ALIGNING WITHIN PAIR-SPECIFIC SUBSPACES FOR NOISY CORRESPONDENCE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Noisy correspondence introduces semantically mismatched images and texts as positive training pairs, creating incorrect attraction between their representations and, when true pairs are replaced, removing correct associations available for learning. Existing methods mitigate erroneous supervision through reliability estimation, robust optimization, and correspondence rectification. However, pair-level reweighting primarily adjusts the learning strength of pairs with different confidence levels, without directly identifying the representation directions suitable for auxiliary alignment of a particular pair. Auxiliary constraints derived from similarity relations between samples may also propagate correspondence errors to other pairs and interfere with learning reliable associations. We propose *Aligning within Pair-Specific Subspaces* (APS), which extracts multi-view consistency evidence from each observed pair to determine the space in which auxiliary alignment is applied. Specifically, a frozen teacher model obtained through confidence-weighted warm-up evaluates the consistency of image-text view pairs. Joint representations of selected view pairs then define a pair-specific subspace. Correspondence confidence jointly weights the global contrastive and auxiliary subspace objectives, strengthening image-text agreement for more reliable pairs within their subspaces while reducing the influence of auxiliary constraints from uncertain pairs. Extensive experiments on Flickr30K, MS-COCO, and CC120K demonstrate the advantages of APS over state-of-the-art baselines for cross-modal retrieval under synthetic and real-world correspondence noise.

1 INTRODUCTION

Image-text retrieval aims to establish correspondences between visual content and language descriptions and is a fundamental task in multimodal representation learning. Vision-language pretrained models, exemplified by CLIP, acquire transferable representations through contrastive learning on large-scale image-text data (Radford et al., 2021). However, the co-occurrence of an image and a text description in training data does not always indicate a semantic correspondence. When unrelated images and texts are labeled as positive pairs, they form *noisy correspondence* (NC) (Huang et al., 2021; Qin et al., 2022). Such erroneous supervision encourages representations of mismatched content to move closer together. When incorrect pairs replace true pairs, the correct associations that would otherwise support learning are also lost. Noisy correspondence learning (NCL) therefore seeks to learn reliable cross-modal representations in the presence of incorrect supervision.

Existing methods address unreliable image-text pairs through two strategies that can be combined. **Reliability modeling and robust optimization** estimate pair quality from losses, prediction confidence, or uncertainty, and control the influence of incorrect supervision through sample selection, loss reweighting, adaptive margins, or robust objectives (Huang et al., 2021; Qin et al., 2022; Liu et al., 2024; Zhang et al., 2024; Qin et al., 2023; Wang et al., 2025; Zhong et al., 2026b;a; Dang et al., 2024; Qin et al., 2025). **Correspondence rectification and supervision reconstruction** use intra-modal relations, neighboring anchors, optimal transport, or generated text to correct soft correspondence labels or reconstruct training pairs (Yang et al., 2023; Han et al., 2023; 2024; Zheng et al., 2026; Yang et al., 2026; Yao et al., 2026; Shi et al., 2026; Li et al., 2025; Han & Cao, 2025). The former adjusts the training contributions of pairs with different confidence levels. The latter improves image-text correspondences through soft-label correction, rematching, or text replacement. These methods alleviate the interference of noisy pairs, but reducing the weights of mismatched pairs

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.

does not recover correct associations that have already been lost. Meanwhile, inaccurate soft-label correction or pair reconstruction can introduce further erroneous supervision.

When correct associations are corrupted, strengthening learning from reliable pairs offers a complementary approach. PC^2 introduces pseudo-classification as an auxiliary task, while SPS, GSC, ESC, and ReCon improve representations through stable feature spaces, geometric structure, or similarity consistency (Duan et al., 2024; Xie et al., 2025; Zhao et al., 2024; Yang et al., 2024; Zha et al., 2025). A central question is how to obtain reliable auxiliary supervision for this purpose. Intra-modal and cross-modal similarity relations provide useful information, but similar object content does not guarantee accurate image-text correspondence. Descriptions containing the same object words can still express different attributes and object relations, as highlighted by work on compositional understanding and hard negatives (Yuksekgonul et al., 2023; Doveh et al., 2023). Modality-specific information and the quality of neighboring anchors can further affect the accuracy of relation estimates (Zheng et al., 2026; Yang et al., 2026; Dang et al., 2025; Liu et al., 2026). We therefore turn to the current pair itself and seek auxiliary supervision from its multiple views, reducing dependence on correspondences involving other pairs. Image augmentation and text perturbation generate views that retain different parts of the original content (Chen et al., 2020a; Jiang & Ye, 2023). Related work based on Brownian bridges has also used augmented image features to construct cross-modal geometric constraints for individual image-text pairs (Jiang et al., 2023). Motivated by these observations, we select image-text view pairs with high cross-modal consistency, extract their joint representation directions, and use these directions to construct a pair-specific subspace. The resulting subspace restricts auxiliary alignment to representation directions supported by that pair.

We propose **APS**, which consists of two connected components. **Pair-specific subspace construction** first generates multiple views through random image augmentation and text token deletion. A frozen teacher model obtained through confidence-weighted warm-up evaluates their cross-modal consistency. From the K^2 possible image-text view pairs, we select the m highest-scoring combinations, apply singular value decomposition (SVD) to their joint image-text representations, and retain the leading r right singular vectors to define the pair-specific subspace. **Confidence-weighted auxiliary subspace alignment** then constrains the student's image and text representations within the subspace of each pair. The same correspondence confidence weights both the auxiliary objective and the global contrastive objective, allowing estimated pair quality to regulate the strength of both constraints. At retrieval time, the student image and text encoders extract global representations, and cosine similarity is used for bidirectional retrieval. Our contributions are summarized as follows:

- **We propose a subspace construction method based on multi-view evidence within individual pairs.** By selecting image-text view pairs with high consistency and extracting their joint representation directions, we construct an auxiliary alignment space specific to each observed image-text pair.
- **We propose confidence-weighted auxiliary subspace alignment.** Combining correspondence confidence with pair-specific subspaces jointly controls the strength and the representation directions of auxiliary supervision, strengthening learning from reliable associations while reducing the influence of noise.
- **We conduct extensive experiments on image-text benchmarks with synthetic and real-world correspondence noise.** APS outperforms existing robust methods in most evaluated settings, demonstrating its effectiveness in mitigating noise interference.

