SMOOTHALIGN: PROBABILISTIC SMOOTHING FOR MULTIMODAL REPRESENTATION LEARNING AND ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal representation learning and alignment are fundamental to cross-modal understanding and retrieval. However, common multimodal alignment objectives represent samples as deterministic points and assign every unpaired observation the same negative status: semantically related but identity-different samples are repelled as strongly as unrelated ones. We identify this structural failure as the semantic overlap bias. To correct this bias, we propose a probabilistic method, **SmoothAlign**, which uses a confidence-calibrated distributional prior derived from Gaussian semantic clouds of intermediate encoder representations to assign different repulsion strengths to unpaired samples according to their semantic overlap. We further prove that SmoothAlign admits a conservative mutual-information lower bound. SmoothAlign consistently outperforms the corresponding baselines in both video retrieval and image–text retrieval. A controlled IEMOCAP diagnostic confirms that the correction assigns stronger softening to semantically overlapping non-pairs than to ordinary negatives. The code is available at `https://anonymous.4open.science/r/SmoothAlign-4EF0/`.

1 INTRODUCTION

Multimodal representation learning maps heterogeneous observations into a shared space through cross-modal alignment, enabling retrieval and multimodal understanding. CLIP Radford et al. (2021) aligns images and text, while VAST Chen et al. (2023b) and LanguageBind Zhu et al. (2024) extend alignment to multiple modalities. GRAM Cicchetti et al. (2025) instead models multimodal consistency through the volume of the parallelotope formed by within-sample representations. Despite these architectural differences, existing methods typically use deterministic point representations and binary paired-versus-unpaired supervision, assigning the same negative status to every non-paired sample regardless of semantic relatedness.

We argue that this uniform treatment introduces a structural bias into multimodal alignment objectives. For an anchor, a standard contrastive objective combines attraction toward its matched sample with repulsion from unpaired samples in the batch. Binary supervision does not encode semantic relatedness, so semantically related but identity-different samples are repelled as strongly as unrelated ones. We call this excessive repulsion between semantically overlapping non-pairs the **semantic overlap bias**. The bias follows from the paired-versus-unpaired assignment itself, rather than from limited batches or sampling noise, and appears in both InfoNCE-based objectives and joint geometric objectives such as GRAM.

PCME Chun et al. (2021) and PCME++ Chun (2024) use probabilistic representations to capture uncertainty and multiplicity in cross-modal correspondences beyond a binary view of matching. Hard-negative mining strategies Robinson et al. (2021) regulate which negative samples are selected during training. Uncertainty modeling and negative selection, however, do not by themselves resolve the structural bias introduced by binary paired-versus-unpaired assignments. The issue concerns not only which negatives enter the objective, but also whether their repulsive contributions are calibrated to semantic overlap. Insufficiently calibrated repulsion can push semantically compatible samples apart and distort the learned cross-modal geometry. This observation motivates us

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.

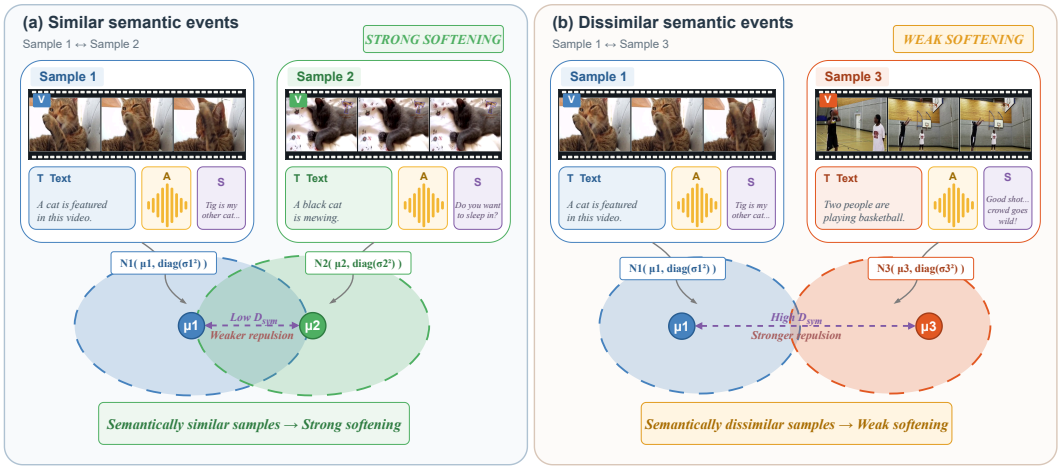


Figure 1: SmoothAlign adapts negative repulsion according to semantic affinity: semantically similar samples receive stronger softening, while dissimilar samples retain stronger repulsion.

to explore objective-level reweighting to reduce excessive repulsion while preserving the original positive supervision and negative sampling scheme.

We revisit multimodal alignment from this bias-correction perspective and propose **SmoothAlign**. As illustrated in Figure 1, SmoothAlign derives a sample-level semantic prior from the intermediate representations produced by different modality encoders and summarizes it as a diagonal Gaussian semantic cloud. SmoothAlign compares these clouds and calibrates the resulting affinity with within-sample cross-modal consistency. The calibrated affinity is converted into a detached negative weight that scales the corresponding loss contribution, while positive supervision and negative sampling remain unchanged. For semantically dissimilar samples, the clouds are more separated and receive weaker softening, thereby preserving stronger repulsion. Because the correction acts on negative terms rather than on a particular similarity function, we can instantiate SmoothAlign with different alignment objectives.

Theoretically, we show that binary paired-versus-unpaired supervision induces semantic overlap bias. Since SmoothAlign corrects this bias by reweighting negative contributions, we further examine the information-theoretic implications of this correction. We show that the resulting weighted objective admits a conservative mutual-information lower bound and that, under ideal calibration, the correction can remove overlap-induced slack. Experimentally, SmoothAlign consistently improves the corresponding video- and image–text retrieval baselines. A controlled IEMOCAP diagnostic further shows that the correction assigns stronger softening to potential false negatives than to ordinary negatives, supporting the intended bias-correction mechanism.

The main contributions of this paper are summarized as follows:

- We identify and formalize the structural bias in contrastive alignment objectives, showing that binary supervision does not distinguish semantically overlapping non-pairs from unrelated negatives when assigning repulsive supervision.
- We propose **SmoothAlign**, which uses confidence-gated affinity between Gaussian semantic clouds to adaptively reweight negative contributions, mitigating the structural bias introduced by binary supervision.
- We evaluate SmoothAlign with GRAM, standard InfoNCE, and PCME++ across multimodal retrieval settings. Ablations and controlled diagnostics show consistent retrieval gains and selective softening of semantically related non-pairs.

2 Related Work

2.1 Multimodal Learning

Multimodal learning commonly organizes heterogeneous observations in a shared space by learning cross-modal correspondences. CLIP Radford et al. (2021) established a scalable contrastive paradigm for image–text correspondence, followed by vision–language models with finer cross-modal interaction and stronger pretrained representations, including ViLT Kim et al. (2021), ALBEF Li et al. (2021), BLIP Li et al. (2022b), and BLIP-2 Li et al. (2023a); CLIP4Clip Luo et al. (2022) extended this paradigm to video–text retrieval. More recent methods broaden pairwise correspondence to multiple sensory modalities, as exemplified by AudioCLIP Guzhov et al. (2022), ULIP Xue et al. (2023), ImageBind Girdhar et al. (2023), LanguageBind Zhu et al. (2024), and VAST Chen et al. (2023b). Whereas these methods predominantly use pairwise or anchor-based relations, GRAM Cicchetti et al. (2025) jointly aligns n modalities by minimizing the parallelotope volume spanned by their representations. Across these designs, binary supervision generally assigns the same negative status to all non-pairs without encoding semantic relatedness. How to correct the uniform repulsion assigned to all non-paired samples under binary supervision is the problem that SmoothAlign is designed to solve.

2.2 Probabilistic Representation Learning

Probability distributions have been widely used to represent uncertainty that is difficult to express with deterministic features. Variational autoencoders Kingma & Welling (2014) and Bayes by Backprop Blundell et al. (2015) model uncertainty through distributions over latent variables or network parameters, while Gaussian embeddings Vilnis & McCallum (2015) use KL divergence to characterize semantic relations between distributions. In multimodal learning, JMVAE-kl Suzuki et al. (2017) uses KL divergence to encourage consistency between multimodal and unimodal posteriors. PCME Chun et al. (2021) and PCME++ Chun (2024) further represent images and texts probabilistically to model the multiplicity and uncertainty of cross-modal correspondences through probabilistic matching distances, pseudo-positive assignments, and soft-label strategies. A related line of research uses hard-negative mining Robinson et al. (2021) to select informative negatives during contrastive training, while threshold-based strategies determine whether candidate pairs should be included or excluded according to their similarity. These approaches address uncertainty at the representation or matching level, or regulate negative samples at the sampling level, but do not continuously calibrate the contribution of every retained negative pair according to its semantic overlap. SmoothAlign instead converts cross-modal distributional similarity into continuous pairwise weights for calibrating the contribution of retained negatives.

