to the candidate relations and estimate the false-negative probability of each relation.

Details of Velocity Field v_ϕ. The velocity field predicts the similarity change required for each noisy relation. To account for differences across noisy relations, we condition the field on the image and text representations and detected noise type of each relation. During velocity-field training, the auxiliary encoders produce the image and text representations. We concatenate these representations and project them to 64 dimensions with a linear layer. We represent the detected noise type (false positive or false negative), with a learnable 8-dimensional embedding vector. The projected image-text representation and the noise-type embedding constitute the relation context c_m. During flow-guided adaptation, we construct c_m at every epoch for the detected noisy relations, using their detected noise types and the image and text representations produced by the encoders being adapted. Although the representations are produced by different encoders, we use the same architecture and initialization for the auxiliary and backbone encoders to maintain compatibility between their representation spaces.

B Additional Results

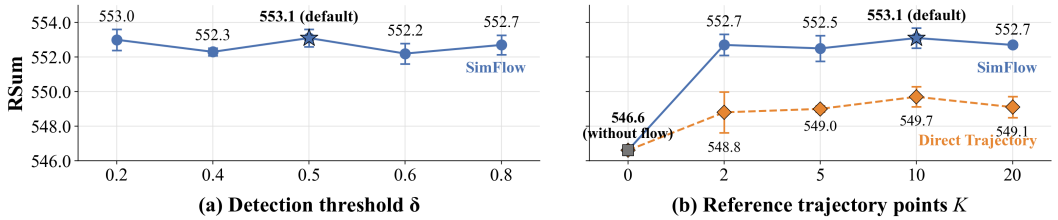


Figure 4: Sensitivity to the detection threshold δ (left) and the number of reference trajectory points K (right) on Flickr30K under 20% noise.

Results with Varying Detection Threshold. To examine the effect of the detection threshold δ, we conducted additional experiments on Flickr30K under 20% noise. Figure 4 (a) shows that the RSum ranges from 552.2 to 553.1 as δ varies from 0.2 to 0.8. These results show that SimFlow is robust to the choice of the detection threshold. Despite changes in the relations identified as noisy as δ varies, SimFlow applies stronger flow-guided correction and reduces contrastive updates more for relations with higher estimated noise probabilities. Those with lower probabilities receive weaker flow-guided correction and a greater contribution from the standard contrastive objective, which helps maintain stable performance across different detection thresholds.

Results with Varying Numbers of Reference Trajectory Points. To examine the effect of the number of collected trajectory points K and the benefit of the velocity field, we conducted additional experiments on Flickr30K under 20% noise. Figure 4 (b) shows stable performance across $K \in \{2, 5, 10, 20\}$, indicating that SimFlow is not sensitive to the number of collection points. We also compared SimFlow with a variant that directly uses the reference trajectory as the correction target. This variant obtains each target by interpolating the two nearest collected reference similarities, without the velocity field. The performance gap between SimFlow and this variant ranges from 3.4 to 3.9 in RSum for $K \in \{2, 5, 10, 20\}$. These results show that conditioning the velocity field on c_m allows it to predict the required similarity change for the correction path and provide target similarities that better reflect the image-text content and detected noise type than directly using the reference trajectory.

Table 5: Retrieval results on Flickr30K with different backbones under 20% and 80% noise.