2 Related Work

2.1 Vision-Language Representation Learning and Image-Text Retrieval

Image-text retrieval enables bidirectional search by learning semantic correspondences between visual content and language descriptions. Early methods mainly investigated global representations and local cross-modal interactions. VSE++ improves the discriminative ability of global representations through ranking with hard negatives (Faghri et al., 2018). SCAN and SGRAF model fine-grained matching through cross-attention and graph reasoning, respectively (Lee et al., 2018; Diao et al., 2021). Vision-language pretraining further exploits large-scale image-text data to learn transferable cross-modal representations. CLIP adopts contrastive learning with dual encoders (Radford et al., 2021), whereas UNITER and ALBEF combine multiple pretraining objectives or momentum

distillation (Chen et al., 2020b; Li et al., 2021). IRRA uses masked text reconstruction to learn implicit cross-modal relations (Jiang & Ye, 2023). Building on pretrained dual encoders, APS constructs auxiliary alignment constraints regulated by correspondence confidence.

2.2 Noisy Correspondence Learning

NCL studies robust representation learning when observed image-text pairs do not reflect true semantic correspondences. Related problems also arise in text-to-image person retrieval, video-language learning, image captioning, and multi-view clustering (Qin et al., 2024; Lin et al., 2024; Kang et al., 2023; Zhou et al., 2026), as well as language-image pretraining and CLIP adaptation (Huang et al., 2023; Han et al., 2025). Existing approaches mainly address incorrect associations through reliability modeling and robust optimization, or through correspondence rectification and supervision reconstruction. NCR, DECL, and related methods suppress noise using pair-quality estimation, sample weighting, or uncertainty modeling (Huang et al., 2021; Qin et al., 2022; Liu et al., 2024; Zha et al., 2024; Qin et al., 2025). CRCL and DCL reduce the influence of incorrect positive associations through complementary learning (Qin et al., 2023; Zhong et al., 2026a). BiCro, L2RM, REPAIR, BiCro++, CCDCL, and PTRL improve supervision using similarity relations, optimal transport, ranking, clustering, consistency, or pseudo-text (Yang et al., 2023; Han et al., 2024; Zheng et al., 2026; Yang et al., 2026; Yao et al., 2026; Shi et al., 2026). Generative vision-language models such as BLIP-2 provide a basis for obtaining textual supervision (Li et al., 2023). PC^2, GSC, ESC, ReCon, and DisNCL further incorporate auxiliary tasks, geometric constraints, or shared-information learning (Duan et al., 2024; Zhao et al., 2024; Yang et al., 2024; Zha et al., 2025; Dang et al., 2025). Particularly relevant to APS, MGCS observes that image-text modality-gap directions are sample-level, compact, and data-dependent (Wang et al., 2026). It uses the difference direction $v - t$ and priors derived from neighboring gap directions to correct similarity and improve noise partitioning. APS constructs its auxiliary projection space from the joint directions $\mathrm{norm}(v + t)$ of multiple views of the same observed pair, and constrains the student's representations within that space. Their respective uses of modality differences and joint representations serve distinct objectives.

2.3 Multi-View Consistency and Subspace Learning

Multi-view and subspace methods improve learning by identifying common information in representations. CMC uses contrastive learning across views, while CyCLIP imposes similarity consistency within and across modalities (Tian et al., 2020; Goel et al., 2022). Brownian-bridge modeling constrains paths of augmented features between paired image and text representations (Jiang et al., 2023). DCCA and MISA learn correlated or shared representations across modalities (Andrew et al., 2013; Hazarika et al., 2020). SSP already uses feature selection, SVD, and projection to construct a dataset-shared visual subspace and class-specific language subspaces for few-shot classification (Zhu et al., 2024). SimLAP, called Hydra in an earlier version, uses learnable feature gates conditioned on class pairs to capture the common information in arbitrary visual positive pairs (Wu et al., 2025). APS shares with these methods the use of subspaces to select information for alignment. Its distinction lies in constructing a fixed subspace from the multi-view evidence of each observed image-text pair itself and combining this space with correspondence confidence to form an auxiliary constraint during training. Retrieval uses the student's global image and text representations.

3 Method

3.1 Problem Formulation

Consider a training dataset $\mathcal{D} = \{(I_i, T_i)\}_{i=1}^{N}$ containing N observed image-text pairs, where I_i and T_i denote the image and text of the ith pair, respectively. Let $y_i \in \{0, 1\}$ be its true correspondence label: $y_i = 1$ indicates that the image and text are semantically matched, whereas $y_i = 0$ indicates noisy correspondence. The labels y_i are unavailable during training, and observed images and texts with the same index are treated as positive pairs (Huang et al., 2021; Yang et al., 2023; Shi et al., 2026). NCL aims to learn an image encoder f_θ and a text encoder g_θ from these noisy pairs for

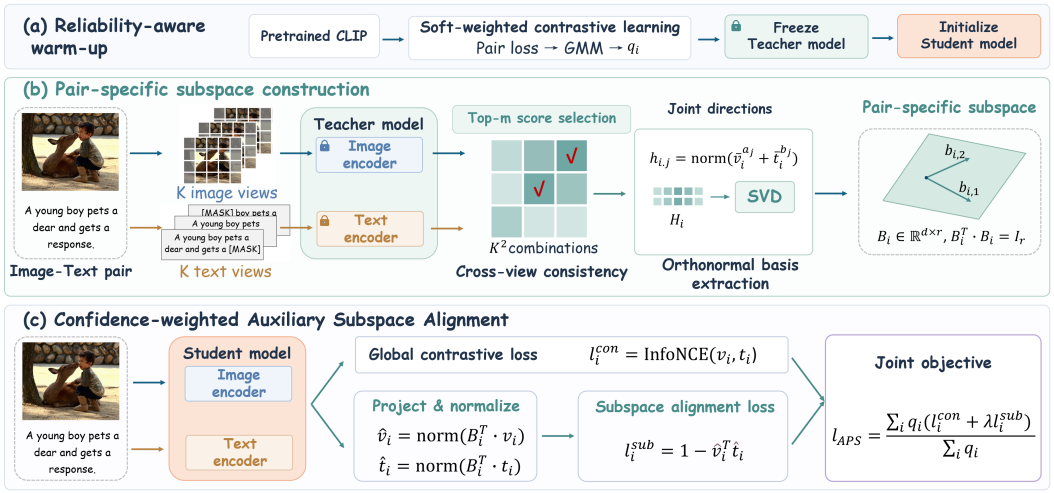


Figure 1: Overview of APS. (a) Confidence-weighted warm-up produces a frozen teacher and initializes the student. (b) The teacher scores K^2 view pairs; the top m joint representations define a subspace through SVD and the leading r right singular vectors. (c) The student learns with global contrastive and auxiliary subspace objectives. `[MASK]` marks content removed by token deletion. The illustrated q_i weighting corresponds to $\alpha = 0$; Eq. (8) gives the general weighting.

reliable cross-modal retrieval. We use cosine similarity to measure image-text correspondence:

$$s_{ij} = s(I_i, T_j) = \frac{f_\theta(I_i)^\top g_\theta(T_j)}{\|f_\theta(I_i)\|_2 \, \|g_\theta(T_j)\|_2}. \tag{1}$$

For a mini-batch $\mathcal{B}$, we use the bidirectional contrastive loss (Radford et al., 2021):

$$\ell_i^{\text{con}} = -\frac{1}{2}\left[\log \frac{\exp(s_{ii}/\tau)}{\sum_{j\in\mathcal{B}} \exp(s_{ij}/\tau)} + \log \frac{\exp(s_{ii}/\tau)}{\sum_{j\in\mathcal{B}} \exp(s_{ji}/\tau)}\right], \tag{2}$$

where τ is the temperature parameter. The two terms correspond to image-to-text and text-to-image matching, respectively.