3 Method

3.1 Semantic Overlap Bias and Information-Theoretic Analysis

Semantic Overlap Bias and Its Correction. Consider a batch $\mathcal{B}$ containing N paired observations. For an anchor x_i, let y_i denote its paired observation and $\{y_j\}_{j\neq i}$ the unpaired observations. We denote their alignment logit by $u_{ij} = u_\theta(x_i, y_j)$, where a larger value indicates stronger alignment. This formulation covers standard InfoNCE when u_{ij} is a temperature-scaled cosine similarity and the volume-based contrastive term of GRAM when u_{ij} is a scaled negative Gramian volume. The standard contrastive loss ℓ_i^{std} and its normalized probability π_{ij} are

$$\ell_i^{\mathrm{std}} = -\log \frac{\exp(u_{ii})}{\sum_{k=1}^{N} \exp(u_{ik})}, \qquad \pi_{ij} = \frac{\exp(u_{ij})}{\sum_{k=1}^{N} \exp(u_{ik})}. \tag{1}$$

Its gradient is

$$\nabla_\theta \ell_i^{\mathrm{std}} = (\pi_{ii} - 1)\nabla_\theta u_{ii} + \sum_{j\neq i} \pi_{ij} \nabla_\theta u_{ij}. \tag{2}$$

The first term attracts the paired observation, whereas every unpaired observation receives a non-zero repulsive coefficient π_{ij}. Thus, semantically related samples are still pushed apart when they are assigned different instance identities.

To formalize this mismatch, let $r_{ij} \in [0, 1]$ denote the latent semantic overlap between two unpaired samples. Relative to an overlap-aware objective that scales their repulsion by $1 - r_{ij}$, Equation 2 can be decomposed as

$$\nabla_\theta \ell_i^{\text{std}} = (\pi_{ii} - 1)\nabla_\theta u_{ii} + \sum_{j \neq i} (1 - r_{ij})\pi_{ij} \nabla_\theta u_{ij} + \underbrace{\sum_{j \neq i} r_{ij}\pi_{ij}\nabla_\theta u_{ij}}_{\mathbf{b}_i}. \tag{3}$$

Here, $\mathbf{b}_i$ represents the **semantic overlap bias**, namely the excessive repulsive contribution assigned to semantically overlapping but unpaired samples. This bias is structural because binary supervision determines whether a pair is positive or negative but provides no mechanism for calibrating a negative pair according to its semantic overlap.

SmoothAlign formalizes the correction by estimating the overlap $\widehat{r}_{ij} \in [0, 1]$ for each retained non-pair and assigning it a detached weight

$$W_{ij} = \max(1 - \widehat{r}_{ij}, w_{\min}). \tag{4}$$

SmoothAlign uses W_{ij} to adjust the contribution of each retained negative pair. A larger estimated semantic overlap produces a smaller weight and hence weaker repulsion, whereas a dissimilar pair receives a weight close to 1 and largely retains its original contribution. This correction preserves the positive supervision, negative candidate set, and base alignment function. The same pairwise reweighting principle can be applied to contrastive, geometric, and probabilistic alignment objectives. Specifically, for InfoNCE and the volume contrastive branch of GRAM, the correction is implemented by adding $\log W_{ij}$ to the negative logit u_{ij}; for DAM and PCME++, W_{ij} directly scales the corresponding negative loss. The weights are estimated from cross-modal statistics of intermediate representations without external semantic supervision or additional trainable parameters, as detailed in Section 3.2.

Mutual-Information Guarantee and Gradient Effect. This pairwise reweighting raises two questions: whether it preserves the original objective's theoretical interpretation and how it changes the relative influence of negative samples. Because a standard mutual-information lower bound is available for InfoNCE, we analyze the first question through the InfoNCE instantiation of SmoothAlign. Let $\widetilde{\ell}_i$ denote the SmoothAlign-weighted loss for anchor i, and let $\mathcal{L}_{\text{SA}} = \mathbb{E}[\widetilde{\ell}_i]$ denote its population objective, where the expectation is taken over positive pairs and sampled negatives and is approximated by averaging over minibatch anchors during training. We use detached weights satisfying $w_{\min} \leq W_{ij} \leq 1$, where $w_{\min} > 0$ is a fixed lower bound on the retained negative weight. This floor prevents any negative term from being removed completely and limits the maximum strength of the correction. Under the standard InfoNCE sampling assumption, the weighted objective admits the following conservative mutual-information lower bound:

$$\widetilde{\ell}_i = -\log \frac{\exp(u_{ii})}{\exp(u_{ii}) + \sum_{j \neq i} W_{ij} \exp(u_{ij})}, \qquad I(X; Y) \geq \log(N w_{\min}) - \mathcal{L}_{\text{SA}}. \tag{5}$$

This bound establishes that bounded negative reweighting retains a valid, although more conservative, mutual-information interpretation. It also quantifies the role of $w_{\min}$: a smaller value permits stronger softening of semantically overlapping negatives but decreases the constant $\log(N w_{\min})$, making the guaranteed lower bound more conservative.

To address the second question, let $\widetilde{\pi}_{ij}$ denote the normalized probability of pair (i, j) under the weighted objective. The negative-to-positive gradient ratio satisfies

$$\frac{\widetilde{\pi}_{ij}}{\widetilde{\pi}_{ii}} = W_{ij} \exp(u_{ij} - u_{ii}). \tag{6}$$

Equation 6 shows that W_{ij} directly controls the relative influence of each negative pair: a smaller weight reduces its repulsive gradient without changing the corresponding positive term. Although the mutual-information bound in Equation 5 is specific to the InfoNCE instantiation, the gradient relation in Equation 6 is equally applicable to the GRAM objective when u_{ij} is defined as its negative Gramian-volume logit. Complete assumptions and proofs are provided in Appendix C.

3.2 Implementation of SmoothAlign

Figure 2 illustrates the SmoothAlign pipeline. Raw multimodal inputs are processed by their respective encoders, whose intermediate representations are aggregated into diagonal Gaussian semantic

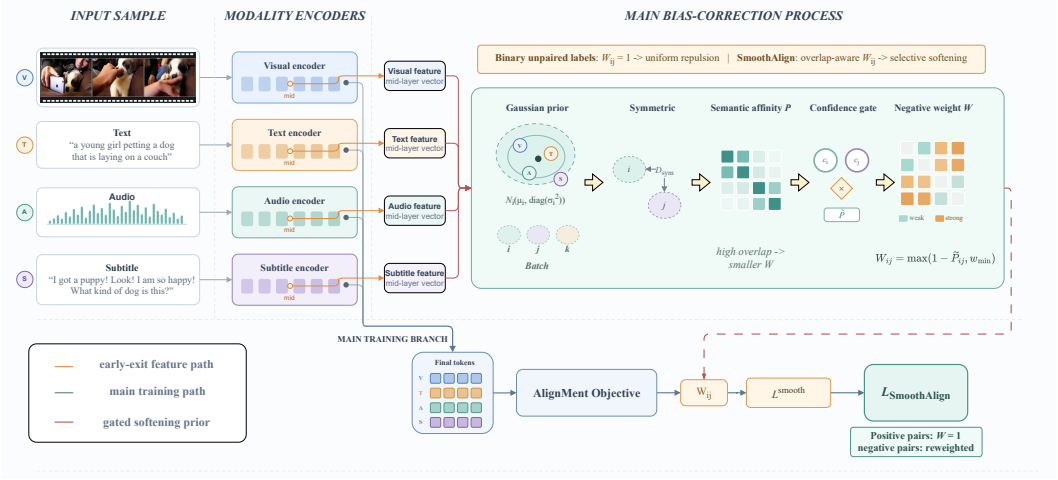


Figure 2: Overview of SmoothAlign. Multimodal intermediate representations form Gaussian semantic clouds for adaptive negative reweighting.

clouds. The distributional affinity between clouds is then converted into adaptive weights, which are applied directly to the negative terms of the underlying alignment objective.

Specifically, for sample i, we extract an intermediate feature from each of its M valid modality encoders and project all features into a shared d-dimensional space. Let $\mathbf{h}_i^{(m)} \in \mathbb{R}^d$ denote the projected feature produced by the encoder of modality m, where $m \in \{1, \ldots, M\}$. After pooling and ℓ_2 normalization, we parameterize the semantic cloud as

$$q_i = \mathcal{N}\big(\boldsymbol{\mu}_i, \mathrm{diag}(\boldsymbol{\sigma}_i^2)\big)\,, \tag{7}$$

where

$$\boldsymbol{\mu}_i = \frac{1}{M}\sum_{m=1}^{M} \mathbf{h}_i^{(m)}, \qquad \boldsymbol{\sigma}_i^2 = \frac{1}{M}\sum_{m=1}^{M}\Big(\mathbf{h}_i^{(m)} - \boldsymbol{\mu}_i\Big)^2. \tag{8}$$

The mean $\boldsymbol{\mu}_i$ represents the semantic center shared across modalities, while the variance $\boldsymbol{\sigma}_i^2$ characterizes their cross-modal variation.

The relationship between samples i and j can be measured by a suitable discrepancy between their semantic clouds, instantiated here with the KL divergence but applicable to other distributional distances or similarity measures as well:

$$D_{\mathrm{KL}}(q_i \| q_j) = \frac{1}{2}\sum_{k=1}^{d}\left[\log\frac{\sigma_{j,k}^2}{\sigma_{i,k}^2} + \frac{\sigma_{i,k}^2 + (\mu_{i,k} - \mu_{j,k})^2}{\sigma_{j,k}^2} - 1\right]. \tag{9}$$

KL divergence is asymmetric. We use its symmetric form and convert the resulting distance into a semantic affinity:

$$D_{\mathrm{sym}}(i,j) = \frac{1}{2}\left[D_{\mathrm{KL}}(q_i \| q_j) + D_{\mathrm{KL}}(q_j \| q_i)\right], \qquad P_{ij} = \exp\left(-\frac{D_{\mathrm{sym}}(i,j)}{\tau_{\mathrm{prior}}}\right). \tag{10}$$

We use P_{ij} as a pair-specific distributional prior for negative reweighting, where it estimates the semantic overlap between samples i and j from their multimodal intermediate representations. The prior temperature τ_{prior} calibrates the numerical scale of the distributional distance. A larger P_{ij} indicates more similar cross-modal statistical structure and hence greater estimated semantic overlap.