Backbone	Pretraining	Method	Flickr30K noise ratio 20%							Flickr30K noise ratio 80%						
			Image → Text			Text → Image			RSum	Image → Text			Text → Image			RSum
			R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
CLIP ViT-B/32	LAION-2B	NCR	87.0	96.6	98.4	69.9	89.2	93.1	534.2	80.5	93.7	96.9	61.8	81.9	87.3	502.1
		BiCro	85.2	96.3	98.2	69.0	88.8	92.9	530.4	76.9	91.5	95.3	58.6	79.3	85.1	486.6
		MSCN	86.4	97.7	**99.2**	71.7	91.4	94.9	541.3	84.8	95.7	98.4	70.5	89.9	94.3	533.6
		GSC	80.9	93.9	96.6	65.8	87.2	91.8	516.1	80.5	95.0	97.5	64.1	86.5	91.9	515.6
		ESC	86.4	96.3	98.3	69.6	89.1	93.2	532.9	79.3	93.7	97.2	61.4	81.4	86.9	500.0
		SPS	87.0	96.7	98.7	72.0	91.2	94.7	540.4	85.7	96.5	98.8	70.9	89.6	94.0	535.5
		AdaCL	83.2	95.2	98.3	69.2	90.6	95.7	532.2	82.9	94.0	98.3	70.2	87.6	92.1	525.1
		DCL	85.7	97.0	98.7	72.9	92.1	95.3	541.6	87.0	**97.9**	98.7	72.6	91.8	95.1	543.2
		Ours	**89.5**	**97.8**	99.1	**75.1**	**93.5**	**96.2**	**551.2**	**89.3**	97.7	**99.2**	**74.0**	**92.8**	**96.2**	**549.2**
CLIP ViT-B/16	WIT-400M	NCR	91.2	98.8	99.5	76.7	93.1	95.7	555.0	87.7	97.3	98.7	68.5	86.7	90.9	529.9
		BiCro	91.1	98.4	99.6	76.3	92.6	95.6	553.6	85.8	96.4	97.6	66.8	85.5	89.9	522.1
		MSCN	92.1	99.4	**100.0**	79.6	94.8	96.9	562.8	91.8	98.9	99.7	79.1	94.5	97.0	561.0
		GSC	85.0	97.1	98.3	71.9	91.0	94.3	537.7	84.7	96.3	98.4	70.1	90.2	94.4	534.1
		ESC	91.5	98.5	99.7	77.0	93.0	95.7	555.4	87.4	97.4	98.4	68.0	86.6	91.2	529.0
		SPS	92.2	98.7	99.9	79.0	94.8	97.1	561.7	90.4	98.7	99.6	78.5	93.9	96.7	557.9
		AdaCL	91.7	98.6	98.7	77.9	93.8	96.8	557.5	91.0	97.8	99.5	75.5	89.6	93.3	546.7
		DCL	92.3	99.3	99.7	78.8	94.8	97.3	562.2	91.8	99.2	**99.9**	78.2	94.6	97.0	560.7
		Ours	**92.5**	**99.5**	99.8	**81.3**	**95.8**	**97.8**	**566.7**	**93.7**	**99.3**	99.8	**80.2**	**95.4**	**97.8**	**566.2**
CLIP ViT-L/14	WIT-400M	NCR	92.9	99.2	99.7	81.2	95.1	97.1	565.2	90.2	98.7	99.5	73.7	90.4	93.4	545.9
		BiCro	93.2	99.0	99.7	80.9	95.2	97.0	565.1	89.3	98.1	98.9	71.6	89.2	92.7	539.8
		MSCN	94.8	**99.7**	**99.9**	83.5	97.0	98.3	573.2	94.0	**99.7**	**99.9**	82.8	96.3	98.1	570.9
		GSC	88.2	98.5	99.4	75.7	91.8	94.3	547.9	84.6	97.4	98.9	67.5	88.2	92.4	529.1
		ESC	93.7	99.4	99.7	81.1	95.2	96.9	566.0	90.1	98.4	99.0	73.9	90.0	93.9	545.3
		SPS	94.3	99.3	99.8	83.4	96.6	98.1	571.5	93.1	99.3	99.7	81.6	95.7	97.9	567.2
		AdaCL	90.1	98.9	99.5	81.1	94.9	97.5	562.0	89.2	98.8	99.5	77.0	91.3	95.0	550.8
		DCL	90.8	98.8	99.6	79.9	95.5	97.9	562.5	91.2	99.4	99.7	79.9	95.5	97.6	563.3
		Ours	**95.7**	**99.7**	99.8	**84.5**	**97.1**	**98.6**	**575.4**	**95.1**	99.6	99.8	**83.1**	**96.7**	**98.5**	**572.8**
EVA-CLIP EVA02-B/16	Merged-2B	NCR	89.4	98.3	98.9	75.8	91.9	94.8	549.0	85.3	96.7	98.4	67.1	86.3	90.3	524.2
		BiCro	88.9	98.4	98.8	74.9	92.0	94.6	547.6	78.0	93.9	97.5	61.1	80.8	84.2	495.5
		MSCN	**91.5**	98.7	99.3	78.9	94.4	96.7	559.4	89.8	97.9	99.0	77.7	94.0	96.4	554.8
		GSC	83.9	96.4	98.1	69.4	88.4	92.0	528.3	82.6	96.6	98.7	69.0	89.1	93.0	529.0
		ESC	89.5	98.5	98.9	75.7	91.9	94.8	549.3	85.3	97.2	98.3	67.5	86.5	90.5	525.3
		SPS	90.6	98.5	99.1	78.1	93.8	96.3	556.3	90.4	98.3	99.2	77.2	93.1	96.1	554.3
		AdaCL	90.5	98.6	99.2	76.5	93.6	96.6	555.0	87.7	96.1	99.0	74.5	92.5	93.3	543.1
		DCL	90.5	98.7	99.3	78.6	94.2	96.8	558.1	90.8	98.6	99.4	78.4	94.5	96.7	558.4
		Ours	91.0	**99.1**	**99.4**	**80.1**	**95.0**	**97.1**	**561.8**	**91.1**	**99.0**	**99.6**	**79.6**	**94.7**	**97.2**	**561.2**