3.2 Correspondence Confidence and Teacher Warm-Up

Estimating correspondence confidence. To reduce the interference of noisy pairs during training, we estimate correspondence confidence from the contrastive loss of each observed pair (Huang et al., 2021; Zheng et al., 2026; Yao et al., 2026). Specifically, we compute per-pair losses using the bidirectional contrastive form in Eq. (2), apply min–max normalization, and fit the resulting loss distribution with a two-component Gaussian mixture model (GMM). Let $\widetilde{\ell}_i$ denote the normalized loss. We associate the component with the lower mean, denoted by $c^\star$, with relatively reliable pairs, and use its posterior probability as the correspondence confidence:

$$q_i = p(c^\star \mid \widetilde{\ell}_i) = \frac{\pi_{c^\star}\mathcal{N}(\widetilde{\ell}_i; \mu_{c^\star}, \sigma_{c^\star}^2)}{\sum_{c=1}^{2} \pi_c \mathcal{N}(\widetilde{\ell}_i; \mu_c, \sigma_c^2)}, \qquad c^\star = \arg\min_c \mu_c, \tag{3}$$

where π_c, μ_c, and σ_c^2 are the mixture proportion, mean, and variance of component c. The value $q_i \in [0, 1]$ expresses the model's estimated confidence that the image and text in the observed pair correspond semantically.

Teacher warm-up. We warm up the model to obtain a cross-modal consistency estimator adapted to the target data. Starting from pretrained image and text encoders, we use correspondence confidence to softly weight the contrastive objective, reducing the influence of uncertain pairs:

$$\mathcal{L}_{\text{warm}} = \frac{\sum_{i\in\mathcal{B}} q_i \ell_i^{\text{con}}}{\sum_{i\in\mathcal{B}} q_i}. \tag{4}$$

Correspondence confidence is estimated at model initialization and updated after each warm-up epoch. Once warm-up is complete, we freeze the resulting encoders as the teacher model $(f_{\bar{\theta}}, g_{\bar{\theta}})$ and fix the final confidence values q_i. The student is initialized with the same encoder parameters and continues learning. The frozen teacher subsequently extracts multi-view representations, evaluates cross-modal consistency, and constructs pair-specific subspaces. It therefore provides a fixed reference for the auxiliary alignment performed during student training.

3.3 Pair-Specific Subspaces Guided by Multi-View Consistency

Constructing cross-modal views. To obtain evidence that retains different parts of the same pair, we perturb the image and text separately, and use cross-modal consistency to select view pairs suitable for constructing the auxiliary alignment space. On the image side, we apply common augmentations, including random cropping, horizontal flipping, color jitter, grayscale conversion, and blur (Chen et al., 2020a). On the text side, body tokens are randomly selected with probability p_t and masked through deletion. The relative order of the remaining tokens and the sequence boundary tokens are preserved, producing multiple text views with different retained content. Text masking is also used in the masked language modeling objectives of IRRA, ALBEF, and UNITER to support cross-modal relation and representation learning (Jiang & Ye, 2023; Li et al., 2021; Chen et al., 2020b). In APS, token deletion serves to generate augmented views; it does not introduce a masked-token prediction objective.

For the ith pair, we generate K image views $\xi_i^a(I_i)$ and K text views $\delta_i^b(T_i)$. The frozen teacher model extracts the global representations $\bar{v}_i^a = \mathrm{norm}(f_{\bar{\theta}}(\xi_i^a(I_i)))$ and $\bar{t}_i^b = \mathrm{norm}(g_{\bar{\theta}}(\delta_i^b(T_i)))$, where $a, b = 1, \ldots, K$ and $\mathrm{norm}(z) = z/\|z\|_2$. Both types of representation lie in $\mathbb{R}^d$. Since perturbations may remove content important for correspondence, we further select image-text view pairs according to the teacher's cross-modal consistency scores.

Subspace construction. We compute the similarity of all K^2 cross-modal view pairs as $C_i^{ab} = (\bar{v}_i^a)^\top \bar{t}_i^b$ and select the m combinations with the highest scores. Their index set is denoted by $\Omega_i = \{(a_j, b_j)\}_{j=1}^m$. Here, m controls the number of view pairs used to construct the subspace, while r controls its retained dimension, with $2 \le r \le \min(m, d)$ and $m \le K^2$. For each selected combination, we combine its image and text representations to obtain a joint representation:

$$h_{i,j} = \mathrm{norm}(\bar{v}_i^{a_j} + \bar{t}_i^{b_j}), \qquad j = 1, \ldots, m. \tag{5}$$

Combining the two modalities allows both visual content and language descriptions to contribute to subspace construction. We stack the selected joint representations into $H_i = [h_{i,1}, \ldots, h_{i,m}]^\top \in \mathbb{R}^{m \times d}$ and perform a thin SVD. The leading r right singular vectors, ordered by descending singular value, form the subspace basis:

$$H_i = U_i \Sigma_i V_i^\top, \qquad B_i = V_i[:, 1:r] \in \mathbb{R}^{d \times r}. \tag{6}$$

The diagonal entries of Σ_i are sorted in descending order. The columns of B_i are orthonormal, satisfying $B_i^\top B_i = \mathbf{I}_r$, where $\mathbf{I}_r$ is the r-dimensional identity matrix. These columns span the pair-specific subspace $\mathcal{S}_i = \mathrm{span}(B_i)$. View-pair selection determines which evidence from the current pair participates in basis construction, while the retained singular directions determine the space in which auxiliary alignment is applied. When $m > r$, the subspace retains the dominant directions of the selected joint representations. When $m = r$ and those representations are linearly independent, SVD provides an orthonormal basis for their span. The teacher model and the resulting B_i remain fixed throughout student learning. We use $r \ge 2$ because unit-normalized one-dimensional projections make the auxiliary cosine objective degenerate; Appendix D states the assumptions and proves this degeneracy.