Distributional affinity can be less reliable for cross-modally inconsistent samples. We calibrate it using bidirectional cross-modal self-matching. For each modality pair (a, b), we compute the inner product between the corresponding representations and scale it by a temperature parameter:

$$s_{ij}^{(a,b)} = \frac{\left\langle \mathbf{h}_i^{(a)}, \mathbf{h}_j^{(b)} \right\rangle}{\tau_{\mathrm{conf}}}, \tag{11}$$

where τ_{conf} controls the sharpness of the matching probabilities. The probabilities that the two modality observations of sample i retrieve each other are

$$p_i^{a\to b} = \frac{\exp(s_{ii}^{(a,b)})}{\sum_{j\in\mathcal{B}}\exp(s_{ij}^{(a,b)})}, \qquad p_i^{b\to a} = \frac{\exp(s_{ii}^{(a,b)})}{\sum_{j\in\mathcal{B}}\exp(s_{ji}^{(a,b)})}. \tag{12}$$

Let $\mathcal{C} = \{(a,b) \mid 1 \le a < b \le M\}$ denote the set of modality pairs. We define the sample-level confidence $c_i \in [0,1]$ by aggregating the bidirectional self-matching consistency of sample i across all modality pairs:

$$c_i = \left[\frac{1}{|\mathcal{C}|}\sum_{(a,b)\in\mathcal{C}}\sqrt{p_i^{a\to b}p_i^{b\to a}}\right]^{\gamma}, \tag{13}$$

where a larger c_i indicates more reliable cross-modal consistency, and γ controls the concentration of the confidence scores. The confidence-gated semantic affinity is

$$\widetilde{P}_{ij} = P_{ij}\sqrt{c_i c_j}. \tag{14}$$

A large $\widetilde{P}_{ij}$ requires both similar semantic clouds and reliable within-sample cross-modal correspondence. Dissimilar or inconsistent samples retain a small affinity.

The confidence-gated affinity $\widetilde{P}_{ij}$ is converted into a pair-specific weight for each negative term:

$$W_{ij} = \max\left(1 - \widetilde{P}_{ij}, w_{\min}\right), \qquad i \neq j, \tag{15}$$

where $w_{\min} > 0$ prevents any negative contribution from being removed completely. For an objective expressed with negative logits, SmoothAlign adjusts the original logit u_{ij} as

$$\widetilde{u}_{ij} = u_{ij} + \log W_{ij}. \tag{16}$$

For an objective with explicit pairwise negative losses, the corresponding negative loss term is instead scaled directly by W_{ij}. Whether W_{ij} is applied to a negative logit or directly to a pairwise negative loss, positive terms retain unit weight, and the original negative sampling procedure remains unchanged. The lower bound $w_{\min}$ ensures that the resulting weighted objective retains a conservative mutual-information lower bound.

As a concrete instantiation, we apply SmoothAlign to both components of GRAM. The negative logits in the volume contrastive objective are adjusted according to Equation 16, while each negative pair selected by the original hard-negative mining procedure has its Data–Anchor Matching (DAM) loss scaled by the corresponding weight. The resulting objective is

$$\mathcal{L}_{\text{SmoothAlign}} = \mathcal{L}_{\text{GRAM}}^{\text{smooth}} + \lambda_{\text{DAM}}\mathcal{L}_{\text{DAM}}^{\text{smooth}}, \tag{17}$$

where $\lambda_{\text{DAM}} = 0.1$, matching the original GRAM configuration.

This implementation provides one concrete instantiation of SmoothAlign. In our experiments, SmoothAlign also performs effectively when integrated with InfoNCE and PCME++, demonstrating that the same negative-reweighting principle can be applied to different alignment objectives. Implementation details are provided in Appendix A.3.

4 Experiments

4.1 Experimental Setup

We evaluate SmoothAlign with the GRAM volume objective on MSR-VTT Xu et al. (2016), DiDeMo Hendricks et al. (2017), ActivityNet Heilbron et al. (2015), and VATEX Wang et al. (2019), using BERT-base Devlin et al. (2019), EVA-CLIP Sun et al. (2023), and BEATs Chen et al. (2023a) as the language, visual, and audio backbones, respectively, and report T2V and V2T R@1 under the available modality configurations. We further evaluate standard InfoNCE and PCME++ on COCO Captions Chen et al. (2015) with CLIP ViT-B/16, reporting the metrics listed in Table 2. IEMOCAP Busso et al. (2008) is used for the controlled structural-bias diagnostic. Dataset statistics, training protocols, and hyperparameter settings are provided in Appendix A.

Table 1: Finetuning multimodal text-to-video (T2V) and video-to-text (V2T) retrieval results in terms of Recall at 1 score (R@1).

		MSR-VTT		DiDeMo		ActivityNet		VATEX	
Method	Modality	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
CLIP4Clip Luo et al. (2022)	T-V	45.6	45.9	43.0	43.6	40.3	41.6	63.0	78.3
InternVideo-L Wang et al. (2022a)	T-V	55.2*	57.9*	57.9*	59.1*	62.2*	62.8*	71.1*	87.2*
UMT-L Li et al. (2023b)	T-V	58.8*	58.6*	70.4*	65.7*	66.8*	64.4*	72.0*	86.0*
HiTeA Ye et al. (2023)	T-V	46.8	-	56.5	-	-	-	-	-
mPLUG-2 Xu et al. (2023)	T-V	53.1	-	56.4	-	-	-	-	-
TEFAL Ibrahimi et al. (2023)	T-VA	52.0	-	-	-	-	-	61.0	-
ViCLIP Wang et al. (2024b)	T-V	52.5	51.8	49.4	50.2	49.8	48.1	-	-
Bimodal T2M Arora et al. (2024)	T-VA	36.8	-	-	-	-	-	-	-
T-MASS Wang et al. (2024a)	T-VA	52.7	-	53.3	-	-	-	65.6	-
vid-TLDR Choi et al. (2024)	T-V	58.5*	-	70.4*	-	65.2*	-	-	-
VALOR-L Liu et al. (2025)	T-VAS	54.4	-	57.6	-	63.4	-	76.9	-
VAST Chen et al. (2023b)	T-VA	55.8	57.6	65.6	62.0	68.8	66.7	86.9	84.1
VAST Chen et al. (2023b)	T-VAS	56.6	57.6	-	-	-	-	87.5	84.0
GRAM Cicchetti et al. (2025)	T-V	55.7	56.4	66.4	63.2	66.5	64.6	84.4	81.6
GRAM Cicchetti et al. (2025)	T-VA	58.4	59.0	67.3	63.5	69.9	66.9	87.0	84.6
GRAM Cicchetti et al. (2025)	T-VAS	64.0	64.8	-	-	-	-	87.7	84.2
SmoothAlign (Ours)	T-V	56.4	57.1	66.7	**63.9**	66.8	64.8	85.7	82.0
SmoothAlign (Ours)	T-VA	59.6	59.1	**67.8**	63.7	**70.0**	**67.2**	88.1	**85.0**
SmoothAlign (Ours)	T-VAS	**65.4**	**65.7**	-	-	-	-	**88.5**	84.2

*Finetuning and evaluation with 12 frames.

Table 2: Comparison of image–text matching methods on ECCV Caption, CxC, and COCO Caption.

	ECCV Caption			CxC	COCO Caption		
Method	mAP@R	R-P	R@1	R@1	1K R@1	5K R@1	RSUM
CLIP ZS Radford et al. (2021)	29.3	39.0	71.1	44.3	62.0	42.7	481.0
VSE∞ Chen et al. (2021)	41.7	50.6	86.3	62.3	79.1	60.7	547.2
P2RM Wang et al. (2022b)	39.7	49.5	80.7	54.2	73.8	52.5	532.7
DAA Li et al. (2022a)	20.7	30.6	50.2	25.4	43.7	23.4	410.2
InfoNCE Radford et al. (2021)	41.1	50.4	84.8	60.9	78.3	59.3	545.5
PCME Chun et al. (2021)	41.0	50.3	84.3	59.9	77.8	58.2	544.2
PCME++ Chun (2024)	42.1	51.2	86.5	62.6	79.3	61.1	548.0
SmoothAlign (InfoNCE)	41.3	50.6	85.5	61.2	78.8	59.5	545.9
SmoothAlign (PCME++)	**42.2**	**51.3**	**87.0**	**63.0**	**79.9**	**61.6**	**549.4**

4.2 Validation Across Alignment Objectives

GRAM volume objective. Table 1 compares SmoothAlign with GRAM using identical modality encoders, pretrained initialization, and retrieval protocols. Under T–V, it improves both retrieval directions on all four datasets; for example, VATEX T2V/V2T R@1 increases from 84.4%/81.6% to 85.7%/82.0%. The gains remain consistent under T–VA, reaching 59.6%/59.1% on MSR-VTT, 67.8%/63.7% on DiDeMo, 70.0%/67.2% on ActivityNet, and 88.1%/85.0% on VATEX. Improvements are particularly visible with richer modality configurations: on MSR-VTT, T–VAS T2V/V2T increases from 64.0%/64.8% to 65.4%/65.7%, while VATEX T–VAS T2V increases from 87.7% to 88.5% with V2T maintained at 84.2%. These results align with the proposed mechanism: cross-modal statistics provide a sample-level estimate of semantic overlap, allowing SmoothAlign to reduce excessive repulsion between related unpaired samples while preserving positive alignment and separating dissimilar negatives. Run-to-run results and qualitative retrieval visualizations on MSR-VTT are provided in Appendices B.3 and B.4.