Results with Different Backbones. To assess the effectiveness of SimFlow across different backbone architectures and pretraining corpora, we evaluated it with four additional pretrained backbones. Specifically, we used CLIP ViT-B/32 pretrained on LAION-2B (Cherti et al., 2023), CLIP ViT-B/16 and ViT-L/14 pretrained on WIT-400M (Radford et al., 2021), and EVA02-B/16 from EVA-CLIP pretrained on Merged-2B (Sun et al., 2023). For NCR, BiCro, MSCN, GSC, ESC, and SPS, which have been proposed using the SGRAF backbone (Diao et al., 2021), we replaced the SGRAF cross-modal matching score with CLIP cosine similarity, following the previous practice in (Zhao et al., 2025). Table 5 reports the retrieval results on Flickr30K under 20% and 80% noise.

SimFlow achieves the highest RSum for every backbone at both noise ratios, demonstrating its effectiveness across different backbone architectures and pretraining corpora.

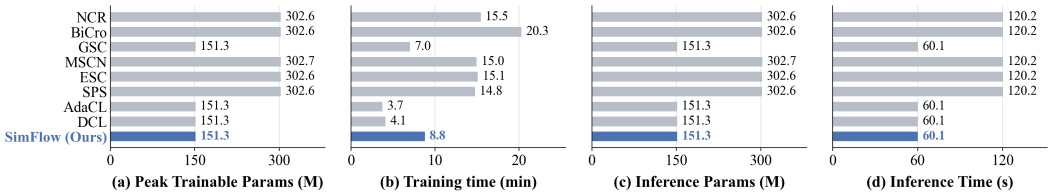


Figure 5: Computational overhead measured on Flickr30K under 20% noise. We report (a) peak trainable parameters, (b) average training time per epoch, (c) inference parameters, and (d) inference time for evaluating retrieval.

Computational Overhead. To examine the computational cost of SimFlow, we compared it with other noisy correspondence methods in terms of peak trainable parameters, average training time per epoch, inference parameters, and inference time. As shown in Figure 5, methods such as NCR, BiCro, MSCN, ESC, and SPS require an additional set of image and text encoders during training, resulting in 302.6M to 302.7M peak trainable parameters. In contrast, SimFlow has 151.3M peak trainable parameters. Its auxiliary encoders are trained separately for 2.7 minutes to obtain the reference trajectories and are discarded afterward. The velocity field contains 0.09M trainable parameters and requires 0.8 minutes of training. The average training time of SimFlow is 8.8 minutes per epoch, slightly longer than the training times of GSC, AdaCL, and DCL due to queries to the velocity field for generating target similarities for detected noisy relations. However, it remains substantially shorter than those of NCR, BiCro, MSCN, ESC, and SPS, which range from 14.8 to 20.3 minutes per epoch. Importantly, at inference, neither the auxiliary encoders nor the velocity field is required; SimFlow uses 151.3M parameters and takes 60.1 seconds for retrieval evaluation.

Table 6: Results of few-shot linear probing using encoders trained on CC140K.

Method	1-shot	2-shot	4-shot	8-shot	16-shot	32-shot
NCR (Huang et al., 2021)	40.0	47.9	55.8	68.2	73.2	76.8
BiCro (Yang et al., 2023)	40.3	47.9	55.6	68.0	73.1	76.6
GSC (Zhao et al., 2024)	29.0	34.9	41.8	62.8	69.4	74.0
ESC (Yang et al., 2024)	39.9	47.8	55.7	68.1	72.8	77.1
SPS (Xie et al., 2025)	42.1	49.4	56.2	68.7	73.6	77.6
MSCN (Han et al., 2023)	40.9	48.6	56.0	68.9	74.4	77.9
AdaCL (Pan et al., 2025)	41.1	47.7	55.1	69.2	74.6	78.3
DCL (Zhong et al., 2026)	42.5	49.5	57.1	70.7	75.5	79.2
SimFlow (Ours)	**43.9**	**51.1**	**58.2**	**71.0**	**76.2**	**79.4**