3.4 Confidence-Weighted Auxiliary Subspace Alignment

Positive alignment within the subspace. For pair i, the student extracts the global representations $v_i = \mathrm{norm}(f_\theta(\xi_i(I_i)))$ from an augmented image and $t_i = \mathrm{norm}(g_\theta(T_i))$ from the full text. The global contrastive and auxiliary objectives reuse this same pair of representations. Using the fixed basis B_i, we obtain unit-length subspace representations $\widehat{v}_i = \mathrm{norm}(B_i^\top v_i)$ and $\widehat{t}_i = \mathrm{norm}(B_i^\top t_i)$.

We define the auxiliary alignment loss as:

$$\ell_i^{\text{sub}} = 1 - \widehat{v}_i^\top \widehat{t}_i. \tag{7}$$

This objective encourages image-text agreement within the subspace of the current pair. It complements global contrastive learning with a positive alignment constraint restricted to the selected representation directions. The loss acts on the student representations through the fixed projection, and its gradients are backpropagated to the student encoders.

Joint confidence weighting. Each pair-specific subspace is constructed from the joint representations of that pair, so its reliability is closely related to the quality of the original image-text correspondence. A low-quality pair may provide both incorrect global supervision and an unreliable auxiliary constraint through its subspace. We therefore use the correspondence confidence q_i estimated for the original pair by the teacher after confidence-weighted warm-up to regulate both learning objectives. This jointly reduces the global and auxiliary loss contributions of low-confidence pairs. We define the student-stage pair weight as $w_i = \alpha + (1-\alpha)q_i$, where $\alpha \in [0, 1)$ controls the minimum learning weight. With $\alpha = 0$, the weight equals the GMM posterior confidence. We combine the two weighted losses as:

$$\mathcal{L}_{\text{APS}} = \frac{\sum_{i\in\mathcal{B}} w_i \left(\ell_i^{\text{con}} + \lambda \ell_i^{\text{sub}}\right)}{\sum_{i\in\mathcal{B}} w_i}, \tag{8}$$

where λ controls the strength of auxiliary alignment. The confidence q_i regulates each pair's contribution, while B_i determines the representation subspace in which its auxiliary constraint acts. Together, they determine how multi-view evidence from the current pair is used for learning. Continuous weighting avoids excluding uncertain pairs solely according to a hard threshold, retaining a learning contribution from information that may still be useful for alignment.

During training, the global contrastive and auxiliary subspace objectives share the student's image-text representations, while the frozen teacher is used to construct the auxiliary spaces. At retrieval time, the student image and text encoders extract global representations, and retrieval uses the global cosine similarity in Eq. (1).

4 Experiments

4.1 Datasets and Metrics

We evaluate APS on Flickr30K (Young et al., 2014), MS-COCO (Lin et al., 2014), and CC120K (Zhang et al., 2024; Sharma et al., 2018). Following the synthetic-noise protocol of NPC (Zhang et al., 2024), we randomly shuffle a proportion of image-text correspondences in the Flickr30K and MS-COCO training sets to construct correspondence noise at rates of 20%, 40%, and 60%. Validation and test pairs remain unchanged. CC120K is a subset of Conceptual Captions (Sharma et al., 2018) introduced by NPC (Zhang et al., 2024), and is used to evaluate retrieval under real-world correspondence noise in web-collected data.

We report R@1, R@5, and R@10 for both image-to-text (I2T) and text-to-image (T2I) retrieval, and use their sum, RSum, to measure overall retrieval performance. Results are displayed to one decimal place. For APS, RSum is calculated by summing the six recall values displayed in the table, while its sample standard deviation is calculated from the unrounded RSum values of individual runs. MS-COCO 1K results are averaged over five test folds, and the 5K results are provided in Appendix C. APS main results, associated ablations, and hyperparameter sensitivity experiments use three independent training runs. Exceptions in the number of runs are specified in the corresponding table notes. Dataset splits and implementation details are given in Appendix A.

4.2 Comparison with State-of-the-Art Methods

We compare APS with NCR (Huang et al., 2021), DECL (Qin et al., 2022), GSC (Zhao et al., 2024), L2RM (Han et al., 2024), ReCon (Zha et al., 2025), CLIP (Radford et al., 2021), NPC (Zhang et al., 2024), DSDMR (Shi et al., 2024), HaNa (Zhong et al., 2026b), and DCL (Zhong et al., 2026a). Results for the comparison methods are taken from their published reports, retaining their original encoders and training settings. APS results are means over three runs. Comparisons conducted under our training configuration are analyzed in Section 4.3.

Table 1: Retrieval results on Flickr30K and MS-COCO 1K under synthetic correspondence noise. Comparison methods use their published results; APS reports means over three runs. Bold and underlined values indicate the best and second-best results for each dataset, noise rate, and metric. Equal displayed values receive the same formatting.

