MULTIMODAL REPRESENTATION ALIGNMENT BASED ON SPHERICAL CONCENTRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal foundation models aim to align heterogeneous modalities into a shared embedding space. Beyond instance-level retrieval, the practical value of such a space also lies in how well it is organized: whether matched modalities are truly aligned without collapsing into a degenerate region, and whether the resulting space transfers to downstream tasks such as clustering and classification. We propose SCORE, a multi-vector similarity that measures the spherical concentration of normalized modality embeddings through a weighted resultant vector. Unlike similarities based on higher-order geometric constructions, SCORE aggregates modalities additively, providing a simple closed-form score for any available subset of modalities and continually pulling every modality toward a shared consensus direction. This additive design yields a better-organized shared space. Across video-text benchmarks, SCORE consistently reduces the modality gap, including non-anchor pairs, without representation collapse, and improves clustering and classification by larger margins than the corresponding retrieval margins. At the same time, SCORE supports variable-modality inference, using the same embeddings for both pairwise and joint retrieval, where it remains competitive with or ahead of recent multi-vector similarities (e.g., 55.8 R@1 for text-to-video/audio/subtitles and 54.0 R@1 for text-to-video on MSR-VTT). Code available at: `https://anonymous.4open.science/r/SCORE-A20F`.

1 INTRODUCTION

Multimodal foundation models aim to align heterogeneous modalities (e.g., vision, audio, language) into a shared embedding space so that semantically matched inputs are close and mismatched ones are far apart. Contrastive learning with cosine similarity, popularized by CLIP (Radford et al., 2021), has been widely adopted for two-modality alignment and extended to multiple modalities by selecting an anchor modality and aligning other modalities to it (Girdhar et al., 2023; Ruan et al., 2023; Zhu et al., 2024).

Recent work goes beyond pairwise cosine similarity by introducing multi-vector similarities that directly score a set of modality embeddings, enabling joint multimodal retrieval without explicit fusion layers. GRAM (Cicchetti et al., 2025b) aligns K modalities by minimizing the volume of the K-parallelotope spanned by the embeddings (via the Gram determinant), while TRIANGLE (Cicchetti et al., 2025a) focuses on the $K = 3$ case and uses the triangle area as an interpretable alignment measure. These higher-order similarities have demonstrated strong performance when multiple modalities are available at inference time.

While retrieval is the most common yardstick for such models, the practical value of a shared multimodal space is broader. It should be well-organized. Concretely, (i) matched modalities of the same instance should be meaningfully aligned according to shared semantics; (ii) this alignment should not be achieved by representation collapse, i.e., embeddings within a modality should remain spread across the space; and (iii) the space should preserve semantic neighborhood structure, capturing not only instance-level matches but also the relations among semantically related instances, so that the same embeddings transfer to downstream uses such as clustering and classification. These properties jointly characterize the quality of the shared space, yet they are only partially reflected by retrieval R@1: for instance, a modality gap can persist even when instance-level retrieval is strong (Liang et al., 2022), and it can degrade group-wise organization such as clustering (Grassucci et al., 2026).

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.

To promote such a well-organized space, we propose SCORE (**S**pherical **CO**ncentration of **RE**presentations), a multi-vector similarity that measures the spherical concentration of normalized modality embeddings via a weighted resultant vector. Given normalized modality embeddings, SCORE measures agreement via the length of a (weighted) resultant vector, inspired by the mean resultant length (MRL) and its connection to spherical variance in directional statistics (Mardia & Jupp, 2000). Unlike similarities that rely on higher-order geometric constructions, SCORE aggregates modality embeddings *additively* through a resultant vector, so that the score depends on all modalities jointly and keeps pulling each of them, including non-anchor pairs, toward a shared consensus direction. This yields a simple closed-form score that applies to any available subset of modalities through a single functional form, enabling the same embeddings to be reused for both pairwise and joint retrieval under variable-modality inference. As we show, this additive design also encourages a better-organized shared space. In addition, to improve flexibility when multiple modalities are combined, we introduce instance-pair-adaptive weights that modulate modality contributions to the resultant vector, while preserving multi-way alignment.

Empirically, SCORE improves the organization of the learned space while remaining comparable to, or slightly better than, recent multi-vector similarities in retrieval. Across video-text benchmarks, it reduces the modality gap on all pairs (including non-anchor audio-video) without representation collapse, and improves clustering and classification on MSR-VTT, while remaining competitive in retrieval under variable-modality inference (reaching 55.8 R@1 for text-to-video/audio/subtitles and 54.0 R@1 for text-to-video on MSR-VTT).

Our main contributions can be summarized as follows:

1. We introduce SCORE, an additive multi-vector similarity that measures spherical concentration, enabling a single representation space to support both pairwise and joint retrieval under variable-modality inference. We analyze its properties, showing that additive aggregation is maximized only under genuine multi-way consensus, whereas multiplicative dispersion logits can be maximized by partial agreement of a modality subset.
2. We propose an instance-pair-adaptive weighting mechanism that increases the expressivity of the joint score, using bounded weights to prevent collapse and preserve multi-way alignment.
3. We show empirically that additive aggregation yields a better-organized shared multimodal space: it reduces the modality gap across all modality pairs (including non-anchor ones) without representation collapse, and improves downstream clustering and classification, while remaining comparable in retrieval.

2 Related Work

Contrastive alignment and scaling beyond two modalities. CLIP (Radford et al., 2021) established contrastive learning with cosine similarity for image-text alignment, inspiring extensions to other pairs such as audio-text (Elizalde et al., 2023) and video-text (Luo et al., 2021), as well as improvements in scale and loss design (Jia et al., 2021; Zhai et al., 2023). To go beyond two modalities, many approaches select an anchor modality and align the others to it (Girdhar et al., 2023; Ruan et al., 2023; Zhu et al., 2024), optionally adding fusion layers when three or more modalities are available at inference (Chen et al., 2023b; Xu et al., 2023). Related large-scale video-language and omni-modal models include VALOR (Liu et al., 2024), InternVideo2 (Wang et al., 2024), and VideoPrism (Zhao et al., 2024).

Multi-vector alignment. Beyond anchor-based alignment and fusion, several works score an entire multimodal tuple with a single logit. GRAM (Cicchetti et al., 2025b) uses the volume of the parallelotope spanned by K modality embeddings (via the Gram determinant), and TRIANGLE (Cicchetti et al., 2025a) specializes this to three modalities using the triangle area. These multiplicative dispersion measures enable joint scoring without explicit fusion and perform well in genuinely multimodal settings. In a complementary line, Symile (Saporta et al., 2024) targets higher-order dependence (total correlation), and PMRL (Liu et al., 2026) maximizes the top singular value of the per-instance representation matrix, i.e., the best shared direction across modalities. In contrast, SCORE maximizes the length of a (weighted) resultant vector, yielding a simple closed-form logit and gradients that pull each modality toward the current consensus. Closely related to our moti-

vation, GMC (Poklukar et al., 2022) learns representations robust to missing modalities by aligning each modality-specific representation with a separately encoded joint representation. SCORE instead defines a single closed-form multi-vector logit that directly scores any available modality subset, removing the need for a dedicated joint encoder and naturally supporting variable-modality inference at the score level.

Modality gap and global organization. Contrastive multimodal models can exhibit a structural mismatch where representations remain partially separated by modality even after training, known as the modality gap (Liang et al., 2022). While such a mismatch may be less visible in instance-wise retrieval, it can substantially affect the global (group-wise) organization of the shared space, e.g., clustering (Grassucci et al., 2026). More broadly, these phenomena relate to the geometry of contrastive learning, including the interplay between alignment and uniformity (Wang & Isola, 2020). Motivated by this perspective, we explicitly analyze how additive versus multiplicative aggregation shapes the modality gap and the global organization of the learned multimodal space.

3 PRELIMINARIES

3.1 SETUP AND NOTATION

We consider K modalities and a shared embedding dimension d. We train with mini-batches of size B, indexing instances by $i \in \{1, \dots, B\}$. For each instance, we observe one anchor modality and $K-1$ non-anchor modalities. Let

$$\mathbf{a}_i \in \mathbb{S}^{d-1}, \qquad \mathbf{m}_{2i}, \dots, \mathbf{m}_{Ki} \in \mathbb{S}^{d-1} \tag{1}$$

denote the normalized embeddings of the anchor and the non-anchor modalities, where $\mathbb{S}^{d-1} := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$ is the unit hypersphere. At inference time, only a subset of modalities may be available. We therefore aim to learn a shared space in which the same embeddings support both pairwise retrieval (two modalities) and joint retrieval (three or more) by scoring the available subset.

3.2 ANCHOR-FIXED BIDIRECTIONAL INFONCE (D2A AND A2D)

Following anchor-fixed multi-vector objectives, we construct negatives by replacing either (i) the anchor while keeping the non-anchor modalities fixed (D2A; data-to-anchor), or (ii) all non-anchor modalities while keeping the anchor fixed (A2D; anchor-to-data). We optimize both directions and average the two losses to symmetrize training and promote consistent alignment in both retrieval directions.

Let $\ell(\cdot) \in \mathbb{R}$ be a multi-vector logit function on a tuple $(\mathbf{a}, \mathbf{m}_2, \dots, \mathbf{m}_K)$, and let τ denote the learnable temperature. We have:

$$\mathcal{L}_{\text{D2A}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\ell(\mathbf{a}_i, \mathbf{m}_{2i}, \dots, \mathbf{m}_{Ki})/\tau)}{\sum_{j=1}^{B} \exp(\ell(\mathbf{a}_j, \mathbf{m}_{2i}, \dots, \mathbf{m}_{Ki})/\tau)}. \tag{2}$$

$$\mathcal{L}_{\text{A2D}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\ell(\mathbf{a}_i, \mathbf{m}_{2i}, \dots, \mathbf{m}_{Ki})/\tau)}{\sum_{j=1}^{B} \exp(\ell(\mathbf{a}_i, \mathbf{m}_{2j}, \dots, \mathbf{m}_{Kj})/\tau)}. \tag{3}$$

$$\mathcal{L}_{\text{bi}} = \frac{1}{2}\left(\mathcal{L}_{\text{D2A}} + \mathcal{L}_{\text{A2D}}\right). \tag{4}$$

4 METHOD

Prior multi-vector objectives quantify multimodal agreement via geometric dispersion. TRIANGLE (Cicchetti et al., 2025a) uses triangle area (for $K = 3$), and GRAM (Cicchetti et al., 2025b) uses parallelotope volume for general K. In contrast, we quantify agreement via spherical concentration on the unit hypersphere, using a weighted generalization of the mean resultant length (MRL), which is closely related to spherical variance in directional statistics (Mardia & Jupp, 2000). This yields an additive-aggregation-based multi-vector logit defined for any set of K modality embeddings.