Results on Linear Probing. To further evaluate the transferability of the representations learned from noisy image-text data, we conducted few-shot linear probing on the ten classification datasets used for zero-shot evaluation in Table 3, using the encoders trained on CC140K and following the protocol in (Radford et al., 2021). For each classification dataset, we froze the image encoder and trained a dataset-specific linear classification layer using 1, 2, 4, 8, 16, and 32 samples per class. Table 6 reports the average accuracy across the ten datasets. Overall, SimFlow achieves the highest average accuracy across different numbers of training samples per class. Although dataset-specific linear classifiers can compensate for differences in the learned representations, SimFlow consistently achieves the highest accuracy from 1-shot to 32-shot. This suggests that the proposed method learns rich semantic representations despite substantial noisy correspondence in the training data, supporting strong generalization to unseen domains and categories.

Visualization of Reference Relations. To qualitatively examine whether the constructed reference relations reflect the semantic correspondence of detected noisy relations, we additionally visualize examples of false-positive and false-negative relations with their corresponding references in Figure 6. The reference relations show patterns of correspondence between image and text similar to those

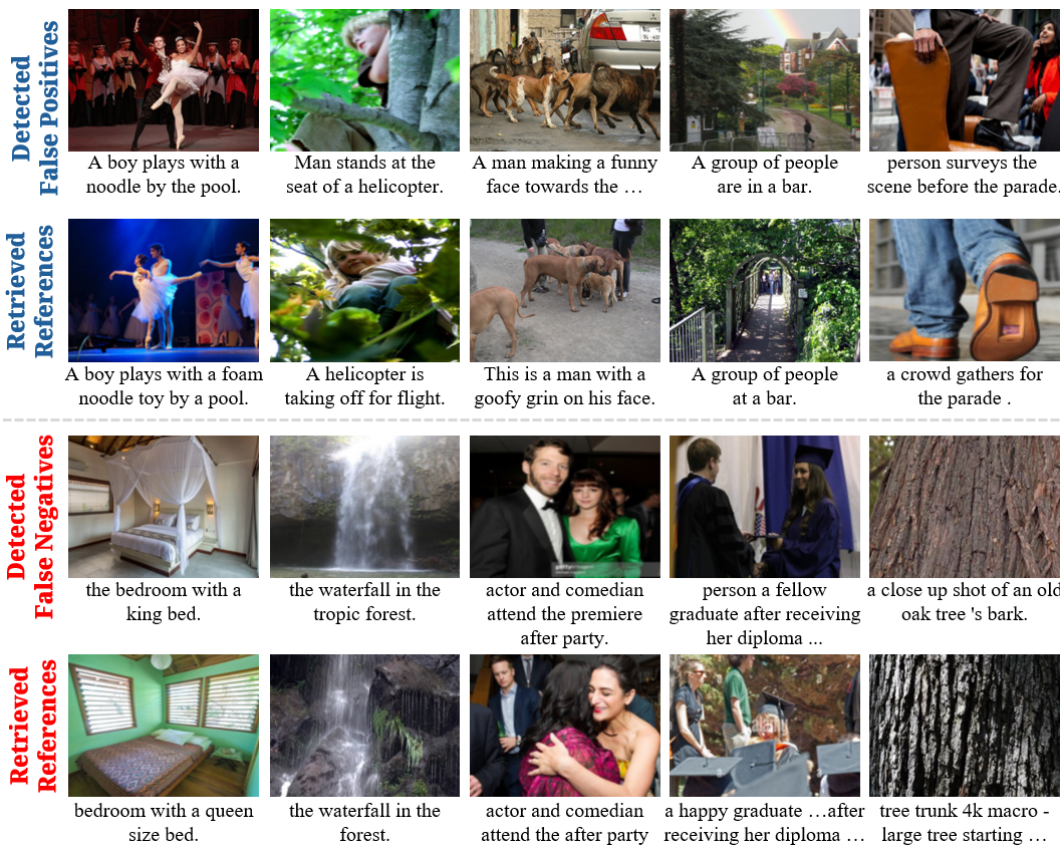


Figure 6: Visualization of detected false-positive and false-negative relations and their corresponding reference relations.

of the detected noisy relations, with semantic inconsistency and semantic correspondence for false positives and false negatives, respectively. These examples show that reference relations constructed from detected clean samples can serve as semantic proxies for the noisy relations when learning their correction trajectories.

Use of LLMs. In accordance with the ICLR 2027 policy on AI assistance disclosure, generative AI tools, including ChatGPT, were used exclusively to edit and refine the prose of small portions of human-written text in the manuscript.