Noise	Method	Flickr30K							MS-COCO 1K						
		Image-to-text			Text-to-image			RSum	Image-to-text			Text-to-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	
20%	NCR	73.5	93.2	96.6	56.9	82.4	88.5	491.1	77.7	95.5	98.2	62.5	89.3	95.3	518.5
	DECL	77.5	93.8	97.0	56.1	81.8	88.5	494.7	77.5	95.9	98.4	61.7	89.3	95.4	518.2
	GSC	78.3	94.6	97.8	60.1	84.5	90.5	505.8	79.5	96.4	**98.9**	64.4	90.6	95.9	525.7
	L2RM	77.9	95.2	97.8	59.8	83.6	89.5	503.8	80.2	96.3	98.5	64.2	90.1	95.4	524.7
	ReCon	80.3	95.3	97.8	61.6	85.5	91.3	511.8	80.9	96.6	98.8	65.2	91.0	96.0	528.5
	CLIP	82.3	95.5	98.3	66.0	88.5	93.5	524.1	75.0	93.1	97.2	58.7	86.1	97.2	507.3
	NPC	87.3	97.5	98.8	72.9	92.1	95.8	544.4	79.9	95.9	98.4	66.3	90.8	**98.4**	529.7
	DSDMR	85.1	97.0	99.2	69.7	90.3	94.8	536.1	80.5	95.8	98.4	65.8	90.9	96.2	527.6
	HaNa	89.0	97.9	99.3	75.1	93.7	96.8	551.8	81.7	96.6	98.8	**69.0**	92.2	96.6	534.9
	DCL	89.0	97.6	99.1	74.2	92.2	95.4	547.5	82.2	**96.8**	**98.9**	66.7	92.2	96.8	533.6
	APS	**89.2**	**98.4**	**99.6**	**76.2**	**94.1**	**97.1**	**554.6**	**82.3**	96.7	**98.9**	68.8	**92.4**	97.1	**536.2**
40%	NCR	68.1	89.6	94.8	51.4	78.4	84.8	467.1	74.7	94.6	98.0	59.6	88.1	94.7	509.7
	DECL	72.7	92.3	95.4	53.4	79.4	86.4	479.6	75.6	95.5	98.3	59.5	88.3	94.8	512.0
	GSC	76.5	94.1	97.6	57.5	82.7	88.9	497.3	78.2	95.9	98.2	62.5	89.7	95.4	519.9
	L2RM	75.8	93.2	96.9	56.3	81.0	87.3	490.5	77.5	95.8	98.4	62.0	89.1	94.9	517.7
	ReCon	79.4	94.3	97.6	59.9	83.9	90.1	505.2	79.9	96.2	98.6	63.5	90.5	95.9	524.6
	CLIP	76.2	93.3	96.5	59.4	85.0	90.9	501.3	70.7	91.7	96.2	54.7	83.4	96.2	492.9
	NPC	85.6	97.5	98.4	71.3	91.3	95.3	539.4	79.4	95.1	98.3	65.0	90.1	**98.3**	526.2
	DSDMR	85.2	97.0	99.0	69.6	90.2	94.9	535.9	79.9	95.9	98.3	64.8	90.3	95.9	525.1
	HaNa	88.5	98.3	99.3	75.1	93.6	96.8	551.6	81.9	**96.7**	98.9	**68.2**	91.6	96.5	533.8
	DCL	**89.8**	97.9	99.0	74.9	92.4	95.6	549.6	**82.2**	96.3	**99.1**	67.1	91.4	96.6	532.7
	APS	88.8	**98.5**	**99.6**	**75.4**	**93.9**	**97.0**	**553.2**	82.0	96.4	98.7	**68.2**	**92.0**	96.9	**534.2**
60%	NCR	13.9	37.7	50.5	11.0	30.1	41.4	184.6	0.1	0.3	0.4	0.1	0.5	1.0	2.4
	DECL	65.2	88.4	94.0	46.8	74.0	82.2	450.6	73.0	94.2	97.9	57.0	86.6	93.8	502.5
	GSC	70.8	91.1	95.9	53.6	79.8	86.8	478.0	75.6	95.1	98.0	60.0	88.3	94.6	511.6
	L2RM	70.0	90.8	95.4	51.3	76.4	83.7	467.6	75.4	94.7	97.9	59.2	87.4	93.8	508.4
	ReCon	74.3	93.6	96.6	55.7	81.6	88.1	489.9	77.2	95.9	98.4	61.8	89.3	95.2	517.8
	CLIP	66.3	87.3	93.0	52.1	78.8	87.4	464.9	67.0	88.8	95.0	49.7	79.6	95.0	475.1
	NPC	83.0	95.9	98.6	68.1	89.6	94.2	529.4	78.2	94.4	97.7	63.1	89.0	**97.7**	520.1
	DSDMR	85.8	96.9	99.1	69.5	90.1	94.7	536.1	78.9	95.2	98.2	63.5	89.5	95.7	521.0
	HaNa	87.1	98.0	99.4	**74.6**	93.2	96.4	548.7	80.8	96.1	98.6	67.0	91.1	96.1	529.7
	DCL	**89.3**	98.0	99.0	73.2	92.1	95.6	547.2	**81.5**	**96.3**	**98.9**	66.9	91.3	96.4	**531.3**
	APS	87.9	**98.1**	**99.5**	74.1	**93.5**	**96.6**	**549.7**	81.1	96.1	98.6	**67.1**	**91.4**	96.6	530.9

Synthetic correspondence noise. Table 1 summarizes the results on Flickr30K and MS-COCO 1K. On Flickr30K, APS obtains RSum of 554.6, 553.2, and 549.7 at noise rates of 20%, 40%, and 60%, respectively. These results exceed the strongest listed comparison method at each noise level by 2.8, 1.6, and 1.0. On MS-COCO 1K, APS achieves RSum of 536.2 and 534.2 under 20% and 40% noise, improving on the strongest listed comparison method by 1.3 and 0.4, respectively. Under 60% noise, APS reaches 530.9, which is below DCL's 531.3 but above the results of the other listed methods. These comparisons demonstrate competitive retrieval performance across different noise conditions, with the advantage of APS varying across datasets and noise levels.

4.3 Ablation Study

We conduct ablation experiments on Flickr30K at all three noise rates. Table 3 organizes five configurations according to the use of four components: GMM-based reliability estimation, continuous confidence weighting, auxiliary subspace alignment, and auxiliary global-space alignment. At each noise level, the rows correspond to CLIP fine-tuning under our configuration, a GMM baseline, subspace alignment with student-stage weighting removed, confidence-weighted global-space alignment, and full APS. The GMM column indicates the use of GMM-estimated pair reliability. The

Table 2: Retrieval on CC120K with real-world correspondence noise. APS reports means over three runs. Formatting follows Table 1.

Method	Image-to-text			Text-to-image			RSum
	R@1	R@5	R@10	R@1	R@5	R@10	
CLIP	68.8	87.0	92.9	67.8	86.4	90.9	493.8
NPC	71.1	92.0	**96.2**	73.0	90.5	94.8	517.6
HaNa	**74.4**	92.2	95.8	72.5	**92.0**	94.6	521.5
APS	72.7	**93.2**	96.1	**73.4**	91.7	**95.6**	**522.7**

Real-world noise. Table 2 reports the CC120K results. APS achieves an RSum of 522.7, exceeding CLIP, NPC, and HaNa by 28.9, 5.1, and 1.2. While different methods attain the best individual recall values, APS achieves the highest overall RSum. These results demonstrate its effectiveness under real-world web correspondence noise.

Table 3: Component ablation on Flickr30K. CW denotes continuous confidence weighting during student learning. Except for the GMM baseline, all rows report three-run means. Bold marks the best result at each noise level.