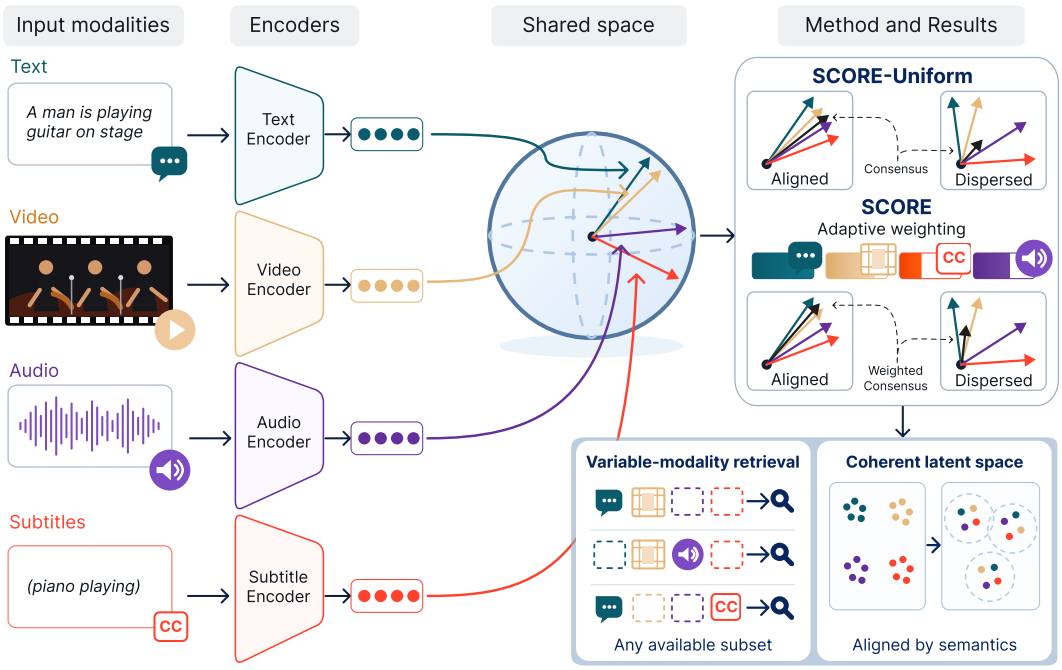


Figure 1: **SCORE as spherical concentration.** SCORE measures multimodal agreement by the length of a (weighted) resultant vector on the unit hypersphere. When modality embeddings concentrate around a common direction, the resultant is long ($R \approx 1$, variance $V = 1 - R \approx 0$); when they disagree, the resultant shortens.

Intuition: concentration on the hypersphere. Consider K unit vectors $\{\mathbf{z}_k\}_{k=1}^K$ on the hypersphere. The more they agree in direction, the larger the norm of their sum. Disagreement causes partial cancellation and shortens the resultant. Directional statistics formalizes this via the mean resultant length (Mardia & Jupp, 2000):

$$R(\{\mathbf{z}_k\}_{k=1}^K) = \left\| \frac{1}{K} \sum_{k=1}^{K} \mathbf{z}_k \right\|_2 \in [0, 1]. \tag{5}$$

A common dispersion measure is the spherical variance $V = 1 - R$, which is 0 when all vectors are perfectly aligned and increases as they spread out. We define SCORE as the length of a (weighted) resultant vector, generalizing the MRL by allowing non-uniform modality contributions to increase the expressivity of the joint score. Figure 1 illustrates this intuition.

4.1 SCORE: Weighted Spherical Concentration

Given a mixed tuple $(\mathbf{a}_i, \mathbf{m}_{2j}, \ldots, \mathbf{m}_{Kj})$ with normalized embeddings and weights $\{w_{k,ij}\}_{k=1}^K$ satisfying $w_{k,ij} \geq 0$ and $\sum_{k=1}^K w_{k,ij} = 1$, we define SCORE via a weighted resultant on the hypersphere:

$$\boxed{\mathbf{r}_{ij} = w_{1,ij}\, \mathbf{a}_i + \sum_{k=2}^{K} w_{k,ij}\, \mathbf{m}_{kj},} \tag{6}$$

$$\boxed{\ell_{\text{SCORE}}(\mathbf{a}_i, \mathbf{m}_{2j}, \ldots, \mathbf{m}_{Kj}) = \|\mathbf{r}_{ij}\|_2.} \tag{7}$$

Since $\mathbf{r}_{ij}$ is a convex combination of unit vectors, $\ell_{\text{SCORE}} \in [0, 1]$. Uniform weights yield a simple, parameter-free score. However, treating all modalities equally may be limiting when their relevance varies across instances, and we empirically observe that adaptive weighting improves joint retrieval (Section 5). We therefore consider adaptive weights that modulate each modality contribution, while constraining them to avoid collapse (Section 4.3).

Score Uniform Setting $w_{k,ij} = 1/K$ for all k, SCORE reduces to the mean resultant length (MRL):

$$\ell_{\text{SCORE-U}}(\mathbf{a}_i, \mathbf{m}_{2j}, \ldots, \mathbf{m}_{Kj}) = \left\| \frac{\mathbf{a}_i + \sum_{k=2}^{K} \mathbf{m}_{kj}}{K} \right\|_2 . \tag{8}$$

We use either ℓ_{SCORE} or $\ell_{\text{SCORE-U}}$ as the multi-vector logit $\ell(\cdot)$ in Eqs. (2-4) and optimize the same anchor-fixed bidirectional InfoNCE objective with a learnable temperature τ.

4.2 Why resultant length? Properties of SCORE

SCORE is designed to (i) apply to any subset of modalities through a single functional form (the norm of a weighted sum) without higher-order geometric constructions, and (ii) reward genuine multi-way consistency when three or more modalities are present.

One score for any modality subset (variable-modality inference). Given any subset of available modalities, SCORE forms the corresponding (weighted) resultant and returns its length. The same rule thus applies to pairwise retrieval (e.g., T-V, T-A) and joint retrieval (e.g., T-V-A-S) without ad-hoc fusion of pairwise similarities.

Rewarding multi-way consistency. Since ℓ_{SCORE} is the length of a weighted mean direction, it increases as embeddings concentrate around a shared consensus direction. It admits the closed-form expansion

$$\ell_{\text{SCORE}}^2 = \sum_{k=1}^{K} w_k^2 + 2\sum_{k=2}^{K} w_1 w_k \langle \mathbf{a}, \mathbf{m}_k \rangle + 2 \sum_{2 \le p < q \le K} w_p w_q \langle \mathbf{m}_p, \mathbf{m}_q \rangle, \tag{9}$$

which shows that SCORE depends on *all* pairwise agreements within the tuple, including non-anchor pairs.

Additive aggregation shapes a better-organized space. This additive structure is the key design choice behind SCORE. By the triangle inequality, $\|\mathbf{r}\|_2 \le \sum_k w_k \|\mathbf{z}_k\|_2 = 1$, with equality *iff* all modality embeddings with positive weight are codirectional. Thus ℓ_{SCORE} is maximized only under genuine multi-way agreement. In contrast, multiplicative dispersion logits (triangle area or parallelotope volume) are maximized whenever any two embeddings coincide, regardless of the others, so they can be satisfied by partial agreement. Because SCORE keeps pulling every modality, including non-anchor pairs, toward a shared consensus direction, it reduces the modality gap across all pairs and promotes a better-organized shared space (Section 5.1). We formalize these geometric properties and their consequences for anchor-fixed contrastive training in Appendix C.

4.3 Adaptive weighting

For each compared pair (i, j), we predict non-negative pre-weights that modulate each modality contribution in the weighted resultant (Eq. 6). To ensure that no modality is ignored and to preserve multi-way alignment (crucial for variable-modality inference), we constrain the predicted pre-weights to a bounded range before normalization.

We form the tuple feature $\mathbf{a}_{ij} = [\mathbf{a}_i; \mathbf{m}_{2j}; \ldots; \mathbf{m}_{Kj}] \in \mathbb{R}^{Kd}$ and compute raw logits for the non-anchor modalities via an MLP g_ϕ: $\tilde{\mathbf{w}}_{ij} = g_\phi(\mathbf{a}_{ij}) \in \mathbb{R}^{K-1}$.

Bounded non-anchor weights. Unbounded weights admit degenerate patterns that effectively ignore one or more modalities, reducing the objective to a lower-modality one and destabilizing InfoNCE logits. We fix the anchor pre-weight to one to remove a global scaling ambiguity, $\hat{w}_{1,ij} = 1$, and map the remaining outputs $(\tilde{w}_{2,ij}, \ldots, \tilde{w}_{K,ij})$ to bounded pre-weights:

$$\hat{w}_{k,ij} = \text{clip}(1 + \tilde{w}_{k,ij},\, w_{\min}, w_{\max}), \qquad k \in \{2, \ldots, K\}, \tag{10}$$

with $0 \le w_{\min} \le 1 \le w_{\max}$. These bounds are chosen for balanced performance across pairwise and joint retrieval, not to maximize any single metric (Appendix E).