Image–text matching objectives. Table 2 evaluates SmoothAlign with standard InfoNCE and PCME++ under the same backbone, data splits, and retrieval protocol. When applied to InfoNCE, SmoothAlign improves all seven metrics: ECCV Caption mAP@R, R-P, and R@1 increase from 41.1%, 50.4%, and 84.8% to 41.3%, 50.6%, and 85.5%, respectively; CxC R@1 increases from

Table 3: Component and distance-function ablations of SmoothAlign on MSR-VTT. Checkmarks indicate the loss components to which SmoothAlign is applied, confidence gate, and distributional distance. Results are reported as Recall at 1 (R@1).

Smoothed loss		Confidence	Distance function				T–V		T–VA		T–VAS	
$\mathcal{L}_{\text{GRAM}}$	$\mathcal{L}_{\text{DAM}}$	Gating	CSD	Mean cosine	Wasserstein-2	Symmetric KL	T2V	V2T	T2V	V2T	T2V	V2T
–	–	–	–	–	–	–	55.7	56.4	58.4	59.0	64.0	64.8
✓	–	✓	–	–	–	✓	56.9	54.8	59.4	59.4	65.3	65.6
–	✓	✓	–	–	–	✓	56.5	55.2	59.3	59.0	64.7	64.9
✓	✓	–	–	–	–	✓	56.2	56.7	59.1	58.9	65.1	65.1
✓	✓	✓	✓	–	–	–	56.0	55.3	58.9	59.5	65.1	65.6
✓	✓	✓	–	✓	–	–	56.3	55.2	59.1	58.9	65.5	65.7
✓	✓	✓	–	–	✓	–	56.1	55.0	58.9	58.8	64.9	65.6
✓	✓	✓	–	–	–	✓	56.4	57.1	59.6	59.1	65.4	65.7

60.9% to 61.2%; and COCO 1K R@1, 5K R@1, and RSUM increase from 78.3%, 59.3%, and 545.5 to 78.8%, 59.5%, and 545.9. SmoothAlign also improves PCME++ across all metrics, reaching 42.2% mAP@R, 51.3% R-P, 87.0% ECCV Caption R@1, 63.0% CxC R@1, 79.9% COCO 1K R@1, 61.6% COCO 5K R@1, and 549.4 RSUM. The consistent improvements under enriched relevance annotations and standard COCO retrieval indicate that the prior is not tied to a particular evaluation protocol. More importantly, the gains over InfoNCE and PCME++ reflect two complementary roles: InfoNCE benefits from pair-specific correction of uniformly weighted negatives, while PCME++ already models correspondence uncertainty through probabilistic embeddings and further benefits from calibrating the contribution of its remaining non-matched pairs.

4.3 Ablation Studies

Impact of Smoothing Different Loss Functions. Table 3 compares the complete SmoothAlign model with variants that apply smoothing to only one loss component. The first row reports the GRAM baseline without SmoothAlign. The two single-component variants apply SmoothAlign only to $\mathcal{L}_{\text{GRAM}}$ or $\mathcal{L}_{\text{DAM}}$, respectively. Smoothing $\mathcal{L}_{\text{GRAM}}$ is particularly effective under richer modality configurations, improving T–VAS T2V/V2T R@1 from 64.0%/64.8% to 65.3%/65.6%. Smoothing $\mathcal{L}_{\text{DAM}}$ improves T2V R@1 from 58.4% to 59.3% under T–VA and from 64.0% to 64.7% under T–VAS. These results indicate that the two smoothed objectives provide complementary but configuration- and direction-dependent supervision, while their joint application produces the most consistent overall gains. Results on the remaining datasets are provided in Appendix B.2.

Impact of Confidence Gating. Removing confidence gating consistently reduces performance relative to the complete SmoothAlign model in Table 3. Restoring the gate improves T2V/V2T R@1 by 0.2/0.4, 0.5/0.2, and 0.3/0.6 points under T–V, T–VA, and T–VAS, respectively. These results indicate that cross-modal consistency provides a useful reliability signal, limiting the influence of uncertain semantic affinities while retaining stronger correction for samples with consistent multimodal observations.

Impact of Distributional Distance Functions. Table 3 compares symmetric KL divergence with CSD Chun (2024), mean cosine, and Wasserstein-2 under the same training protocol, confidence gating, negative weighting, and temperature-selection procedure. Symmetric KL outperforms CSD in five of the six retrieval settings and Wasserstein-2 in all six settings. Mean cosine is competitive under T–VAS, producing 65.5% T2V R@1 compared with 65.4% for symmetric KL and matching its 65.7% V2T R@1, but performs less consistently under T–V and T–VA. Overall, symmetric KL provides the most consistent balance across modality configurations and retrieval directions, consistent with its joint sensitivity to differences in semantic centers and variance-normalized cross-modal variation.

4.4 Controlled Diagnostic of Structural Bias Correction

To directly assess the selectivity of SmoothAlign's negative reweighting, we conduct a controlled diagnostic on IEMOCAP Busso et al. (2008). Each utterance is associated with one of six fixed

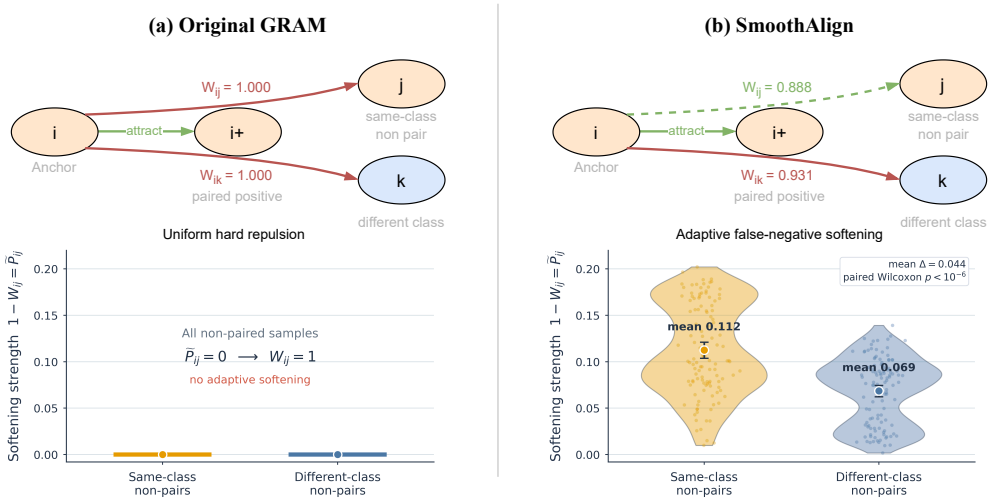


Figure 3: Controlled diagnostic of semantic-overlap correction on IEMOCAP.

emotion-text anchors, creating class-level semantic overlap among otherwise non-paired samples while keeping the anchor vocabulary and pairing protocol fixed. We initialize GRAM and SmoothAlign from the same released five-modality GRAM checkpoint and adapt them on 5,747 training utterances for 30 epochs. We freeze the modality encoders and optimize only the text, video, and audio projection heads and the contrastive temperature, limiting differences caused by encoder updates. We sample two video frames per utterance and use the provided audio features. From the held-out split, we select 40 examples from each of the three most frequent classes—frustration, neutral, and anger—yielding a class-balanced set of 120 samples. Same-class non-pairs serve as controlled proxies for semantically overlapping negatives, whereas different-class non-pairs serve as lower-overlap controls.

Figure 3 illustrates the structural bias and its correction. GRAM assigns the same explicit negative coefficient $W_{ij} = 1$ to every non-paired sample, whereas SmoothAlign adjusts this coefficient according to estimated semantic overlap. The anchor-level analysis shows that same-class non-pairs receive stronger mean softening than different-class non-pairs (0.112 versus 0.069), corresponding to average negative weights of 0.888 and 0.931, with an approximately 0.044 paired mean gap. These results support SmoothAlign's selective correction of excessive negative repulsion. Further analyses show that softening varies with baseline GRAM negative hardness while remaining semantically selective: it increases for more competitive negatives, and the same-class advantage persists across hardness bins and adaptation checkpoints. Detailed results are provided in Appendix B.1.

5 Conclusion

We identified semantic overlap bias in multimodal alignment, where binary supervision assigns repulsive gradients to semantically related but unpaired samples, and proposed SmoothAlign to correct it. By using cross-modal statistical structure to construct a pair-specific semantic-overlap prior, SmoothAlign effectively reduces excessive negative repulsion. It can be integrated into a range of existing alignment objectives, improving the performance of the underlying models without introducing additional trainable parameters. The controlled IEMOCAP diagnostic further shows that the correction acts in the intended direction, assigning stronger softening to same-class non-pairs than to different-class non-pairs. Notably, pretrained models already possess a basic understanding of each modality, providing a useful prior for SmoothAlign to estimate semantic overlap and contributing to its effectiveness during fine-tuning. Extending SmoothAlign to the pretraining stage itself remains an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools for several writing and presentation-related tasks. First, they helped edit and polish the manuscript text by improving grammar, clarity, readability, and organization, and by identifying or removing redundant exposition to make the presentation more concise. Second, they helped refine scientific figures visually, including suggestions for color palettes, typography, text placement, spacing, and overall visual consistency. These suggestions concerned presentation only; the authors determined the scientific content and final design of each figure. Third, generative AI tools assisted with literature retrieval and discovery by identifying potentially related work and helping compare the scope of prior methods with that of this paper. The authors independently checked all candidate references and comparisons against the original publications.