Noise	Components				Image-to-text			Text-to-image			RSum
	GMM	CW	Auxiliary alignment Subspace	Global	R@1	R@5	R@10	R@1	R@5	R@10	
20%					86.6	97.2	99.2	71.6	91.9	95.5	542.0
	✓	✓			87.9	**98.4**	99.4	75.4	93.9	97.0	552.0
	✓		✓		86.5	97.7	99.2	72.7	92.2	95.9	544.2
	✓	✓		✓	86.2	97.7	99.5	73.6	92.9	96.6	546.5
	✓	✓	✓		**89.2**	**98.4**	**99.6**	**76.2**	**94.1**	**97.1**	**554.6**
40%					86.0	96.9	98.8	70.3	90.7	95.0	537.7
	✓	✓			87.7	98.3	99.4	74.6	93.6	96.9	550.5
	✓		✓		86.8	97.1	99.1	71.2	91.3	95.2	540.7
	✓	✓		✓	85.9	97.9	99.3	73.3	92.7	96.6	545.7
	✓	✓	✓		**88.8**	**98.5**	**99.6**	**75.4**	**93.9**	**97.0**	**553.2**
60%					82.5	96.5	98.3	66.8	88.2	93.1	525.4
	✓	✓			86.2	97.6	99.3	73.6	93.0	**96.6**	546.3
	✓		✓		85.0	96.6	98.5	68.2	89.2	93.8	531.3
	✓	✓		✓	84.9	97.3	99.2	72.6	92.4	96.4	542.8
	✓	✓	✓		**87.9**	**98.1**	**99.5**	**74.1**	**93.5**	**96.6**	**549.7**

CW column specifically indicates continuous confidence weighting during student training. Consequently, the APS variant without CW still retains the warmed-up teacher and pair-specific subspaces. Complete configurations, standard deviations, and extended ablations are provided in Appendix B.

Reliability estimation and confidence weighting. The GMM-soft baseline (GMM+CW) achieves RSum of 552.0, 550.5, and 546.3 at the three noise levels, outperforming ordinary CLIP fine-tuning in each case. Among the APS variants that share the same teacher warm-up and student training budget, removing continuous weighting during student learning reduces RSum by 10.4, 12.5, and 18.4 relative to full APS. These results indicate that a warmed-up teacher and pair-specific subspaces alone are insufficient to control the influence of mismatched pairs. Continuing to regulate each pair's contribution during student learning is important for effective auxiliary alignment.

The space used for auxiliary alignment. When the same continuous confidence weights and auxiliary loss coefficient are retained, replacing subspace alignment with a cosine constraint in the global space decreases RSum by 8.1, 7.5, and 6.9 at the three noise levels. Applying the auxiliary constraint within a pair-specific subspace therefore yields better retrieval performance in these experiments than directly reinforcing global image-text similarity. This comparison supports using evidence from the pair to determine the representation directions on which auxiliary alignment acts. Appendix B further compares single-modality subspace construction, shuffled subspaces, Top-2, Random-2, and Bottom-2 view-pair selection, as well as batch-shared bases, to examine the source and scope of the evidence used for the auxiliary space.

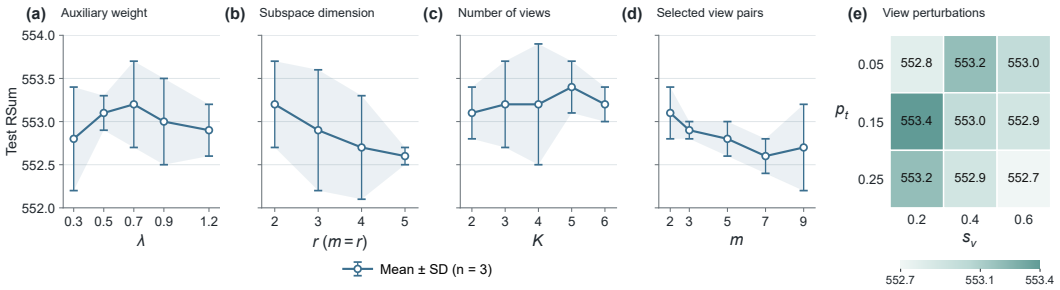


Figure 2: Hyperparameter sensitivity on Flickr30K with 40% noise: (a) auxiliary weight λ; (b) subspace dimension r, with $m = r$; (c) views per modality K, with $m = r = 2$; (d) selected view pairs m, with $K = 3, r = 2$; (e) token deletion probability p_t and minimum crop area s_v. Panels (a)–(d) show three-run mean test RSum ± 1 sample standard deviation; panel (e) shows means. Panel (d) uses an independent supplementary experiment.

4.4 Hyperparameter Sensitivity

Figure 2 reports three-run test results on Flickr30K with 40% correspondence noise. Unspecified parameters retain their default values ($K = 3, m = r = 2$), and the dimension scan varies $m = r$ jointly. Tables S7 and S8 provide complete results.

Auxiliary weight and subspace dimension. For $\lambda \in \{0.3, 0.5, 0.7, 0.9, 1.2\}$, RSum varies from 552.8 to 553.2, with the highest mean at the default $\lambda = 0.7$. Varying $m = r$ from 2 to 5 yields RSum between 552.6 and 553.2, with the highest mean at $m = r = 2$. We exclude $r = 1$ because the unit-normalized auxiliary cosine loss has zero gradients when both projections are nonzero; Appendix D gives the proof.

Number of views. For $K \in \{2, 3, 4, 5, 6\}$, RSum ranges from 553.1 to 553.4 without monotonic improvement. The default $K = 3$ is within 0.2 of the highest mean while requiring only nine cross-modal view-pair scores.

Number of selected view pairs. With $K = 3, r = 2$, we vary $m \in \{2, 3, 5, 7, 9\}$ and retain the leading two right singular vectors. RSum ranges from 552.6 to 553.1, with the highest mean at $m = 2$, where SVD orthonormalizes the two selected directions. Larger m does not improve mean performance, although these small differences do not establish statistical superiority.

Perturbation strength. Across the nine combinations of $p_t \in \{0.05, 0.15, 0.25\}$ and $s_v \in \{0.20, 0.40, 0.60\}$, RSum ranges from 552.7 to 553.4. The default $p_t = 0.25, s_v = 0.20$ achieves 553.2, only 0.2 below the highest mean.

5 Conclusion

We proposed APS, a pair-specific subspace alignment method for noisy correspondence learning, which extracts auxiliary supervision from within each observed image-text pair to strengthen learning from reliable associations. A frozen teacher model obtained through warm-up evaluates multi-view consistency, and joint image-text representations are used to construct pair-specific subspaces that determine where auxiliary alignment is applied. Correspondence confidence jointly weights the global contrastive and auxiliary subspace objectives, reducing the influence of low-confidence pairs while promoting image-text agreement for more reliable pairs within their constructed spaces. Experiments on three benchmark datasets demonstrate competitive retrieval performance under synthetic and real-world correspondence noise, providing a practical approach to robust cross-modal learning based on evidence within individual pairs.

AI USE STATEMENT

We used generative large language models to assist with language editing to improve the clarity and readability of the manuscript, code development, and data auditing. All AI-assisted text, code, and data-auditing outputs were reviewed and verified by the authors before being incorporated into this work. The authors take full responsibility for the final content of the paper and its associated research artifacts, including the implementation, data processing, and reported results.