Normalized resultant. Given normalized embeddings $\{\mathbf{z}_{k,ij}\}_{k=1}^{K}$, we compute

$$\mathbf{r}_{ij} = \sum_{k=1}^{K} \hat{w}_{k,ij} \mathbf{z}_{k,ij}, \qquad W_{ij} = \sum_{k=1}^{K} \hat{w}_{k,ij}, \qquad \tilde{\ell}_{ij} = \frac{\|\mathbf{r}_{ij}\|_2}{W_{ij}}. \tag{11}$$

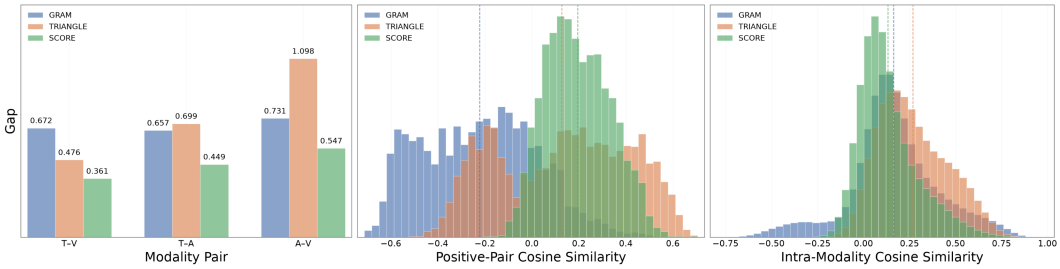


Figure 2: Multimodal embedding representation metrics among GRAM, TRIANGLE, and the proposed SCORE on the MSR-VTT dataset. Left: modality gap metric, with SCORE obtaining the lowest value across cross-modal pairs. Middle: cosine similarity between positive pairs (T-V, T-A, and V-A); higher is better. The cosine distribution of SCORE has the highest mean among the methods. Right: intra-modal cosine similarity, which is lower when embeddings of the same modality remain spread across the space and do not collapse into modality clusters. In this case, SCORE has an average close to 0.0, still supporting our interpretation that SCORE promotes a more coherent shared space without representation collapse.

Dividing by W_{ij} is algebraically equivalent to using normalized weights $w_{k,ij} = \hat{w}_{k,ij}/W_{ij}$ (so that $w_{k,ij} \geq 0$ and $\sum_k w_{k,ij} = 1$), but yields a simpler computation graph since the weights need not be materialized explicitly. Together with $\hat{w}_{1,ij} = 1$, the clipping ensures $W_{ij} > 0$, so $\tilde{\ell}_{ij}$ is well-defined.

Initialization. We initialize the last layer of g_ϕ to output zeros, so that $\tilde{w}_{k,ij} = 0$ and hence $\hat{w}_{k,ij} = 1$ (Eq. 10). This recovers uniform weights, $\tilde{\ell}_{ij} = \|\frac{1}{K}\sum_{k=1}^{K} \mathbf{z}_{k,ij}\|_2$, providing a conservative starting point before the module learns to modulate modality contributions.

4.4 Final objective

Following prior multimodal pretraining works (Chen et al., 2023b; Cicchetti et al., 2025b;a), we optionally add a data-text matching (DTM) loss to stabilize training:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{bi}} + \lambda \mathcal{L}_{\text{DTM}}, \tag{12}$$

where $\lambda \geq 0$ is a tunable hyperparameter and $\mathcal{L}_{\text{bi}}$ is the anchor-fixed bidirectional InfoNCE objective in Eq. 4. DTM is a widely adopted, binary cross-modal matching objective that predicts, via cross-attention, whether the text and the remaining modalities come from the same instance, acting as an implicit regularizer for the multimodal embeddings (Li et al., 2021). We provide the full formulation, negative construction, and hard-negative mining in Appendix F.

4.5 Computational cost

SCORE computes a weighted sum and a normalization, requiring $O(Kd)$ operations per logit (plus a lightweight MLP if adaptive weights are used). In contrast, GRAM and PMRL logits typically form a $K \times K$ Gram matrix ($O(K^2 d)$) and require determinant- or SVD-related computations (often $O(K^3)$), so SCORE scales more smoothly with the number of modalities. In contrastive training, all methods evaluate $O(B^2)$ logits per batch; when adaptive weights are enabled, SCORE additionally predicts weights per compared pair. We report end-to-end wall-clock time and memory in Section 5.3.

5 Experiments

We evaluate SCORE from two complementary angles. First, we assess the organization of the learned shared space: the modality gap, the absence of representation collapse, and downstream clustering/classification (Section 5.1). Second, we verify that this well-organized space remains

effective for retrieval under variable-modality inference (Sections 5.2-5.2), and is computationally efficient (Section 5.3).

Setup. We finetune three encoders (BERT-Base (Devlin et al., 2019) for text, EvaCLIP-ViT-G (Sun et al., 2023) for video, BEATs (Chen et al., 2023a) for audio) on a 150k subset of VAST27M and evaluate clustering, classification, and zero-shot text-video retrieval. Furthermore, we perform zero-shot retrieval on MSR-VTT Xu et al. (2016), DiDeMo (Hendricks et al., 2017), ActivityNet (Caba Heilbron et al., 2015), and VATEX (Wang et al., 2019), following the recipe of GRAM, TRIANGLE, and PMRL (Cicchetti et al., 2025b;a; Liu et al., 2026) for fair comparison. We use 8 uniformly sampled frames and a 10s audio clip, and report R@1 for T2V and V2T (details in Appendix D). Additionally, we select a video-text-based benchmark, MSR-VTT, and an audio-text benchmark, VGG-Sound+ (Cho et al., 2025; Cho, 2024), as representative datasets of these two tasks, and further train all the encoders in a configuration from scratch. Specifically, on VGG-Sound+ we start from pretrained unimodal encoders (without intermediate multimodal pretraining) and optimize all parameters end-to-end. Full setups (datasets, encoders, hyperparameters) are in Appendix D.

Throughout our experimental validation, we validate the organization of the learned latent space through ad-hoc metrics and group-wise tasks, multimodal retrieval (video-text-based and audio-text-based benchmarks, and variable-modality settings), and the computational efficiency of the proposed SCORE method.

5.1 Organization of the learned shared space

In this section, we prove our claim that additive aggregation yields a better-organized shared space. We assess this with three complementary measures on the from-scratch MSR-VTT models: (i) the modality gap, the average distance between the centroids of different modalities. The lower the gap, the more the space is organized according to semantic content rather than by modality; (ii) the mean intra-modality cosine similarity, the average within-modality similarity across instances, where lower values indicate that embeddings of the same modality remain spread rather than collapsed into modality-wise clusters; and (iii) downstream clustering and classification over the 20 semantic classes of MSR-VTT. Figure 2 shows that SCORE consistently reduces the modality gap across all pairs relative to the other multi-vector models. The improvement on the audio-video pair is especially notable, as it is a non-anchor pair that anchor-based objectives do not optimize directly, consistent with SCORE's additive design, which depends on all pairwise agreements (Eq. 9). In addition, positive-pair cosine similarity highlights that matching pairs have, on average, higher cosine similarity in SCORE than in other methods. Therefore, matching pairs are closer in the learned space. At the same time, the low intra-modality cosine similarities indicate that modality representations do not collapse into single points but remain spread across the space, mitigating the creation of modality clusters. Embeddings remain spread across the space. To complement the latent space organization analysis, we report group-wise task results in Tab. 1. On both clustering and linear probing, Table 1 shows that SCORE outperforms GRAM and TRIANGLE according to all V-Measure, K-NN accuracy, and accuracy metrics. We do not evaluate PMRL here as there are no released checkpoints. Together, Figure 2 and Table 1 provide direct and quantitative evidence that SCORE improves the global organization of the shared space with respect to multiplicative-based multi-vector models.

Table 1: Group-wise tasks on the MSR-VTT dataset. We evaluate the clustering task (using V-Measure and K-NN) and the linear probing task with accuracy among the three multi-vector models.

Method	V-Measure ↑	K-NN ↑	Acc ↑
GRAM	16.76	29.94	37.28
TRIANGLE	17.89	30.50	35.59
SCORE	**20.73**	**38.98**	**38.41**

5.2 Zero-shot multimodal retrieval

Retrieval on multimodal video-text benchmarks. Table 2 shows that SCORE is strong in joint retrieval, reaching 55.8/53.3 (T2V/V2T) R@1 on MSR-VTT (T-VAS) and the best results on VATEX (84.8/83.6). On DiDeMo and ActivityNet, it remains competitive with prior multi-vector

Table 2: Zero-shot multimodal text-to-video (T2V) and video-to-text (V2T) retrieval R@1 results.

		MSR-VTT		DiDeMo		ActivityNet		VATEX	
	Modality	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
Cosine-based									
UMT (Liu et al., 2022)	T-V	33.3	-	34.0	-	31.9	-	-	-
OmniVL (Wang et al., 2022a)	T-V	34.6	-	33.3	-	-	-	-	-
UMT-L (Li et al., 2023)	T-V	40.7	37.1	48.6	49.9	41.9	39.4	-	-
TVTSv2 (Zeng et al., 2023)	T-V	38.2	-	34.6	-	-	-	-	-
ViCLIP (Wang et al., 2023)	T-V	42.4	41.3	18.4	27.9	15.1	24.0	-	-
VideoCoCa (Yan et al., 2022)	T-V	34.3	64.7	-	-	34.5	33.0	53.2	73.6
ImageBind (Girdhar et al., 2023)	T-V	36.8	-	-	-	-	-	-	-
InternVideo-L (Wang et al., 2022b)	T-V	40.7	39.6	31.5	33.5	30.7	31.4	49.5	69.5
HiTeA (Ye et al., 2022)	T-V	34.4	-	43.2	-	-	-	-	-
mPLUG-2 (Xu et al., 2023)	T-V	47.1	-	45.7	-	-	-	-	-
VideoPrism-b (Zhao et al., 2024)	T-V	51.4	50.2	-	-	49.6	47.9	62.5	77.1
LanguageBind (Zhu et al., 2024)	T-V	44.8	40.9	39.9	39.8	41.0	39.1	-	-
Multi-vector									
VAST (Chen et al., 2023b)	T-VA	49.3	43.7	49.5	48.2	51.4	46.8	80.0	77.3
VAST (Chen et al., 2023b)	T-VAS	50.7	49.0	-	-	-	-	82.1	78.7
TRIANGLE (Cicchetti et al., 2025a)	T-VA	55.2	52.5	53.1	52.5	<u>57.9</u>	**53.1**	<u>83.9</u>	80.9
GRAM (Cicchetti et al., 2025b)	T-VAS	54.8	52.9	**54.2**	52.2	**58.0**	50.4	83.5	<u>82.7</u>
PMRL (Liu et al., 2026)	T-VAS	54.5	52.4	50.6	48.4	56.0	49.6	80.5	75.2
SCORE-Uniform	T-VAS	53.5	51.9	50.3	48.9	52.2	46.2	82.9	81.7
SCORE (ours)	T-VAS	**55.8**	**53.3**	<u>53.7</u>	**52.7**	56.6	<u>51.5</u>	**84.8**	**83.6**

Table 3: Retrieval performance on MSR-VTT with a different number of available modalities. For each baseline (but PMRL having only one entry), we compute paired deltas over the comparable entries of this table and assess them with a one-sided exact Wilcoxon signed-rank test; all improvements over the baselines are statistically significant ($p < 0.05$).