Generative AI tools were not used to generate synthetic datasets, formulate or prove mathematical claims, implement the method, design or execute experiments, process or interpret experimental data, or generate the reported results. The authors reviewed and revised all AI-assisted text, figure suggestions, and literature-related outputs. The authors take full responsibility for the final manuscript, including its content, claims, figures, citations, and conclusions.

REFERENCES

Pranav Arora, Selen Pehlivan, and Jorma Laaksonen. Text-to-multimodal retrieval with bimodal input fusion in shared cross-modal transformer. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pp. 15823–15834. ELRA and ICCL, 2024.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *CoRR*, abs/1505.05424, 2015.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359, 2008.

Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 15789–15798. Computer Vision Foundation / IEEE, 2021.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5178–5193. PMLR, 2023a.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015.

Joonmyung Choi, Sanghyeok Lee, Jaewon Chu, Minhyuk Choi, and Hyunwoo J. Kim. vid-tldr: Training free token merging for light-weight video transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 18771–18781. IEEE, 2024.

Sanghyuk Chun. Improved probabilistic image-text representations. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 8415–8424. Computer Vision Foundation / IEEE, 2021.

Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, and Seong Joon Oh. ECCV caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for MS-COCO. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pp. 1–19. Springer, 2022.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 15180–15190. IEEE, 2023.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pp. 976–980. IEEE, 2022.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 961–970. IEEE Computer Society, 2015.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 5804–5813. IEEE Computer Society, 2017.

Sarah Ibrahimi, Xiaohang Sun, Pichao Wang, Amanmeet Garg, Ashutosh Sanan, and Mohamed Omar. Audio-enhanced text-to-video retrieval using text-conditioned feature alignment. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 12020–12030. IEEE, 2023.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3128–3137. IEEE Computer Society, 2015.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5583–5594. PMLR, 2021.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Hao Li, Jingkuan Song, Lianli Gao, Pengpeng Zeng, Haonan Zhang, and Gongfu Li. A differentiable semantic metric approximation in probabilistic embedding for cross-modal retrieval. In

Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022a.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 9694–9705, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023a.

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 19891–19903. IEEE, 2023b.

Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. VALOR: vision-audio-language omni-perception pretraining model and dataset. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(2):708–724, 2025.

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

Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pp. 2855–2870. Association for Computational Linguistics, 2021.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

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

Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Jiamian Wang, Pichao Wang, Guohao Sun, Dongfang Liu, Sohail A. Dianat, Raghuveer Rao, Majid Rabbani, and Zhiqiang Tao. Text is MASS: modeling as stochastic embedding for text-video retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 16551–16560. IEEE, 2024a.

Xin Wang, Jiawei Wu, Jun-Kun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 4580–4590. IEEE, 2019.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *CoRR*, abs/2212.03191, 2022a.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024b.

Zheng Wang, Zhenwei Gao, Xing Xu, Yadan Luo, Yang Yang, and Heng Tao Shen. Point to rectangle matching for image text retrieval. In João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (eds.), *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pp. 4977–4986. ACM, 2022b.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. mplug-2: A modularized multi-modal foundation model across text, image and video. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38728–38748. PMLR, 2023.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 5288–5296. IEEE Computer Society, 2016.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP: learning a unified representation of language, images, and point clouds for 3d understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 1179–1189. IEEE, 2023.

Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 15359–15370. IEEE, 2023.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Caiwan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

A Experimental Details Appendix

A.1 Datasets and Evaluation Protocols

We evaluate SmoothAlign on four multimodal video retrieval benchmarks and one image–text retrieval benchmark. We additionally use IEMOCAP for the controlled diagnostic analysis of potential false negatives. Table 4 summarizes the modalities, dataset splits, and experimental usage. Here, I, T, V, A, and S denote image, text, video, audio, and subtitles, respectively.

MSR-VTT. MSR-VTT Xu et al. (2016) contains 10,000 web video clips accompanied by 200,000 natural-language descriptions, with 20 captions provided for each video. The dataset covers a broad range of visual content, including human activities, sports, music, news, and natural scenes. Following the standard 1KA retrieval split adopted by GRAM, we use 9,000 videos for training and 1,000 videos for evaluation. Following GRAM's modality preprocessing, we evaluate text–video, text–video–audio, and text–video–audio–subtitle configurations.

VATEX. VATEX Wang et al. (2019) originally consists of 41,250 video clips collected from Kinetics-600 and 825,000 sentence-level descriptions in English and Chinese. Since a substantial portion of the original videos is no longer publicly accessible, we follow the available subset used by GRAM, which contains 14,491 videos. Among them, 14,060 videos are used for training and 431 videos are used for evaluation. We conduct retrieval experiments using bimodal, trimodal, and quadrimodal configurations.

DiDeMo. DiDeMo Hendricks et al. (2017) contains approximately 10,000 untrimmed videos collected from Flickr. Each video is annotated with multiple natural-language descriptions corresponding to temporally localized moments, providing fine-grained associations between language and video content. We follow the GRAM split: 8,394 training videos, 1,065 validation videos, and 1,003 test videos. We extract audio tracks from the videos to construct the trimodal text–video–audio retrieval setting.

ActivityNet. ActivityNet Heilbron et al. (2015) is a large-scale video benchmark for human activity understanding. Following GRAM, we use its text–video retrieval setup with 10,009 videos for training and 4,917 videos for evaluation. We also extract audio tracks to support trimodal retrieval.

COCO Captions. COCO Captions Chen et al. (2015) contains 123,287 images, each paired with five human-written captions. Following the Karpathy split Karpathy & Fei-Fei (2015), we use 113,287 images for training, 5,000 images for validation, and 5,000 images for testing. We evaluate image-to-text and text-to-image retrieval on both the COCO 1K and COCO 5K settings. COCO 1K averages results over five folds, each containing 1,000 images and 5,000 captions, whereas COCO 5K evaluates retrieval over the complete test split of 5,000 images and 25,000 captions. Following PCME++ Chun (2024), we additionally evaluate the same test images and captions using the enriched relevance annotations of CxC Parekh et al. (2021) and ECCV Caption Chun et al. (2022).

IEMOCAP. IEMOCAP Busso et al. (2008) is a multimodal emotion dataset containing approximately 12 hours of scripted and improvised dyadic interactions recorded from 10 actors across five sessions. The recordings include synchronized speech, visual information, text transcriptions, and categorical emotion annotations. We use IEMOCAP only for controlled diagnostic analysis of potential false-negative repulsion, not for the main retrieval evaluation. We assign each utterance one of six fixed emotion-text anchors based on its annotation. These anchors serve as controlled text inputs for both GRAM and SmoothAlign, while the video and audio observations remain instance-specific. From the held-out test split, we select 40 examples from each of the three most frequent classes, yielding 120 evaluation samples. We introduce no explicit emotion-classification loss or additional class-label target; we use same-class and different-class labels only to organize the diagnostic analysis.

A.2 Evaluation Metrics

For the video retrieval experiments, we report Recall at 1 (R@1) in both text-to-video (T2V) and video-to-text (V2T) directions. R@1 measures the percentage of queries for which a relevant item is retrieved at the first rank. For image–text retrieval, Recall at K (R@K) measures the percentage of queries having at least one relevant item among the top-K retrieved results. Following PCME++,

Table 4: Dataset statistics and experimental usage. The IEMOCAP entry reports the 120 held-out examples used for the false-negative diagnostic, not the full dataset size.

Dataset	Modalities	Train	Validation	Test/Evaluation	Usage
MSR-VTT	T–V–A–S	9,000	–	1,000	Video retrieval
VATEX	T–V–A–S	14,060	–	431	Video retrieval
DiDeMo	T–V–A	8,394	1,065	1,003	Video retrieval
ActivityNet	T–V–A	10,009	–	4,917	Video retrieval
COCO Captions	I–T	113,287	5,000	5,000	Image–text retrieval
IEMOCAP	T–V–A	5,747	–	120	Diagnostic analysis

we average the image–text results reported in the main comparison over the image-to-text (I2T) and text-to-image (T2I) directions. On ECCV Caption, we additionally report R-Precision (R-P), which measures precision among the top R results when a query has R relevant items, and mAP@R, which averages precision over the first R ranks and then over all queries. For COCO 1K, RSUM sums R@1, R@5, and R@10 in both retrieval directions. Higher values indicate better performance for all metrics.

A.3 SmoothAlign Integration for InfoNCE and PCME++

For both image–text objectives, SmoothAlign constructs a detached pairwise affinity matrix from intermediate image and text features. We normalize the features and summarize them as diagonal Gaussian semantic clouds using their element-wise mean and variance. We compute the symmetric KL divergence between the clouds, convert it into an affinity, and calibrate it with bidirectional image–text matching confidence. The resulting affinity is converted into a negative-pair weight

$$W_{ij} = \max(1 - \widetilde{P}_{ij}, w_{\min}), \tag{18}$$

where the prior is detached from the optimization graph. Thus, the prior changes only the contribution of non-matched pairs and does not introduce an additional gradient path through the feature representations.