REPRODUCIBILITY STATEMENT

Section 3 specifies the confidence estimation, teacher warm-up, subspace construction, and student objective. Appendix A provides dataset splits and implementation settings. Appendices B, C, and F report additional results and run-to-run variability, Appendix D gives the assumptions and proof for the one-dimensional degeneracy result, and Appendix E defines the scope of the cost measurement.

REFERENCES

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep Canonical Correlation Analysis. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1247–1255. PMLR, 2013. URL https://proceedings.mlr.press/v28/andrew13.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020a. URL https://proceedings.mlr.press/v119/chen20j.html.

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 *Proceedings of the European Conference on Computer Vision*, pp. 104–120, 2020b. doi: 10.1007/978-3-030-58577-8_7. URL https://doi.org/10.1007/978-3-030-58577-8_7.

Zhuohang Dang, Minnan Luo, Chengyou Jia, Guang Dai, Xiaojun Chang, and Jingdong Wang. Noisy Correspondence Learning with Self-Reinforcing Errors Mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1463–1471, 2024. doi: 10.1609/aaai.v38i2.27911. URL https://doi.org/10.1609/aaai.v38i2.27911.

Zhuohang Dang, Minnan Luo, Jihong Wang, Chengyou Jia, Haochen Han, Herun Wan, Guang Dai, Xiaojun Chang, and Jingdong Wang. Disentangled Noisy Correspondence Learning. *IEEE Transactions on Image Processing*, 34:2602–2615, 2025. doi: 10.1109/tip.2025.3559457. URL https://doi.org/10.1109/tip.2025.3559457.

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. doi: 10.1609/aaai.v35i2.16209. URL https://doi.org/10.1609/aaai.v35i2.16209.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. Teaching Structured Vision & Language Concepts to Vision & Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2657–2668, 2023. doi: 10.1109/cvpr52729.2023.00261. URL https://doi.org/10.1109/cvpr52729.2023.00261.

Yue Duan, Zhangxuan Gu, Zhenzhe Ying, Lei Qi, Changhua Meng, and Yinghuan Shi. PC^2: Pseudo-Classification Based Pseudo-Captioning for Noisy Correspondence Learning in Cross-Modal Retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9397–9406, 2024. doi: 10.1145/3664647.3680860. URL https://doi.org/10.1145/3664647.3680860.

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*, 2018. URL https://www.bmva-archive.org.uk/bmvc/2018/contents/papers/0344.pdf. Paper 344.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. CyCLIP: Cyclic Contrastive Language-Image Pretraining. In *Advances in Neural Information Processing Systems*, volume 35, pp. 6704–6719, 2022. doi: 10.52202/068431-0486. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/2cd36d327f33d47b372d4711edd08de0-Abstract-Conference.html.

Ao Han and Yang Cao. Cross-Modal Matching with Noisy Correspondence via Neighbor Replacing. In *Advanced Intelligent Computing Technology and Applications: International Conference on Intelligent Computing*, pp. 309–320, 2025. doi: 10.1007/978-981-96-9794-6_26. URL https://doi.org/10.1007/978-981-96-9794-6_26.

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. doi: 10.1109/cvpr52729.2023.00726. URL https://doi.org/10.1109/cvpr52729.2023.00726.

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. 26669–26678, 2024. doi: 10.1109/cvpr52733.2024.02519. URL https://doi.org/10.1109/cvpr52733.2024.02519.

Haochen Han, Alex Jinpeng Wang, Peijun Ye, and Fangming Liu. Unlearning the Noisy Correspondence Makes CLIP More Robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4518–4528, 2025. doi: 10.1109/iccv51701.2025.00430. URL https://doi.org/10.1109/iccv51701.2025.00430.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020. doi: 10.1145/3394171.3413678. URL https://doi.org/10.1145/3394171.3413678.

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, pp. 926–934, 2023. doi: 10.1609/aaai.v37i1.25172. URL https://doi.org/10.1609/aaai.v37i1.25172.

Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with Noisy Correspondence for Cross-modal Matching. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29406–29419, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f5e62af885293cf4d511ceef31e61c80-Abstract.html.

Ding Jiang and Mang Ye. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2787–2797, 2023. doi: 10.1109/cvpr52729.2023.00273. URL https://doi.org/10.1109/cvpr52729.2023.00273.

Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and Constructing Latent Modality Structures in Multi-Modal Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7661–7671, 2023. doi: 10.1109/cvpr52729.2023.00740. URL https://doi.org/10.1109/cvpr52729.2023.00740.

Wooyoung Kang, Jonghwan Mun, Sungjun Lee, and Byungseok Roh. Noise-aware Learning from Web-crawled Image-Text Data for Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2930–2940, 2023. doi: 10.1109/iccv51070.2023.00275. URL https://doi.org/10.1109/iccv51070.2023.00275.

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. 212–228, 2018. doi: 10.1007/978-3-030-01225-0_13. URL https://doi.org/10.1007/978-3-030-01225-0_13.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9694–9705, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/505259756244493872b7709a8a01b536-Abstract.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Ruoxuan Li, Xiangyu Wu, and Yang Yang. Noise Self-Correction via Relation Propagation for Robust Cross-Modal Retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4748–4757, 2025. doi: 10.1145/3746027.3755585. URL https://doi.org/10.1145/3746027.3755585.

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 *Proceedings of the European Conference on Computer Vision*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Yijie Lin, Jie Zhang, Zhenyu Huang, Jia Liu, Zujie Wen, and Xi Peng. Multi-granularity Correspondence Learning from Long-term Noisy Videos. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/c0f1eed3f3ca4b2e19ab639a8426270b-Paper-Conference.pdf.

Fan Liu, Chenwei Dong, Chuanyi Zhang, Hualiang Zhou, and Jun Zhou. Robust Noisy Correspondence Learning via Self-Drop and Dual-Weight. *arXiv preprint arXiv:2412.06172*, 2024. doi: 10.48550/arXiv.2412.06172. URL https://arxiv.org/abs/2412.06172.

Yang Liu, Wentao Feng, Shu-Dong Huang, Yalan Ye, and Jiancheng Lv. Intra-Modal Neighbors Never Lie: Rectifying Inter-Modal Noisy Correspondence via Graph-Based Intra-Modal Reasoning. *arXiv preprint arXiv:2606.04061*, 2026. doi: 10.48550/arXiv.2606.04061. URL https://arxiv.org/abs/2606.04061.

Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep Evidential Learning with Noisy Correspondence for Cross-modal Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4948–4956, 2022. doi: 10.1145/3503161.3547922. URL https://doi.org/10.1145/3503161.3547922.