Method	T-V		T-A		T-AV		T-VAS	
	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
VAST	50.5	80.5	4.0	20.6	49.3	80.0	50.7	80.5
GRAM	<u>52.8</u>	82.9	5.4	22.4	54.2	83.9	<u>54.8</u>	<u>82.9</u>
TRIANGLE	50.9	82.7	5.0	23.4	**55.2**	<u>84.2</u>	-	-
PMRL	-	-	-	-	-	-	54.5	-
SCORE-Uniform	52.1	<u>83.3</u>	<u>5.9</u>	<u>23.8</u>	53.1	79.6	53.5	80.4
SCORE (ours)	**54.0**	**83.8**	**7.2**	**23.9**	<u>55.1</u>	**84.8**	**55.8**	**85.3**

similarities. SCORE also consistently improves over SCORE-Uniform, supporting the benefit of instance-pair-adaptive weighting.

Variable-modality retrieval. Using the same pretrained checkpoints as Table 2, we score different modality subsets at test time: pairwise (T-V, T-A) and joint (T-AV, T-VAS), reporting R@1/R@10. Table 3 shows that SCORE achieves the best or comparable R@1 across all subsets, improving both pairwise and joint retrieval (e.g., 54.0 R@1 on T-V, 55.8 on T-VAS, and 7.2 on T-A vs. 4.0 for VAST). This confirms that the same embeddings, learned for a well-organized joint space, remain effective when only a subset of modalities is available. Compared to SCORE-Uniform, adaptive weighting yields consistent gains, especially in joint retrieval (e.g., T-AV R@10: 84.8 vs. 79.6).

From-scratch training across scales. We evaluate T2V, T2A, and joint T2AV retrieval. Pairwise baselines (Text-Vision Only, Text-Audio Only) use cosine InfoNCE on the corresponding pair; Text-Audio + Text-Vision averages the two pairwise losses with shared text embeddings. Table 4 shows that SCORE achieves the best overall average on both datasets, driven by improved joint retrieval, and outperforms the strong Text-Audio + Text-Vision baseline that averages pairwise (hub-style) objectives. The strongest T2A result comes from the Text-Audio-only model: on these datasets, audio is a less reliable cue for joint alignment but remains useful for audio-only retrieval, so our

Table 4: Retrieval results on MSR-VTT and VGG-sound+, with average performance (AVG).

Method	MSR-VTT (from scratch)							VGG-sound+ (from scratch)						
	T2V		T2A		T2AV		AVG	T2V		T2A		T2AV		AVG
	R@1	R@10	R@1	R@10	R@1	R@10		R@1	R@10	R@1	R@10	R@1	R@10	
Text-Vision Only	31.6	**74.4**	-	-	-	-	-	**58.3**	**90.7**	-	-	-	-	-
Text-Audio Only	-		**7.5**	**34.7**	-	-	-	-	-	**13.9**	51.4	-	-	-
Text-Audio + Text-Vision	31.0	65.7	4.4	30.5	31.2	72.9	39.3	57.4	90.4	12.0	52.7	57.6	90.6	59.1
GRAM	29.8	63.0	5.1	24.9	37.9	76.1	39.5	57.7	88.0	11.6	50.2	56.3	90.9	59.1
TRIANGLE	30.1	64.1	4.8	24.7	37.6	74.0	39.2	58.5	85.8	10.2	44.7	57.0	90.6	57.8
SCORE-Uniform (ablation)	30.8	72.2	6.2	30.3	38.0	76.5	42.3	58.0	88.7	13.8	**52.9**	56.6	89.9	60.0
SCORE (ours)	**32.1**	74.2	5.1	28.5	**38.9**	**81.5**	**43.4**	57.9	89.1	13.0	51.5	**59.5**	**91.4**	**60.4**

adaptive weighting downweights audio when combining modalities. This improves joint retrieval at a small cost to the audio-only direction, reflecting the trade-off between optimizing a single pair and learning representations effective under variable-modality inference (analysis in Appendix G).

5.3 Computational Efficiency

Figure 3 compares the wall-clock time and memory of SCORE against cosine-based and multi-vector methods as the number of modalities grows. SCORE remains stable in both time and memory, whereas cosine-based and multi-vector methods grow markedly with the number of modalities, with PMRL being the slowest due to its SVD computation. This smoother scaling makes SCORE well-suited for large-scale multimodal settings, where efficiency becomes a critical bottleneck, and is consistent with the complexity analysis in Section 4.5.

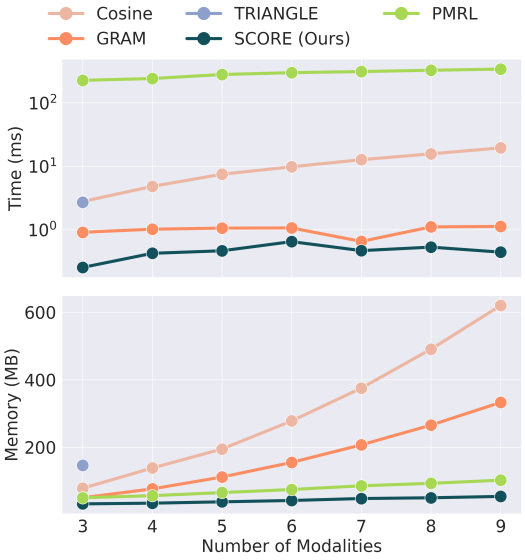


Figure 3: Computational analysis. SCORE is the most efficient among the cosine-based and multi-vector methods, scaling smoothly with the number of modalities.

6 Conclusion

We presented SCORE, an additive-aggregation-based multi-vector similarity that quantifies the spherical concentration of multimodal embeddings via the length of a (weighted) resultant vector. Because SCORE aggregates all modalities additively, its score depends on every pairwise agreement within a tuple and keeps pulling each modality, including non-anchor pairs, toward a shared consensus direction. Both analytically and empirically, we showed that this design yields a better-organized shared space: SCORE reduces the modality gap across all modality pairs without representation collapse, and improves downstream clustering and classification by margins substantially larger than the corresponding retrieval margins. We contrasted this with multiplicative dispersion logits, which can be maximized by partial agreement of a modality subset. At the same time, SCORE provides a simple closed-form score for any available subset of modalities, enabling the same embeddings to be reused for both pairwise and joint retrieval under variable-modality inference, where it remains competitive with or ahead of recent multi-vector similarities while scaling efficiently with the number of modalities. We further introduced instance-pair-adaptive weighting to increase the flexibility of joint scoring while constraining weights to prevent collapse and preserve multi-way alignment. Future work includes exploring richer weighting mechanisms, scaling to larger modality sets, and extending SCORE beyond retrieval to other multimodal tasks that benefit from a well-organized shared space.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance, i.e., polishing and proofreading the text (grammar and formatting). We have not used generative AI tools for research ideation or methodology design. These were carried out entirely by the authors. Retrieval of related work and literature review were conducted manually by the authors. All technical claims, experimental results, and artifacts were produced by the authors. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

We have reviewed the ICLR Code of Ethics and believe our work conforms to it. Our experiments use only publicly available benchmark datasets under their respective licenses and standard evaluation protocols; we do not collect new personal data or conduct research involving human subjects. We properly credit the creators of all datasets, code, and models used in this work by citing the original sources. Our contribution is a similarity objective for multimodal representation learning; we do not release any high-risk model or new dataset, so no additional safeguards beyond the responsible use of the referenced assets are required. As with retrieval and representation learning in general, improved cross-modal alignment could be misused in downstream applications such as large-scale content search or surveillance. While our work does not introduce new data or deployment-specific mechanisms, we acknowledge these potential downstream risks. On the positive side, our method can improve robustness to missing or noisy modalities and reduce compute overhead, lowering the cost of deploying multimodal retrieval systems.

REPRODUCIBILITY STATEMENT

We have made several efforts to ensure the reproducibility of our results. The SCORE similarity and the adaptive weighting mechanism are fully specified in Section 4 (Eqs. 6-11), and the training objective is defined in Section 3 and Section 4.4. Formal statements, assumptions (ℓ_2-normalized embeddings and non-negative bounded weights), and complete proofs of the analytical properties of SCORE (including the closed-form expansion) are provided in Appendix A. Full experimental details for both the vanilla and large-scale settings, including dataset splits, preprocessing, encoders, hyperparameters (and how they were chosen), optimizer, and hardware, are given in Appendix D. The statistical significance protocol used in Table 3 is described in its caption. Anonymized source code and hyperparameters to reproduce our experiments are available at the repository linked in the abstract (`https://anonymous.4open.science/r/SCORE-A20F`).

REFERENCES

Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. AudioMNIST: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 361(1):418–428, 2024. doi: 10.1016/j.jfranklin.2023.11.038.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

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 *International Conference on Machine Learning (ICML)*, pp. 5178–5193, 2023a.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Ming-Ting Sun, Xinxin Zhu, and J. Liu. VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Neural Information Processing Systems (NeurIPS)*, 2023b.

Sangyeon Cho. vgg_sound_plus. IEEE DataPort, 2024. URL `https://ieee-dataport.org/documents/vggsoundplus`.

Sangyeon Cho, Jangyeong Jeon, Mingi Kim, and Junyeong Kim. Synergy-CLIP: Extending clip with multi-modal integration for robust representation learning. *IEEE Access*, 13:65630–65642, 2025. doi: 10.1109/ACCESS.2025.3559663.