For InfoNCE, we first compute the original bidirectional image–text logits from the mean embeddings. For each non-matched pair, SmoothAlign modifies the logit as

$$\widetilde{z}_{ij} = z_{ij} + \log W_{ij}, \qquad j \neq i, \tag{19}$$

while keeping the matched logit unchanged, $\widetilde{z}_{ii} = z_{ii}$. The training loss is then the average of the image-to-text and text-to-image cross-entropy losses computed from the softened logits. Consequently, positive supervision, the gathered negative pool, the learnable temperature, and the inference-time cosine-similarity scoring remain unchanged.

For PCME++, we apply SmoothAlign to the existing probabilistic pairwise matching objective. We retain the original PCME++ probabilistic encoders, distributional matching distance, matched-pair targets, and regularization terms. Let ℓ_{ij} denote the PCME++ pairwise logit and m_{ij} its matched or non-matched target. We compute the element-wise binary cross-entropy loss and apply the SmoothAlign weight only to non-matched entries:

$$\mathcal{L}^{\text{smooth}}_{\text{PCME++}} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}\ \text{BCEWithLogits}(\ell_{ij}, m_{ij}), \qquad A_{ij} = \begin{cases} 1, & m_{ij} = 1, \\ W_{ij}, & m_{ij} = 0. \end{cases} \tag{20}$$

Thus, matched-pair supervision and the native PCME++ retrieval scores are preserved, while semantically overlapping non-matched pairs receive reduced loss contributions.

A.4 Implementation Details

Experiments with the GRAM volume objective. We follow the GRAM training protocol and initialize the model from its released five-modality pretrained checkpoint. We fine-tune the model for four epochs using AdamW with a learning rate of 2×10^{-5}, a weight decay of 0.01, and a global

training batch size of 64. We use a linear learning-rate scheduler with a warm-up ratio of 0.1 and set the maximum gradient norm to 2.0. We sample four video frames during training and eight frames during evaluation, with a global evaluation batch size of 128. The language-side branches use BERT-base, the visual branch uses EVA-CLIP ViT-G, and the audio branch uses BEATs. For configurations with subtitles, the subtitle stream uses the same BERT-based language encoder. All experiments use mixed-precision distributed training on two NVIDIA A100 GPUs.

Experiments with image–text matching objectives. All image–text experiments use CLIP ViT-B/16 visual and textual backbones initialized from the corresponding pretrained CLIP checkpoint. We evaluate both standard InfoNCE and PCME++'s probabilistic matching objective under the same COCO Captions split and retrieval protocol. Following the PCME++ training protocol, we train the models for 25 epochs using AdamP with an initial learning rate of 5×10^{-4}, a weight decay of 10^{-4}, and a global batch size of 128. We use a layer-wise learning-rate decay of 0.7, with learning-rate multipliers of 0.01 and 0.1 for the visual and textual backbones, respectively. We freeze the visual backbone for the first two epochs, then reduce the learning rate by a factor of 0.1 during the final ten epochs. Cross-rank feature gathering ensures that all 128 samples participate in the negative pool. For PCME++, we retain its probabilistic image and text encoders, CSD matching distance, VIB regularization, MSDA augmentation, and SWA evaluation protocol. All image–text experiments use mixed-precision distributed training on two NVIDIA A100 GPUs.

Temperature settings. The prior temperature τ_{prior} controls the conversion of the distributional distance between Gaussian semantic clouds into a semantic affinity. The confidence temperature τ_{conf} controls the sharpness of the bidirectional cross-modal self-matching probabilities, while the exponent γ controls the concentration of the resulting confidence scores. We fix these hyperparameters for each objective and share them across all SmoothAlign variants within that objective. For the GRAM volume objective, we use $\tau_{\text{prior}} = 1.0$, $\tau_{\text{conf}} = 0.07$, and $\gamma = 0.5$. For both image–text objectives, we use $\tau_{\text{prior}} = 10^4$, $\tau_{\text{conf}} = 0.07$, and $\gamma = 0.5$. We use the prior temperature only for SmoothAlign variants, while the corresponding baseline objectives retain their original formulations.

We select prior temperatures separately for the GRAM volume objective and the image–text matching objectives because their corresponding distributional distances have different numerical scales. This difference calibrates the distance-to-affinity mapping and does not change the model architecture, positive supervision, negative pool, or reweighting formulation. For each objective, all SmoothAlign variants use the same values of τ_{prior}, τ_{conf}, and γ, while the corresponding baseline is trained with its original objective and protocol without the additional prior. Therefore, we hold the temperature settings fixed for all SmoothAlign variants within each objective, while evaluating the corresponding baseline under its original training protocol. The reported comparisons are thus made under consistent training and evaluation settings, without introducing an additional temperature-based confounding factor.

B Additional Experiments

B.1 Optimization Dynamics and Hardness-Stratified Correction

We use the same held-out IEMOCAP subset as in the main-text diagnostic: 120 samples, with 40 examples from each of frustration, neutral, and anger. Softening strength is defined as $1-W_{ij}$, where W_{ij} is the SmoothAlign weight assigned to a non-paired sample; larger values therefore indicate a larger reduction in negative repulsion.

We examine two properties of the reweighting. First, we evaluate the same-class and different-class groups at six SmoothAlign adaptation checkpoints: epochs 5, 10, 15, 20, 25, and 30. Second, at the final checkpoint, we rank each anchor's non-paired candidates by the original GRAM negative logit within that anchor and divide the ranks into five equal-count bins, from easiest to hardest. All reported means are computed from per-anchor group averages, and confidence intervals are obtained by nonparametric bootstrap over the 120 anchors.

Figure 4(a) shows that the same-class softening advantage is already present at the earliest measured checkpoint. At epoch 5, same-class and different-class non-pairs receive mean softening of 0.106 and 0.067, respectively, giving an anchor-level gap of 0.039 ($[0.033, 0.046]$). The gap remains stable

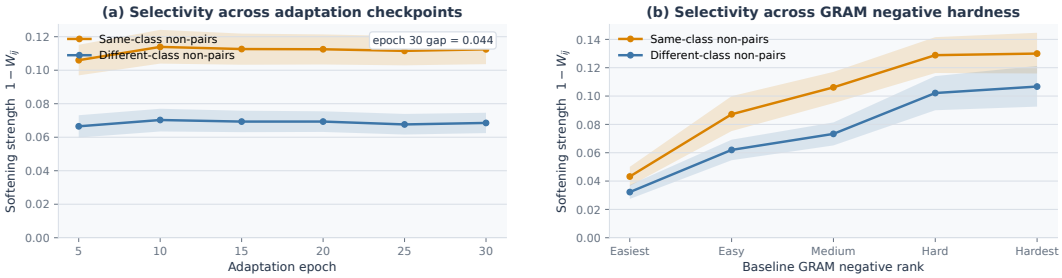


Figure 4: Optimization dynamics and hardness-stratified analysis of SmoothAlign softening on the controlled IEMOCAP split. Curves show mean softening strength $1 - W_{ij}$; shaded regions denote 95% bootstrap confidence intervals.

across the subsequent checkpoints and reaches 0.044 ($[0.037, 0.050]$) at epoch 30. This persistence indicates that the selective correction is not specific to a single final-checkpoint estimate.

Figure 4(b) shows that both groups receive stronger softening as the original GRAM negative logit becomes more competitive. At the same time, same-class non-pairs receive stronger softening within every hardness bin. From the easiest to the hardest bin, the paired anchor-level gaps are 0.011, 0.025, 0.033, 0.027, and 0.023; all five bootstrap confidence intervals for these paired gaps exclude zero, and the one-sided paired Wilcoxon tests remain significant after Bonferroni correction ($p \leq 2.32 \times 10^{-4}$ before correction). Thus, the reweighting is hardness-aware but not hardness-only: conditional on the original GRAM hardness range, it consistently prefers non-pairs that share the anchor's semantic class. This pattern supports the proposed mechanism that SmoothAlign uses semantic overlap information to selectively relax excessive negative repulsion.

B.2 Loss-Component Ablations

We extend the loss-component ablation to DiDeMo, ActivityNet, and VATEX using the same experimental setting as the corresponding GRAM baseline on each dataset. In each experiment, SmoothAlign is instantiated by applying its probabilistic prior exclusively to either the Gramian contrastive loss $\mathcal{L}_{\text{GRAM}}$ or the Data–Anchor Matching loss $\mathcal{L}_{\text{DAM}}$, while leaving the other loss component unchanged. Table 5 reports T2V and V2T R@1 under all available modality configurations.

The additional results reveal a clear asymmetry between the two retrieval directions. On DiDeMo, smoothing $\mathcal{L}_{\text{GRAM}}$ improves T2V R@1 from 66.4% to 66.9% under T–V and from 67.3% to 68.1% under T–VA, whereas the corresponding V2T performance decreases slightly. In contrast, smoothing $\mathcal{L}_{\text{DAM}}$ produces smaller but more balanced improvements, increasing T2V and V2T under both modality configurations. These results suggest that the two loss components influence the two retrieval directions differently.