Yang Qin, Yuan Sun, Dezhong Peng, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Cross-modal Active Complementary Learning with Self-refining Correspondence. In *Advances in Neural Information Processing Systems*, volume 36, pp. 24829–24840, 2023. doi: 10.52202/075280-1078. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4e580cdd54fe38ca9a5b8ea6fe99bb44-Abstract-Conference.html.

Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-Correspondence Learning for Text-to-Image Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27187–27196, 2024. doi: 10.1109/cvpr52733.2024.02568. URL https://doi.org/10.1109/cvpr52733.2024.02568.

Yang Qin, Lifu Huang, Dezhong Peng, Bohan Jiang, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Trustworthy Visual-Textual Retrieval. *IEEE Transactions on Image Processing*, 34:4515–4526, 2025. doi: 10.1109/tip.2025.3587575. URL https://doi.org/10.1109/tip.2025.3587575.

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. URL https://proceedings.mlr.press/v139/radford21a.html.

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 (Volume 1: Long Papers)*, pp. 2556–2565, 2018. doi: 10.18653/v1/p18-1238. URL https://doi.org/10.18653/v1/p18-1238.

Dan Shi, Zechao Li, Lei Zhu, and Jinhui Tang. Pseudo-Text Guided Robust Learning for Noisy Correspondence in Cross-Modal Retrieval. *IEEE Transactions on Image Processing*, 35:3299–3310, 2026. doi: 10.1109/tip.2026.3675497. URL https://doi.org/10.1109/tip.2026.3675497.

Haitao Shi, Meng Liu, Xiaoxuan Mu, Xuemeng Song, Yupeng Hu, and Liqiang Nie. Breaking Through the Noisy Correspondence: A Robust Model for Image-Text Matching. *ACM Transactions on Information Systems*, 42(6):1–26, 2024. doi: 10.1145/3662732. URL https://doi.org/10.1145/3662732.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Multiview Coding. In *Proceedings of the European Conference on Computer Vision*, pp. 776–794, 2020. doi: 10.1007/978-3-030-58621-8_45. URL https://doi.org/10.1007/978-3-030-58621-8_45.

Wuyuqing Wang, Zeyuan Gu, and Erkun Yang. Noisy Correspondence Learning with Modality Gap Direction Correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10163–10171, 2026. doi: 10.1609/aaai.v40i12.37984. URL https://doi.org/10.1609/aaai.v40i12.37984.

Yunbo Wang, YuJie Wu, Zhien Dai, Can Tian, Jun Long, and Jianhai Chen. Noisy Correspondence Rectification via Asymmetric Similarity Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21384–21392, 2025. doi: 10.1609/aaai.v39i20.35439. URL https://doi.org/10.1609/aaai.v39i20.35439.

Jiantao Wu, Sara Atito, Zhenhua Feng, Shentong Mo, Josef Kitler, and Muhammad Awais. Rethinking Positive Pairs in Contrastive Learning. *arXiv preprint arXiv:2410.18200*, 2025. doi: 10.48550/arXiv.2410.18200. URL https://arxiv.org/abs/2410.18200v2. Version 2, revised May 29, 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. URL https://doi.org/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*, pp. 19883–19892, 2023. doi: 10.1109/cvpr52729.2023.01904. URL https://doi.org/10.1109/cvpr52729.2023.01904.

Shuo Yang, Yancheng Long, Yujie Wei, Zeke Xie, Hongxun Yao, Min Xu, and Liqiang Nie. Noisy Correspondence Rectification in Multimodal Clustering Space for Cross-Modal Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(4):4657–4672, 2026. doi: 10.1109/tpami.2025.3646184. URL https://doi.org/10.1109/tpami.2025.3646184.

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*, pp. 17700–17709, 2024. doi: 10.1109/cvpr52733.2024.01676. URL https://doi.org/10.1109/cvpr52733.2024.01676.

Tao Yao, Yizhen Wu, Linliang Zhang, Guorui Sheng, Ying Li, and Qi Tian. Correspondence Calibrating and Dynamic Consistency Learning for Noisy Cross-Modal Retrieval. *IEEE Transactions on Multimedia*, 28:5190–5202, 2026. doi: 10.1109/tmm.2026.3664974. URL https://doi.org/10.1109/tmm.2026.3664974.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://doi.org/10.1162/tacl_a_00166.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and Why Vision-Language Models Behave like Bags-of-Words, and What to Do about It? In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Quanxing Zha, Xin Liu, Yiu-ming Cheung, Xing Xu, Nannan Wang, and Jianjia Cao. UGNCL: Uncertainty-Guided Noisy Correspondence Learning for Efficient Cross-Modal Matching. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 852–861, 2024. doi: 10.1145/3626772.3657806. URL https://doi.org/10.1145/3626772.3657806.

Quanxing Zha, Xin Liu, Shu-Juan Peng, Yiu-Ming Cheung, Xing Xu, and Nannan Wang. ReCon: Enhancing True Correspondence Discrimination through Relation Consistency for Robust Noisy Correspondence Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29680–29689, 2025. doi: 10.1109/cvpr52734.2025.02763. URL https://doi.org/10.1109/cvpr52734.2025.02763.

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. doi: 10.1609/aaai.v38i7.28564. URL https://doi.org/10.1609/aaai.v38i7.28564.

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*, pp. 27371–27380, 2024. doi: 10.1109/cvpr52733.2024.02585. URL https://doi.org/10.1109/cvpr52733.2024.02585.

Ruochen Zheng, Jiahao Hong, Changxin Gao, and Nong Sang. REPAIR: Rank Correlation and Noisy Pair Half-Replacing With Memory for Noisy Correspondence. *IEEE Transactions on Multimedia*, 28:465–478, 2026. doi: 10.1109/tmm.2025.3623547. URL https://doi.org/10.1109/tmm.2025.3623547.

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, 2026a. doi: 10.1016/j.inffus.2026.104156. URL https://doi.org/10.1016/j.inffus.2026.104156.

Fangming Zhong, Haiquan Yu, Cun Zhu, and Suhua Zhang. HaNa: Hardness and Noise-Aware Robust Cross-modal Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16495–16503, 2026b. doi: 10.1609/aaai.v40i19.38689. URL https://doi.org/10.1609/aaai.v40i19.38689.

Haochen Zhou, Guofeng Ding, Mouxing Yang, Peng Hu, Yijie Lin, and Xi Peng. Uncover Underlying Correspondence for Robust Multi-view Clustering. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8c1d92835eb4e601f396c97ec60439fe-Abstract-Conference.html.

Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. Selective Vision-Language Subspace Projection for Few-shot CLIP. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3848–3857, 2024. doi: 10.1145/3664647.3680885. URL https://doi.org/10.1145/3664647.3680885.