Giordano Cicchetti, Eleonora Grassucci, and Danilo Comminiello. A TRIANGLE enables multimodal alignment beyond cosine similarity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *International Conference on Learning Representations (ICLR)*, 2025b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186, 2019.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740, 2020.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

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

Eleonora Grassucci, Giordano Cicchetti, Emanuele Frasca, Aurelio Uncini, and Danilo Comminiello. Closing the modality gap aligns group-wise semantics. In *International Conference on Learning Representations (ICLR)*, 2026.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

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

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Colored-MNIST, 2025. URL https://dx.doi.org/10.21227/sxtx-qh90.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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 Transactions on Pattern Analysis and Machine Intelligence*, 47(2):708–724, 2024.

Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Principled multimodal representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. doi: 10.1109/TPAMI.2026.3675685.

Ye Liu, Siyuan Li, Yang Wu, Chang Wen Chen, Ying Shan, and Xiaohu Qie. UMT: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3042–3051, 2022.

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. *Neurocomputing*, 2021.

Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. Wiley, 2000.

Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S. Melo, Ana Paiva, and Danica Kragic. Geometric multimodal contrastive representation learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 17782–17800, 2022.

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 International Conference on Machine Learning (ICML)*, 2021.

Ludan Ruan, Anwen Hu, Yuqing Song, Liang Zhang, S. Zheng, and Qin Jin. Accommodating audio modality in CLIP for multimodal processing. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

Armand Saporta, A. M. Puli, Mark Goldstein, and Rajesh Ranganath. Contrasting with Symile: Simple model-agnostic representation learning for unlimited modalities. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

John L. Singleton. Stroke based MNIST data. 2016. URL https://zenodo.org/records/201035.

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

Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. OmniVL: One foundation model for image-language and video-language tasks. In *Advances in Neural Information Processing Systems*, 2022a.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Xin Wang, Jianwei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. VATEX: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 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. *ArXiv preprint: arXiv:2212.03191*, 2022b.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Jian Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Y. Qiao. InternVid: A large-scale video-text dataset for multimodal understanding and generation. *ArXiv preprint: arXiv:2307.06942*, 2023.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 396–416. Springer, 2024.

Haiyang Xu, Qinghao Ye, Mingshi Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qiuchen Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Feiran Huang, and Jingren Zhou. mPLUG-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning (ICML)*, 2023.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. VideoCoCa: Video-text modeling with zero-shot transfer from contrastive captioners. *ArXiv preprint: arXiv:2212.04979*, 2022.

Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. HiTeA: Hierarchical temporal-aware video-language pre-training. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15359–15370, 2022.

Ziyun Zeng, Yixiao Ge, Zhan Tong, Xihui Liu, Shutao Xia, and Ying Shan. TVTSv2: Learning out-of-the-box spatiotemporal visual representations at scale. *ArXiv preprint: arXiv:2305.14173*, 2023.

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

Long Zhao, Nitesh Bharadwaj Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J. Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, Rachel Hornung, Florian Schroff, Ming Yang, David A. Ross, Huisheng Wang, Hartwig Adam, Mikhail Sirotenko, Ting Liu, and Boqing Gong. VideoPrism: A foundational visual encoder for video understanding. In *International Conference on Machine Learning (ICML)*, 2024.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Liejie Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations (ICLR)*, 2024.

A CLOSED-FORM EXPANSION OF THE RESULTANT LENGTH (PROOF)

Proposition 1 (Closed-form expansion). *Let $\mathbf{a}, \mathbf{m}_2, \ldots, \mathbf{m}_K \in \mathbb{S}^{d-1}$ be unit-normalized embeddings and let $w_1, \ldots, w_K$ satisfy $w_k \geq 0$ and $\sum_{k=1}^{K} w_k = 1$. Define*

$$\mathbf{r} = w_1\mathbf{a} + \sum_{k=2}^{K} w_k\mathbf{m}_k, \qquad \ell_{\mathrm{SCORE}} = \|\mathbf{r}\|_2. \tag{13}$$

Then

$$\ell_{\mathrm{SCORE}}^2 = \sum_{k=1}^{K} w_k^2 + 2\sum_{k=2}^{K} w_1 w_k \langle \mathbf{a}, \mathbf{m}_k \rangle + 2 \sum_{2 \leq p < q \leq K} w_p w_q \langle \mathbf{m}_p, \mathbf{m}_q \rangle, \tag{14}$$

which matches Eq. 9 in the main text.

Proof. By definition,

$$\ell_{\mathrm{SCORE}}^2 = \|\mathbf{r}\|_2^2 = \left\langle w_1\mathbf{a} + \sum_{k=2}^{K} w_k\mathbf{m}_k,\ w_1\mathbf{a} + \sum_{t=2}^{K} w_t\mathbf{m}_t \right\rangle. \tag{15}$$

Expanding the inner product yields

$$\ell_{\mathrm{SCORE}}^2 = w_1^2\langle \mathbf{a}, \mathbf{a} \rangle + 2\sum_{k=2}^{K} w_1 w_k \langle \mathbf{a}, \mathbf{m}_k \rangle + \sum_{k=2}^{K}\sum_{t=2}^{K} w_k w_t \langle \mathbf{m}_k, \mathbf{m}_t \rangle. \tag{16}$$

Since $\langle \mathbf{a}, \mathbf{a} \rangle = 1$ and $\langle \mathbf{m}_k, \mathbf{m}_k \rangle = 1$, the last double sum splits into diagonal and off-diagonal terms:

$$\sum_{k=2}^{K}\sum_{t=2}^{K} w_k w_t \langle \mathbf{m}_k, \mathbf{m}_t \rangle = \sum_{k=2}^{K} w_k^2 + 2 \sum_{2 \leq p < q \leq K} w_p w_q \langle \mathbf{m}_p, \mathbf{m}_q \rangle. \tag{17}$$

Combining terms gives the claim. □

B CONNECTIONS

B.1 CONNECTION TO CLIP WHEN $K = 2$

Let $\mathbf{a}, \mathbf{m}_2 \in \mathbb{S}^{d-1}$ be unit-normalized embeddings. When $K = 2$ and uniform weights are used ($w_1 = w_2 = \frac{1}{2}$), Eq. 7 becomes

$$\ell_{\mathrm{SCORE}}(\mathbf{a}, \mathbf{m}_2) = \left\| \frac{\mathbf{a} + \mathbf{m}_2}{2} \right\|_2. \tag{18}$$

Let θ be the angle between $\mathbf{a}$ and $\mathbf{m}_2$, i.e., $\mathbf{a}^\top \mathbf{m}_2 = \cos\theta$. Then

$$\begin{aligned}
\left\| \frac{\mathbf{a} + \mathbf{m}_2}{2} \right\|_2 &= \frac{1}{2}\|\mathbf{a} + \mathbf{m}_2\|_2 && (19)\\
&= \frac{1}{2}\sqrt{\langle \mathbf{a} + \mathbf{m}_2,\ \mathbf{a} + \mathbf{m}_2 \rangle} && (20)\\
&= \frac{1}{2}\sqrt{\langle \mathbf{a}, \mathbf{a} \rangle + 2\langle \mathbf{a}, \mathbf{m}_2 \rangle + \langle \mathbf{m}_2, \mathbf{m}_2 \rangle} && (21)\\
&= \frac{1}{2}\sqrt{\|\mathbf{a}\|_2^2 + 2\,\mathbf{a}^\top \mathbf{m}_2 + \|\mathbf{m}_2\|_2^2} && (22)\\
&= \frac{1}{2}\sqrt{1 + 2\cos\theta + 1} && (23)\\
&= \sqrt{\frac{1 + \cos\theta}{2}} && (24)\\
&= \cos\left(\frac{\theta}{2}\right). && (25)
\end{aligned}$$

B.2 CONNECTION TO PMRL

For a single instance with K modalities, let $\mathbf{z}_1, \ldots, \mathbf{z}_K \in \mathbb{S}^{d-1}$ and define the representation matrix $Z = [\mathbf{z}_1, \ldots, \mathbf{z}_K] \in \mathbb{R}^{d \times K}$. PMRL (Liu et al., 2026) maximizes the top singular value $\sigma_1(Z)$, which admits the variational form

$$\sigma_1(Z) = \max_{\|\mathbf{v}\|_2=1} \|Z\mathbf{v}\|_2. \tag{26}$$

SCORE as an ℓ_1-constrained projection. Consider weights $\{w_k\}_{k=1}^K$ satisfying $w_k \geq 0$ and $\sum_{k=1}^K w_k = 1$, and define the weighted resultant

$$\mathbf{r} = \sum_{k=1}^K w_k \mathbf{z}_k \;=\; Z\mathbf{w}, \qquad \mathbf{w} = (w_1, \ldots, w_K)^\top. \tag{27}$$

SCORE uses the (normalized) resultant length

$$\tilde{\ell} \;=\; \|\mathbf{r}\|_2 \;=\; \|Z\mathbf{w}\|_2, \tag{28}$$

which is scale-invariant with respect to the pre-weights under the constraint $\sum_{k=1}^K w_k = 1$. In our adaptive variant, we implement this normalization implicitly by predicting positive pre-weights $\{\hat{w}_k\}_{k=1}^K$ and computing $\tilde{\ell} = \|\sum_k \hat{w}_k \mathbf{z}_k\|_2 / \sum_k \hat{w}_k$.

Relation to PMRL's best-direction objective. Let $\bar{\mathbf{w}} = \mathbf{w}/\|\mathbf{w}\|_2$ (for $\mathbf{w} \neq \mathbf{0}$). Plugging $\mathbf{v} = \bar{\mathbf{w}}$ into Eq. 26 yields

$$\sigma_1(Z) \;\geq\; \|Z\bar{\mathbf{w}}\|_2 \;=\; \frac{\|Z\mathbf{w}\|_2}{\|\mathbf{w}\|_2} \;=\; \frac{\|\mathbf{r}\|_2}{\|\mathbf{w}\|_2}. \tag{29}$$

Moreover, since $w_k \geq 0$ and $\sum_{k=1}^K w_k = 1$ imply $\|\mathbf{w}\|_2 \leq 1$, we also have the simple upper bound

$$\|Z\mathbf{w}\|_2 \;=\; \|\mathbf{w}\|_2 \, \|Z\bar{\mathbf{w}}\|_2 \;\leq\; \|\mathbf{w}\|_2 \, \sigma_1(Z) \;\leq\; \sigma_1(Z). \tag{30}$$

Therefore, maximizing SCORE with weights $\{w_k\}_{k=1}^K$ satisfying $w_k \geq 0$ and $\sum_{k=1}^K w_k = 1$ optimizes a scale-invariant proxy that is always bounded above by PMRL's best-direction objective $\sigma_1(Z)$, while Eq. 29 shows how any particular choice of weights induces a fixed-direction approximation to PMRL.