On ActivityNet, the effects are more configuration-dependent. Smoothing $\mathcal{L}_{\text{GRAM}}$ improves T–V T2V from 66.5% to 67.1% while maintaining V2T at 64.6%. Under T–VA, it improves V2T from 66.9% to 67.2% but slightly reduces T2V. Smoothing $\mathcal{L}_{\text{DAM}}$ is most effective in the T–VA setting, where T2V and V2T increase from 69.9% and 66.9% to 70.3% and 67.2%, respectively. However, applying the same mechanism to $\mathcal{L}_{\text{DAM}}$ under T–V reduces both retrieval scores.

On VATEX, both smoothing variants consistently improve T2V across all modality configurations. Applying the prior to $\mathcal{L}_{\text{GRAM}}$ increases T2V from 84.4%, 87.0%, and 87.7% to 85.1%, 87.7%, and 88.1% under T–V, T–VA, and T–VAS, respectively. Smoothing $\mathcal{L}_{\text{DAM}}$ yields the strongest T2V results, reaching 85.3%, 87.9%, and 88.4%. In contrast, the corresponding V2T results remain unchanged or decrease slightly. This direction-dependent behavior suggests that the effectiveness of loss-specific smoothing depends on the query–candidate structure and the distribution of hard or potentially false-negative samples.

Overall, the results demonstrate that SmoothAlign can operate through both $\mathcal{L}_{\text{GRAM}}$ and $\mathcal{L}_{\text{DAM}}$. Applying the prior to $\mathcal{L}_{\text{GRAM}}$ generally provides greater improvements in geometric T2V alignment, particularly with richer modality configurations, whereas applying it to $\mathcal{L}_{\text{DAM}}$ produces competitive

Table 5: Loss-component ablations on multiple multimodal retrieval datasets in terms of Recall at 1 (R@1). The SmoothAlign prior is applied exclusively to either the Gramian contrastive loss ($\mathcal{L}_{\mathrm{GRAM}}$) or the Data–Anchor Matching loss ($\mathcal{L}_{\mathrm{DAM}}$).

		MSR-VTT		DiDeMo		ActivityNet		VATEX	
Method	Modality	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
GRAM Baseline	T-V	55.7	56.4	66.4	63.2	66.5	64.6	84.4	81.6
GRAM Baseline	T-VA	58.4	59.0	67.3	63.5	69.9	66.9	87.0	84.6
GRAM Baseline	T-VAS	64.0	64.8	-	-	-	-	87.7	84.2
Smooth $\mathcal{L}_{\mathrm{GRAM}}$	T-V	56.9	54.8	66.9	62.9	67.1	64.6	85.1	81.4
Smooth $\mathcal{L}_{\mathrm{GRAM}}$	T-VA	59.4	59.4	68.1	62.7	69.3	67.2	87.7	84.3
Smooth $\mathcal{L}_{\mathrm{GRAM}}$	T-VAS	65.3	65.6	-	-	-	-	88.1	84.2
Smooth $\mathcal{L}_{\mathrm{DAM}}$	T-V	56.5	55.2	66.9	63.4	65.9	63.1	85.3	81.1
Smooth $\mathcal{L}_{\mathrm{DAM}}$	T-VA	59.3	59.0	67.7	63.7	70.3	67.2	87.9	84.0
Smooth $\mathcal{L}_{\mathrm{DAM}}$	T-VAS	64.7	64.9	-	-	-	-	88.4	83.9

Table 6: Run-to-run stability of SmoothAlign with the GRAM objective on MSR-VTT. Results are reported as mean $\pm$ standard deviation of Recall at 1 (R@1) over three random seeds.

	T-V		T-VA		T-VAS	
Method	T2V	V2T	T2V	V2T	T2V	V2T
SmoothAlign (Ours)	56.4 ± 0.6	57.1 ± 0.2	59.6 ± 0.3	59.1 ± 0.1	65.4 ± 0.4	65.7 ± 0.2

improvements in several T–VA settings. Nevertheless, neither loss component universally dominates across all datasets and retrieval directions, indicating that they provide complementary pathways through which SmoothAlign can influence multimodal alignment.

B.3 Run-to-Run Stability on MSR-VTT

To evaluate run-to-run stability, we repeat the MSR-VTT experiments under the T–V, T–VA, and T–VAS modality configurations using three random seeds. Table 6 reports the mean and standard deviation of T2V and V2T R@1. Across all configurations and retrieval directions, the standard deviations range from 0.1 to 0.6. The bimodal T–V setting exhibits the largest variation, with a standard deviation of 0.6 for T2V, whereas the corresponding V2T deviation is only 0.2. After introducing audio, the T–VA configuration yields T2V and V2T deviations of 0.3 and 0.1, respectively. The richer T–VAS configuration remains similarly consistent, with deviations of 0.4 for T2V and 0.2 for V2T. Moreover, V2T exhibits lower variation than T2V under all three modality configurations.

This limited variation is consistent with SmoothAlign's design as a bounded adaptive reweighting mechanism. In this instantiation, SmoothAlign preserves the original GRAM pairing and hard-negative sampling protocol and continuously reweights the sampled negative losses using its probabilistic prior. Consequently, changing the random seed affects parameter initialization and data ordering but does not alter the underlying positive and negative supervision. The confidence gate further limits the influence of uncertain semantic affinities, reducing the risk that noisy prior estimates dominate optimization. The consistently small standard deviations therefore indicate that the improvements obtained by applying SmoothAlign to GRAM are not tied to a particular random initialization.

B.4 Qualitative Multimodal Retrieval Visualizations

Following the confusion-matrix visualization used in GRAM, we qualitatively analyze how SmoothAlign changes the joint multimodal representation space under the GRAM objective. While the original GRAM visualization demonstrates how jointly considering multiple modalities can resolve ambiguities in individual cosine-based retrieval scores, our analysis examines whether

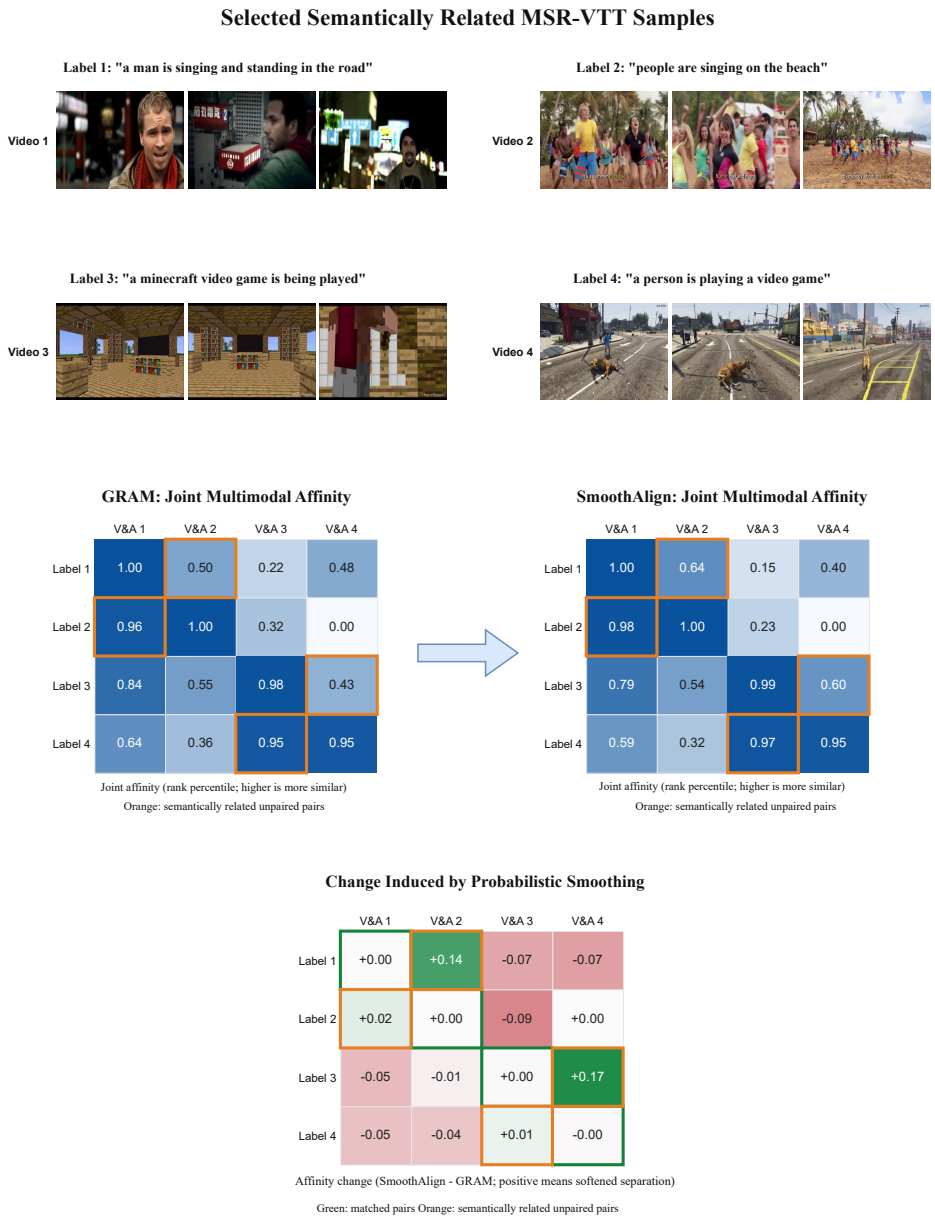


Figure 5: Qualitative comparison of the joint multimodal representation spaces learned by GRAM and its SmoothAlign-enhanced counterpart on MSR-VTT. The blue matrices report percentile-normalized joint affinities, while the bottom matrix shows the signed affinity changes introduced by SmoothAlign. Affinity differences between SmoothAlign and GRAM are computed before rounding the percentile-normalized joint affinities. SmoothAlign increases affinity within the two semantically related groups while leaving cross-group affinities unchanged or reduced in this example.