Uniform weights (SCORE-Uniform). For uniform weights $w_k = 1/K$, we have $\|\mathbf{w}\|_2 = 1/\sqrt{K}$ and $\mathbf{r} = \frac{1}{K}\sum_{k=1}^K \mathbf{z}_k$. Eq. 29 reduces to

$$\sigma_1(Z) \;\geq\; \sqrt{K} \left\| \frac{1}{K} \sum_{k=1}^K \mathbf{z}_k \right\|_2, \tag{31}$$

recovering the connection used for SCORE-Uniform.

Computational note. Computing a resultant length requires only a weighted sum and a norm, i.e., $O(dK)$ operations per instance. In contrast, objectives that explicitly manipulate singular values typically require an SVD/eigendecomposition of a $d \times K$ matrix (or equivalently an eigendecomposition of the $K \times K$ Gram matrix), which is $O(dK^2 + K^3)$ in general.

Discussion. PMRL optimizes within-instance multi-way alignment by explicitly manipulating the spectrum of the per-instance representation matrix (maximizing its dominant singular value) and complements this objective with instance-wise regularization on the leading directions to avoid collapse. SCORE instead defines a closed-form multi-vector logit based on the length of a (weighted) resultant vector, which is directly applicable to scoring arbitrary modality subsets and naturally admits instance-pair-adaptive weighting at the score level. These differences make SCORE particularly well-suited to variable-modality inference while avoiding SVD-based computations.

From an alignment-measurement perspective, SCORE and PMRL emphasize different notions of multi-way agreement. PMRL maximizes a best-direction objective $\sigma_1(Z) = \max_{\|\mathbf{v}\|_2=1} \|Z\mathbf{v}\|_2$, which allows the maximizing direction to downweight modalities that disagree, and in general is not restricted to non-negative coefficients. Moreover, the spectral objective is invariant to sign

flips of individual modality columns (i.e., $Z \mapsto ZD$ for a diagonal D with ± 1 entries), since $(ZD)^\top(ZD) = D(Z^\top Z)D$ has the same eigenvalues as $Z^\top Z$. As a result, when considering the within-instance spectral term in isolation, the objective does not uniquely fix the orientation of each modality embedding, and different sign conventions can yield the same singular values. In contrast, SCORE uses non-negative weights that sum to one and measures the length of a weighted mean resultant; disagreement across modalities leads to cancellation, so the score more directly rewards multi-way consistency and admits a directional-statistics interpretation in terms of concentration around a consensus direction. In our adaptive variant, we additionally bound the predicted (pre-)weights to discourage degenerate solutions that effectively ignore some modalities, further promoting multi-way alignment.

Behavior in the pairwise case ($K = 2$) and distinction from cosine ranking. The pairwise case highlights a qualitative difference between SCORE and PMRL. For $K = 2$ with $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{S}^{d-1}$ and $c := \langle \mathbf{z}_1, \mathbf{z}_2 \rangle$, the representation matrix $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2]$ has Gram matrix

$$\mathbf{Z}^\top \mathbf{Z} = \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}, \tag{32}$$

whose eigenvalues are $1 \pm c$. Hence the dominant singular value is

$$\sigma_1(\mathbf{Z}) = \sqrt{1 + |c|}, \tag{33}$$

which depends on c only through its absolute value $|c|$. Consequently, PMRL best-direction objective is *monotone in* $|c|$ and takes the same value at $c = +1$ and $c = -1$ (both anti-aligned and aligned pairs are treated as maximally consistent). In contrast, SCORE with uniform weights gives $\ell_{\mathrm{SCORE}} = \sqrt{(1 + c)/2} = \cos(\theta/2)$ (Appendix B.1), which is *strictly increasing in c over the full range* $c \in [-1, 1]$ and is minimized at $c = -1$. Therefore, SCORE induces exactly the cosine ranking for pairwise retrieval, whereas PMRL spectral objective does not (it is invariant to the sign of c). This full-range consistency with cosine when $K = 2$ distinguishes SCORE from PMRL, and, together with the additive (rather than volumetric) aggregation, from GRAM as well.

B.3 Relation to hub-based pairwise alignment objectives

A common multi-modal training strategy (e.g., ImageBind-style alignment) optimizes multiple pairwise cosine-InfoNCE losses between a chosen hub modality (image in ImageBind; text in LanguageBind) and each other modality, typically aggregating these losses across modality pairs. Such objectives directly involve only hub-other similarities, while non-hub modality pairs are not explicitly optimized.

In contrast, Section 4.2 shows that SCORE induces a quantity that depends on a weighted sum of all pairwise cosine similarities within the tuple, including non-hub pairs. This provides a direct training signal for non-anchor alignment, which is desirable for variable-modality inference where the hub modality may be absent at test time. The same conclusion holds for the normalized length $\tilde{\ell}$ (Eq. 11), since it differs only by a positive scaling factor $1/W$.

C Why Additive Aggregation? Degenerate Optima of Multiplicative Dispersion Logits

This section formalizes a structural property that motivates SCORE's additive design. We contrast the behavior of additive concentration logits with that of multiplicative dispersion logits (e.g., triangle area or parallelotope volume) under the standard A→D negative construction (Eq. 3), focusing on the degenerate regime in which a subset of modalities becomes nearly coincident. We emphasize that the statements below are purely geometric properties of the respective logits. They characterize how each logit responds to degeneracy and thereby clarify the rationale behind SCORE's additive aggregation.

Setup (three modalities). Consider $K = 3$ modalities: anchor $\mathbf{a} \in \mathbb{S}^{d-1}$ and two non-anchor modalities $\mathbf{m}_2, \mathbf{m}_3 \in \mathbb{S}^{d-1}$. In the A→D direction (Eq. 3), negatives keep the anchor fixed and replace both non-anchor modalities with those from another instance j. Thus, for a fixed anchor index i, the A→D denominator contains logits of the form

$$\ell(\mathbf{a}_i, \mathbf{m}_{2j}, \mathbf{m}_{3j}), \qquad j \in \{1, \ldots, B\}, \tag{34}$$

where $(\mathbf{m}_{2j}, \mathbf{m}_{3j})$ is a matched non-anchor pair from the same instance j.

TRIANGLE logit. We consider a TRIANGLE-style multiplicative dispersion logit defined as the negative triangle area:

$$\ell_{\triangle}(\mathbf{a}, \mathbf{m}_2, \mathbf{m}_3) := -A_{\triangle}(\mathbf{a}, \mathbf{m}_2, \mathbf{m}_3), \qquad A_{\triangle}(\mathbf{a}, \mathbf{m}_2, \mathbf{m}_3) := \frac{1}{2}\sqrt{\|\mathbf{u}\|_2^2 \, \|\mathbf{v}\|_2^2 - \langle \mathbf{u}, \mathbf{v} \rangle^2}, \quad (35)$$

where $\mathbf{u} := \mathbf{m}_2 - \mathbf{a}$ and $\mathbf{v} := \mathbf{m}_3 - \mathbf{a}$.

The following lemma applies to any non-negative dispersion measure $\mathrm{Disp}(\cdot)$ (e.g., triangle area or parallelotope volume). We instantiate it for TRIANGLE here and revisit GRAM in the remark below.

Lemma 1 (Upper bound of negated dispersion logits)**.** *For any non-negative dispersion measure* $\mathrm{Disp}(\cdot) \geq 0$*, the negated logit* $\ell(\cdot) = -\mathrm{Disp}(\cdot)$ *satisfies* $\ell(\cdot) \leq 0$*. In particular, for TRIANGLE,* $\ell_{\triangle}(\mathbf{a}, \mathbf{m}_2, \mathbf{m}_3) \leq 0$ *with equality iff* $A_{\triangle} = 0$*.*

Proof. Immediate from $\mathrm{Disp}(\cdot) \geq 0$. □

Both propositions below follow from the same geometric fact: $A_{\triangle}(\mathbf{a}, \mathbf{m}_2, \mathbf{m}_3)$ becomes small when any two of the three embeddings $\mathbf{a}, \mathbf{m}_2, \mathbf{m}_3$ become close.

Proposition 2 (Anchor–non-anchor near-coincidence saturates the logit)**.** *For any* $\mathbf{a}, \mathbf{m}_2, \mathbf{m}_3 \in \mathbb{S}^{d-1}$*,*

$$A_{\triangle}(\mathbf{a}, \mathbf{m}_2, \mathbf{m}_3) \leq \frac{1}{2} \, \|\mathbf{m}_2 - \mathbf{a}\|_2 \, \|\mathbf{m}_3 - \mathbf{a}\|_2. \quad (36)$$

In particular, if $\|\mathbf{m}_2 - \mathbf{a}\|_2 \to 0$*, then* $A_{\triangle} \to 0$ *and hence* $\ell_{\triangle} \to 0$*, i.e., the logit approaches its maximum (Lemma 1) even if* $\mathbf{m}_3$ *remains misaligned.*

Proof. By definition,

$$A_{\triangle}(\mathbf{a}, \mathbf{m}_2, \mathbf{m}_3) = \frac{1}{2}\sqrt{\|\mathbf{u}\|_2^2 \, \|\mathbf{v}\|_2^2 - \langle \mathbf{u}, \mathbf{v} \rangle^2} \leq \frac{1}{2} \, \|\mathbf{u}\|_2 \, \|\mathbf{v}\|_2 = \frac{1}{2} \, \|\mathbf{m}_2 - \mathbf{a}\|_2 \, \|\mathbf{m}_3 - \mathbf{a}\|_2, \quad (37)$$

where the inequality follows from $\langle \mathbf{u}, \mathbf{v} \rangle^2 \geq 0$ (equivalently, Cauchy–Schwarz). □

Proposition 3 (Non-anchor-side degeneracy yields anchor-insensitive logits)**.** *Let* $\mathbf{a}, \mathbf{m}_2, \mathbf{m}_3 \in \mathbb{S}^{d-1}$ *and define* $A_{\triangle}$ *and* $\ell_{\triangle}$ *as in Eq. 35. If* $\mathbf{m}_2 = \mathbf{m}_3$*, then* $A_{\triangle}(\mathbf{a}, \mathbf{m}_2, \mathbf{m}_3) = 0$ *and hence* $\ell_{\triangle}(\mathbf{a}, \mathbf{m}_2, \mathbf{m}_3) = 0$*, independently of* $\mathbf{a}$*. More generally, for any* $\mathbf{a}$*,*

$$A_{\triangle}(\mathbf{a}, \mathbf{m}_2, \mathbf{m}_3) \leq \|\mathbf{m}_2 - \mathbf{m}_3\|_2, \quad (38)$$

and therefore if $\|\mathbf{m}_2 - \mathbf{m}_3\|_2 \leq \varepsilon$ *then* $\ell_{\triangle}(\mathbf{a}, \mathbf{m}_2, \mathbf{m}_3) \geq -\varepsilon$*.*

Proof. The triangle area does not depend on which vertex is taken as the origin. Taking $\mathbf{a}$ as the origin, the two edge vectors are $\mathbf{u} := \mathbf{m}_2 - \mathbf{a}$ and $\mathbf{v} := \mathbf{m}_3 - \mathbf{a}$; taking $\mathbf{m}_2$ as the origin, they are $\mathbf{a} - \mathbf{m}_2 = -\mathbf{u}$ and $\mathbf{w} := \mathbf{m}_3 - \mathbf{m}_2$. Both describe the same triangle, so

$$A_{\triangle} = \frac{1}{2}\sqrt{\|\mathbf{u}\|_2^2 \, \|\mathbf{v}\|_2^2 - \langle \mathbf{u}, \mathbf{v} \rangle^2} = \frac{1}{2}\sqrt{\|\mathbf{u}\|_2^2 \, \|\mathbf{w}\|_2^2 - \langle \mathbf{u}, \mathbf{w} \rangle^2}. \quad (39)$$

Therefore, by Cauchy–Schwarz and $\|\mathbf{u}\|_2 = \|\mathbf{m}_2 - \mathbf{a}\|_2 \leq 2$,

$$A_{\triangle} \leq \frac{1}{2} \, \|\mathbf{u}\|_2 \, \|\mathbf{w}\|_2 \leq \|\mathbf{m}_3 - \mathbf{m}_2\|_2. \quad (40)$$

In particular, if $\mathbf{m}_2 = \mathbf{m}_3$ then $\mathbf{w} = \mathbf{0}$ and $A_{\triangle} = 0$. □

Consequence under A→D InfoNCE. Fix an anchor index i and consider the A→D per-sample loss

$$\mathcal{L}_{\mathrm{A2D},i} = -\frac{\ell_{ii}}{\tau} + \log \sum_{j=1}^{B} \exp(\ell_{ij}/\tau), \qquad \ell_{ij} := \ell(\mathbf{a}_i, \mathbf{m}_{2j}, \mathbf{m}_{3j}). \quad (41)$$

By Lemma 1, all logits satisfy $\ell_{ij} \leq 0$. If a negative candidate $j \neq i$ has (nearly) degenerate candidate-side non-anchor modalities, then Proposition 3 gives $\ell_{ij} \approx 0$, i.e., the negative logit is

(nearly) maximal. Since the positive logit ℓ_{ii} is also upper-bounded by 0, it cannot exceed this ceiling. Hence, any reduction of $\mathcal{L}_{\text{A2D},i}$ must come from decreasing such dominating negatives ℓ_{ij}. For multiplicative dispersion logits, decreasing ℓ_{ij} amounts to increasing the candidate-side area/volume, i.e., moving the non-anchor modalities of candidate tuples apart. This is a purely geometric consequence of the logit form.

Contrast: SCORE remains anchor-sensitive under candidate-side degeneracy. Consider SCORE with a fixed anchor weight $w_1 \in (0, 1)$. If the candidate-side non-anchor modalities collapse to a single direction $\mathbf{m} \in \mathbb{S}^{d-1}$ (e.g., $\mathbf{m}_2 = \mathbf{m}_3 = \mathbf{m}$), then the SCORE logit reduces to

$$\ell_{\text{SCORE}}(\mathbf{a}, \mathbf{m}, \mathbf{m}) = \|w_1\mathbf{a} + (1 - w_1)\mathbf{m}\|_2 = \sqrt{w_1^2 + (1 - w_1)^2 + 2w_1(1 - w_1)\langle \mathbf{a}, \mathbf{m}\rangle}, \quad (42)$$

which depends on the anchor–candidate angle through $\langle \mathbf{a}, \mathbf{m}\rangle$. In particular, $\ell_{\text{SCORE}}(\mathbf{a}, \mathbf{m}, \mathbf{m})$ is maximized only when $\mathbf{m}$ aligns with $\mathbf{a}$, and is minimized when $\mathbf{m} = -\mathbf{a}$, yielding

$$|2w_1 - 1| \le \ell_{\text{SCORE}}(\mathbf{a}, \mathbf{m}, \mathbf{m}) \le 1. \quad (43)$$

Thus, candidate-side non-anchor degeneracy alone cannot make the SCORE logit maximal unless the anchor also aligns.

Remark on GRAM (general K). GRAM-style logits are based on the (negative) volume of the K-parallelotope spanned by modality embeddings, expressible via a Gram determinant. The same geometric mechanism applies: if the candidate-side non-anchor embeddings become (nearly) linearly dependent (e.g., $\mathbf{m}_2 = \cdots = \mathbf{m}_K$), the Gram determinant collapses to (near) zero, yielding a (near) maximal logit independently of the anchor. SCORE avoids this anchor-independent collapse because it aggregates all modalities additively through the resultant vector.

Summary. The statements above characterize a purely geometric property: multiplicative dispersion logits attain their maximal value under candidate-side degeneracy independently of the anchor, whereas the additive SCORE logit remains a function of the anchor-candidate agreement through the resultant vector. This anchor-sensitivity under degeneracy is the design rationale behind SCORE additive aggregation.

D Experimental Details

D.1 Vanilla Experiment

All modalities are mapped to a shared latent space of dimension 6, which was also used as the output dimension of the modality-specific encoders. The text branch is implemented with a pretrained GoogleNews word2vec model. Each caption is tokenized, lowercased, stripped of English stopwords, and represented by the average of the corresponding word vectors. The resulting sentence embedding was then passed through a learned linear projection into the common latent space. The audio branch is modeled with a lightweight convolutional encoder composed of three 3×3 convolutional layers with 32, 64, and 128 channels, respectively. Each convolution is followed by a ReLU and a 2×2 max-pooling operation. The final feature map is flattened and processed by a two-layer multilayer perceptron with an intermediate hidden size of 512 before projection to the shared embedding space. The grayscale visual branch uses a simpler convolutional encoder with two 3×3 convolutional layers with 32 and 64 channels, each followed by ReLU and 2×2 max-pooling, and a final linear layer mapping the flattened representation to the common embedding dimension. In the colorized-MNIST setting, the same visual architecture is adopted, with the only modification being the use of three input channels instead of one. For the stroke modality, we use a recurrent encoder. The input is first projected through a linear layer, followed by layer normalization and GELU activation. The projected sequence is then processed by a two-layer bidirectional GRU with hidden size 128 and dropout 0.1. The sequence representation is obtained by combining an attention-weighted pooling of the recurrent states with the final hidden state; their concatenation is finally passed through a projection head consisting of layer normalization, a hidden linear layer with GELU and dropout, and a final linear mapping to the shared latent space.

We keep all the hyperparameters fixed across the EasyCL experiments unless otherwise stated. We use a batch size of 10, trained for 10,000 iterations, and set the learning rate to 10^{-4}. Embeddings are normalized to unit norm, and the contrastive temperature is initialized to 0.07 and treated as

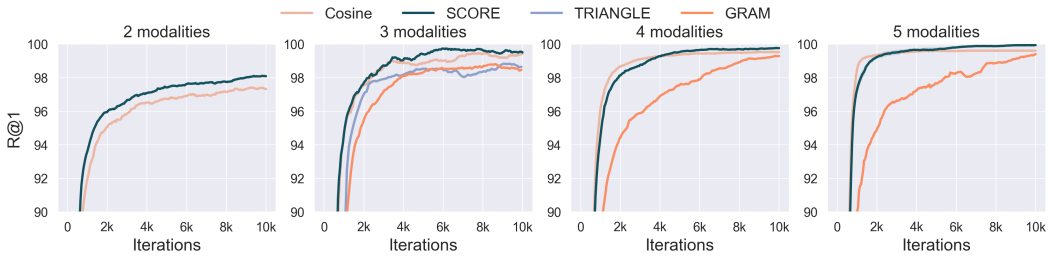


Figure 4: Results on the 5-modal MNIST for the cosine-based method, multi-vector similarities, and the proposed SCORE. Involving more modalities improves R@1 performance.

Table 5: Dataset statistics and hyperparameters. Modalities stand for T: text, V: video, A: audio, S: subtitles. # Frames refers to both training and inference.

Modality	Benchmark	#Video / #Audio		# Frames
		Train	Test	
Threemodal (T-V-A)	DiDeMo	8394	1003	8
	ActivityNet	10009	4917	8
	VGG-sound+	187520	1000	8
Fourmodal (T-V-A-S)	MSR-VTT	9000	1000	8
	VATEX	14060	431	8

learnable, following Radford et al. (2021). For the audio preprocessing, we employ a sampling rate of 48 kHz, an FFT size of 512, 64 Mel bins, and an audio sequence length of 48,000 samples. For the stroke branch, the input dimensionality is set to 3 and the trajectory length to 64. All the experiments are conducted on a single NVIDIA RTX 4080 with 16GB of VRAM, and runs are pretty fast, taking a few minutes to complete.

D.2 Vanilla Experiment

We first study SCORE in a controlled setting where we scale the number of modalities up to five, isolating the effect of modality count under a common encoder and training recipe. Starting from MNIST (LeCun et al., 1998) (text as the digit word, image as the second modality), we successively add AudioMNIST (Becker et al., 2024) (audio), Stroke-MNIST (Singleton, 2016) (stroke trajectories), and Colored-MNIST (Lee et al., 2025) (color) as third to fifth modalities. Across loss configurations, from cosine-based to multi-vector (GRAM, TRIANGLE) and SCORE, adding modalities improves R@1, and SCORE benefits most from richer multimodal evidence.