SmoothAlign relaxes the separation between semantically related but unpaired samples without uniformly increasing the affinity of unrelated samples.

We select four representative examples from the MSR-VTT test set and organize them into two interpretable semantic groups. Labels 1 and 2 correspond to singing scenes in different environments, whereas Labels 3 and 4 describe video-game scenes. Each video is represented by three key frames, and each candidate denoted by V&A consists of the paired video and audio representations of the

corresponding sample. Because the GRAM baseline and its SmoothAlign-enhanced counterpart are trained independently and may produce Gramian volumes on different numerical scales, we convert the negative three-modal Gramian volume into its rank percentile among all 1,000 test candidates. A higher percentile therefore indicates stronger joint multimodal affinity.

As shown in Figure 5, SmoothAlign consistently increases the joint affinity between semantically related samples. For the two singing scenes, the affinity from Label 1 to V&A 2 increases from 0.50 to 0.64, while the reverse direction increases from 0.96 to 0.98. Similarly, for the two video-game scenes, the affinity from Label 3 to V&A 4 increases from 0.43 to 0.60, and the reverse direction increases from 0.95 to 0.97. Meanwhile, the diagonal affinities of the matched samples remain nearly unchanged, indicating that semantic softening does not compromise positive alignment.

Importantly, the increased affinity is concentrated within the two semantic groups. The eight cross-group entries between the singing and video-game samples are either unchanged or reduced after smoothing. This behavior differs from a uniform contraction of the representation space, which would indiscriminately increase most off-diagonal affinities. Instead, the visualization supports SmoothAlign's intended behavior: its probabilistic prior selectively relaxes the geometric separation of semantically related unpaired samples while preserving, or strengthening, the separation of semantically dissimilar samples.

C Theoretical Analysis

This appendix provides the assumptions and derivations underlying the theoretical results summarized in Section 3.1. We analyze the InfoNCE instantiation of SmoothAlign from two perspectives: its mutual-information lower bound and the effect of pairwise reweighting on the relative contribution of negative samples. Throughout the analysis, we detach the SmoothAlign weights and treat them as constants during differentiation.

C.1 Mutual-Information Lower Bound

Let $\widetilde{\ell}_i$ denote the SmoothAlign-weighted InfoNCE loss for anchor i, as defined in Equation 5. Its expected loss over positive pairs and sampled negatives is denoted by

$$\mathcal{L}_{\mathrm{SA}} = \mathbb{E}\left[\widetilde{\ell}_i\right]. \tag{21}$$

In practice, we approximate this expectation by averaging the loss over anchors in a minibatch.

Proposition 1 (Conservative bound). Under the standard InfoNCE sampling assumption, suppose that the detached weights satisfy $0 < w_{\min} \le W_{ij} \le 1$. For an objective with N candidates, the SmoothAlign-weighted loss retains a conservative mutual-information lower bound.

$$0 < w_{\min} \le W_{ij} \le 1.$$

For a contrastive objective with N candidates, SmoothAlign satisfies

$$I(X;Y) \ge \log(N w_{\min}) - \mathcal{L}_{\mathrm{SA}}. \tag{22}$$

Let

$$D_i^{\mathrm{std}} = \sum_{k=1}^{N} \exp(u_{ik}), \qquad D_i^{\mathrm{SA}} = \exp(u_{ii}) + \sum_{j \ne i} W_{ij} \exp(u_{ij}) \tag{23}$$

denote the standard and SmoothAlign-weighted denominators. Since the positive term has unit weight and every negative weight is at least $w_{\min}$,

$$D_i^{\mathrm{SA}} \ge w_{\min} D_i^{\mathrm{std}}. \tag{24}$$

It follows that

$$\widetilde{\ell}_i \ge \ell_i^{\mathrm{std}} + \log w_{\min}, \qquad \mathcal{L}_{\mathrm{std}} \le \mathcal{L}_{\mathrm{SA}} - \log w_{\min}. \tag{25}$$

Combining this inequality with the standard InfoNCE bound $I(X;Y) \ge \log N - \mathcal{L}_{\mathrm{std}}$ gives

$$I(X;Y) \ge \log(N w_{\min}) - \mathcal{L}_{\mathrm{SA}}. \tag{26}$$

This bound is conservative because $w_{\min} \le 1$ reduces the constant term from $\log N$ to $\log(N w_{\min})$. Nevertheless, it shows that bounded negative reweighting preserves a valid mutual-information lower-bound interpretation.

C.2 Effect on Negative Contributions

Using the weighted denominator D_i^{SA}, define

$$\widetilde{\pi}_{ii} = \frac{\exp(u_{ii})}{D_i^{\mathrm{SA}}}, \qquad \widetilde{\pi}_{ij} = \frac{W_{ij}\exp(u_{ij})}{D_i^{\mathrm{SA}}}, \quad j \neq i. \tag{27}$$

Because W_{ij} is detached, it introduces no additional derivative path. The derivatives of the weighted loss are therefore

$$\frac{\partial \widetilde{\ell}_i}{\partial u_{ii}} = \widetilde{\pi}_{ii} - 1, \qquad \frac{\partial \widetilde{\ell}_i}{\partial u_{ij}} = \widetilde{\pi}_{ij}. \tag{28}$$

The softmax odds of negative pair (i, j) relative to the matched pair satisfy

$$\frac{\widetilde{\pi}_{ij}}{\widetilde{\pi}_{ii}} = W_{ij}\exp(u_{ij} - u_{ii}), \tag{29}$$

whereas standard InfoNCE gives

$$\frac{\pi_{ij}}{\pi_{ii}} = \exp(u_{ij} - u_{ii}). \tag{30}$$

Thus, SmoothAlign multiplies the negative-to-positive softmax odds of pair (i, j) by exactly W_{ij}. When $W_{ij} \approx 1 - r_{ij}$, semantically overlapping negatives receive smaller relative coefficients, while dissimilar negatives with $W_{ij} \approx 1$ remain close to the original objective. This reduces the excessive component $\mathbf{b}_i$ identified in Equation 3.

C.3 Ideal Gaussian Calibration

We finally consider an idealized setting that connects the Gaussian semantic clouds to the preceding analysis. Let $F_{ij} = 1$ indicate that a nominal negative pair is semantically overlapping, and assume that the Gaussian clouds are sufficient statistics for this event. Under perfect calibration, the gated Gaussian affinity satisfies

$$\widetilde{P}_{ij} = \Pr(F_{ij} = 1 \mid q_i, q_j). \tag{31}$$

This is an ideal assumption; the learned clouds and affinities used in practice need not satisfy it exactly.

For any non-negative pairwise contribution $g_{ij} = g(q_i, q_j)$, the law of total expectation gives

$$\mathbb{E}\Big[(1 - \widetilde{P}_{ij})g_{ij}\Big] = \mathbb{E}[(1 - F_{ij})g_{ij}]\,. \tag{32}$$

Thus, under ideal Gaussian calibration, the unfloored weight removes the expected contribution of semantically overlapping negatives while retaining that of ordinary negatives. With the practical floor $W_{ij} = \max(1 - \widetilde{P}_{ij}, w_{\min})$, the residual satisfies

$$0 \leq \mathbb{E}[W_{ij}g_{ij}] - \mathbb{E}[(1 - F_{ij})g_{ij}] \leq w_{\min}\mathbb{E}[g_{ij}]. \tag{33}$$

At the same time, Proposition 1 guarantees

$$I(X;Y) \geq \log(N w_{\min}) - \mathcal{L}_{\mathrm{SA}}. \tag{34}$$

Therefore, ideal Gaussian calibration provides an exact population-level interpretation of the semantic correction, while the weight floor preserves the conservative mutual-information lower bound. This result does not assume that the practical Gaussian prior is perfectly calibrated or that its bound is tighter than the standard InfoNCE bound.

C.4 Scope and Future Theoretical Directions

The preceding analysis serves two complementary purposes. First, the conservative bound shows that bounded and detached negative reweighting preserves a valid mutual-information interpretation for contrastive learning. Second, the gradient and ideal-calibration analyses explain how SmoothAlign modifies optimization: it directly scales the relative contribution of each negative pair and, under ideal calibration, removes the expected contribution associated with semantic overlap.

Together, these results characterize SmoothAlign as a controlled correction of negative supervision rather than a change to positive pairing or negative sampling.

The mutual-information bound is derived for the contrastive formulation analyzed in Proposition 1. The softmax-odds result also applies algebraically to the volume-based contrastive branch of GRAM when u_{ij} is instantiated as its negative Gramian-volume logit. However, GRAM jointly measures the geometry of multiple modalities, while PCME++ uses a probabilistic pairwise matching objective; neither objective is directly characterized by the standard InfoNCE mutual-information bound.

A useful direction for future work is therefore to develop objective-specific information-theoretic analyses for these settings. For GRAM, this may involve relating Gramian-volume minimization to multivariate dependence measures, such as total correlation or multivariate mutual information. For probabilistic matching objectives, future work may study how semantic reweighting interacts with distributional uncertainty and alternative variational mutual-information bounds. Such extensions could provide a unified theoretical account of negative correction across geometric, contrastive, and probabilistic alignment objectives.