Figure 4 reports R@1 for the controlled MNIST-based study as we scale the number of modalities from two to five. Across all settings, involving more modalities improves R@1, and SCORE benefits most from richer multimodal evidence compared to cosine-based and multi-vector baselines (GRAM, TRIANGLE).

D.3 Large-scale Experiments

We employ several benchmark datasets for the downstream tasks:

MSR-VTT (Xu et al., 2016) contains 10000 video clips paired with 200000 captions, covering a wide variety of topics such as human activities, sports, and natural scenes.

VATEX (Wang et al., 2019) includes 41250 video clips sourced from the Kinetics-600 dataset (Kay et al., 2017), along with 825000 sentence-level annotations. As a significant portion of the dataset is no longer accessible online due to removed or private videos, we utilize a reduced subset consisting of 14491 samples.

Table 6: Ablation on weight bounds on MSR-VTT (from scratch). The anchor (text) pre-weight is fixed to 1. AVG is the mean of T2VA, T2V, and T2A R@1. $[w_{\min}, w_{\max}] = [1, 1]$ corresponds to SCORE-Uniform. The $[0.8, 1.2]$ interval gives the best balanced performance across pairwise and joint retrieval.

Bounds $[w_{\min}, w_{\max}]$	T2VA	T2V	T2A	AVG
Unbounded	38.5	32.9	0.2	23.87
$[0.0, 2.0]$	38.5	33.6	0.2	24.10
$[0.4, 1.6]$	37.9	**33.7**	3.7	25.10
$[0.8, 1.2]$	**38.9**	32.1	5.1	**25.36**
$[0.9, 1.1]$	38.3	31.7	5.2	25.06
$[1.0, 1.0]$ (SCORE-Uniform)	38.0	30.8	**6.2**	25.00

DiDeMo (Hendricks et al., 2017) is composed of 10000 long videos collected from Flickr, where each video is annotated with four short sentences arranged in temporal order. These descriptions correspond to distinct moments within the video, capturing specific events or scenes. Each video therefore includes four temporally ordered captions describing its progression.

ActivityNet (Caba Heilbron et al., 2015) consists of 20000 long videos (with an average duration of 180 seconds) collected from YouTube, paired with 100000 captions. The dataset spans 200 diverse human activity classes, including daily actions, sports, and complex behaviors. It is annotated with both activity labels and temporal boundaries, enabling precise localization of activities within each video.

VGG-sound+ (Cho, 2024) is an extension of the VGGSound dataset that introduces a triple-modality setting, combining image, text, and audio data. It contains approximately 200000 video samples, where each instance is represented as a triplet (I_i, T_i, A_i) consisting of an image frame, a textual description, and an audio clip. Since the original VGGSound dataset provides only category-level metadata without detailed captions, textual descriptions are generated using two strategies: a semi-handcrafted method based on prompt templates, and an automatic captioning approach using the BLIP-2 model.

For all datasets, we adopt the official splits for retrieval tasks. Details regarding dataset splits and the number of frames used for fine-tuning and evaluation are reported in Tab. 5.

We pretrain the SCORE and SCORE-Uniform models on a subset of the VAST27M (Chen et al., 2023b) dataset, using 150k randomly sampled instances. Training is performed with a learning rate of $1e-4$ using the AdamW optimizer with weight decay and a batch size of 256. We use $w_{\min} = 0.8$ and $w_{\max} = 1.2$ as bounds for the adaptive weighting mechanism, empirically tested and results shown in Tab. 6. During fine-tuning, the batch size is reduced to 64, and the number of epochs is adjusted depending on the dataset. Complete training details are provided in Tab. 5. The experiments are conducted on a cluster with 4×L40 NVIDIA GPUs with 48GB of VRAM.

E Ablation on Weight Bounds

We ablate the clipping bounds $[w_{\min}, w_{\max}]$ applied to the non-anchor pre-weights in Eq. 10. Recall that the anchor (text) pre-weight is fixed to $\hat{w}_{1,ij} = 1$ (Sec. 4.3), and that $[w_{\min}, w_{\max}] = [1, 1]$ recovers SCORE-Uniform (all modalities weighted equally). Widening the bounds increases the expressivity of the adaptive weighting, but also allows the module to assign extreme relative weights that effectively ignore some modalities, which can harm multi-way alignment and thus variable-modality inference.

Setup. We train from scratch on MSR-VTT and vary the bounds, keeping all other hyperparameters fixed. We report R@1 for joint retrieval (T2VA) and pairwise retrieval (T2V, T2A), and summarize the trade-off with AVG, the mean of the three R@1 values.

Results. Table 6 shows a clear trade-off governed by the bounds. Without bounding (Unbounded) or with loose bounds ($[0, 2]$), the adaptive module drives T2A R@1 to a near-degenerate 0.2. This confirms the motivation for bounding the pre-weights (Section 4.3): unconstrained weights admit

degenerate solutions that collapse the multi-modal objective onto a subset of modalities, which is harmful under variable-modality inference. As the bounds tighten, T2A recovers, and the joint T2VA score improves. The interval $[0.8, 1.2]$ attains the best joint retrieval (T2VA 38.9) and the best overall balance (AVG 25.70), while retaining competitive pairwise performance. We therefore adopt $[0.8, 1.2]$ throughout. We note that these bounds are chosen for balanced performance across pairwise and joint retrieval rather than to maximize any single metric: for example, SCORE-Uniform ($[1, 1]$) attains the highest T2A, and looser bounds attain higher T2V, but neither balances the two as well as $[0.8, 1.2]$.

F DTM Formulation

DTM complements the contrastive objective by providing an additional cross-modal matching signal via cross-attention. It is a binary classification objective that predicts whether the text and the remaining modalities come from the same instance. We encode text with cross-attention enabled, conditioning on the concatenated (unpooled) features from the other modalities, and apply an MLP head to obtain a match probability p_i, following Li et al. (2021). We construct negatives by pairing the text of instance i with the non-text modalities of another instance $j \neq i$ within the batch (with hard negative mining following Li et al. (2021)) and optimize the binary cross-entropy loss:

$$\mathcal{L}_{\text{DTM}} = -\frac{1}{B} \sum_{i=1}^{B} \left[y_i \log p_i + (1 - y_i) \log(1 - p_i) \right], \tag{44}$$

where $y_i = 1$ for matched tuples and $y_i = 0$ otherwise.

G Analysis of the Adaptive Weighting Mechanism

We analyze the learned adaptive weights (Section 4.3) to understand how the module modulates modality contributions and when it helps.

Learned weights on MSR-VTT. Training from scratch on MSR-VTT, we record the pre-weights assigned to each non-anchor modality (the anchor/text pre-weight is fixed to 1; non-anchor pre-weights are clipped to $[0.8, 1.2]$). On average, the module assigns a higher weight to video (≈ 1.20) than to audio (≈ 0.84). Inspecting retrieval examples in which audio receives a low weight, we find that many exhibit sparse or noisy audio, or audio only weakly aligned with the caption. This is consistent with MSR-VTT being originally designed for video-text retrieval, where the audio stream is not guaranteed to be informative.

Interpretation and trade-off. On these datasets, audio is a less reliable cue for joint alignment, yet it remains informative for audio-only retrieval. The adaptive module correctly identifies audio as less informative for the joint score and reduces its contribution accordingly, as reflected in the learned weights above. Because SCORE optimizes a single multimodal objective (jointly over text, video, and audio) rather than optimizing each retrieval direction independently, downweighting a noisy or weakly aligned audio stream improves joint retrieval (e.g., T2AV) at a small cost to the audio-only direction (T2A), where audio is the only non-text evidence. The bounded weights ensure that audio is never entirely ignored, preserving multi-way alignment. The largest benefit over SCORE-Uniform is thus observed in the joint setting, whereas at $K = 2$ the module learns near-uniform weights and SCORE behaves like its uniform counterpart (Appendix H, Clotho).

H Additional Evaluation: Cross-dataset Transfer

The clustering and classification results on MSR-VTT are reported in Section 5.1 (Table 1). Here we additionally evaluate SCORE in a standalone audio-text setting to test transfer beyond video-text data.

Cross-dataset evaluation: standalone audio retrieval (Clotho). To test SCORE in a standalone audio-text setting ($K = 2$), we train on Clotho (Drossos et al., 2020) with the same vanilla encoders and hyperparameters for all methods. Table 7 shows that SCORE slightly outperforms the cosine baseline on this audio-centric benchmark. At $K = 2$, the adaptive weighting module learns near-uniform weights, so SCORE behaves similarly to SCORE-Uniform (consistent with Appendix G).

Table 7: Standalone audio-text retrieval on Clotho ($K = 2$). R@1.

Method	T2A R@1	A2T R@1
Cosine	17.42	22.20
SCORE-Uniform	18.03	22.78
SCORE (ours)	**18.05**	**22.78**

I LIMITATIONS AND DISCUSSION

Overhead of instance-pair-adaptive weighting. While the SCORE logit itself is lightweight (a weighted sum followed by an ℓ_2 norm), our adaptive variant predicts weights for each compared pair (i, j). This introduces additional computation on top of the standard $O(B^2)$ within-batch logit evaluation used by InfoNCE-style training. Although our experiments show favorable wall-clock time and memory in the evaluated settings, scaling to substantially larger batch sizes or more modalities may require further engineering, e.g., computing weights only for a subset of negatives (such as hard/top-k negatives) or amortizing weights across candidates.

Scaling to many modalities. SCORE aggregates modality embeddings through a (normalized) weighted resultant. As the number of modalities K increases, simple aggregation can become less effective: the joint score may be dominated by a subset of modalities, or conversely become overly "averaged" and less discriminative. Our adaptive weighting increases the expressivity of the joint score by modulating modality contributions on a per-pair basis, but it may not fully address all challenges at larger K, such as heterogeneous modality availability patterns and the combinatorial diversity of modality subsets at inference time. Designing more structured weighting mechanisms (e.g., hierarchical or group-wise weighting, or weights conditioned on modality availability patterns) is a promising direction for improving robustness at larger K.