STR: AUDITING SHARED-DIRECTION GEOMETRY IN MULTIMODAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal contrastive learning involves inter-modal semantic agreement, inter-modal separation, and intra-modal diversity, yet most alignment objectives fold these concerns into a single aggregate similarity score. This paper decomposes them into separately weighted geometric terms, collectively Surface-Tension Regularization (STR), an auxiliary regularizer with no additional learnable parameters, and uses the decomposition as a measurement instrument. Across four VAST benchmarks and three seeds per benchmark, canonical STR increases common-mode energy and differs from the baseline in its response to shared-direction removal. Separating deletion from residual-norm rescaling in these two models shows that deletion drives the baseline loss, whereas under STR rescaling alone is disruptive and the combined operation approximately preserves aggregate top-1 recall. Replacing Gram-volume attraction with pairwise cosine also yields energy inflation and a similar combined-removal response on MSR-VTT and DiDeMo across three seeds per dataset. Seed-0 factorial and checkpoint-conditional gradient measurements provide complementary evidence about coupling and mean-margin changes. A single fixed STR configuration without per-dataset tuning produces small changes in final top-1 recall on the four benchmarks. The same metric remains similar across configurations with substantially different embedding-space behavior, motivating separate evaluation of the embeddings deployed directly.

1 INTRODUCTION

Multimodal representation learning builds shared embedding spaces across modalities such as video, text, and audio for cross-modal retrieval and multimodal understanding. Strong performance requires both inter-modal alignment (bringing semantically corresponding representations closer) and intra-modal structure preservation (maintaining within-modality expressiveness and diversity), yet the two can conflict and must be balanced. Most prior works address this with contrastive learning (Oord et al., 2018; Radford et al., 2021), and recent methods replace the cosine similarity inside the contrastive objective with a single aggregate geometric or information-theoretic measure (Cicchetti et al., 2025b;a; Saporta et al., 2024). These aggregate objectives fold three different concerns into one similarity score: inter-modal semantic agreement (Wang & Isola, 2020), inter-modal separation (Liang et al., 2022), and intra-modal diversity (Chen et al., 2024; Zhang et al., 2026). When such an objective fails to improve performance, which geometric pressure was misdirected is difficult to determine from the loss value alone.

This paper decomposes the three concerns into separately weighted geometric terms, collectively **Surface-Tension Regularization (STR)**: an auxiliary regularizer with no additional learnable parameters that operates alongside standard contrastive training (here the Image-Text Contrastive (ITC) and Image-Text Matching (ITM) losses) without modifying positives, negatives, or the similarity function. The three terms are an inter-modal agreement term L_{vol} (attraction), an inter-modal separation term L_{sep} (repulsion), and an intra-modal diversity term L_{div} (dispersion); each is constructed so that its embedding-level gradient acts on a structurally different part of the embedding space (App. K.2). The terms are separately weighted but coupled: they share encoders and embeddings, so separate coefficients provide intervention access, not outcome-level independent control.

The decomposition serves here to audit how shared-direction deletion and residual rescaling affect retrieval, rather than to establish a uniformly superior regularizer. Applying it as a measurement

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.

instrument on four video-text retrieval benchmarks (VATEX (Wang et al., 2019), MSR-VTT (Xu et al., 2016), DiDeMo (Hendricks et al., 2017), ActivityNet (Heilbron et al., 2015)) exposes an asymmetry that a single aggregate score does not: alignment pressure concentrates embedding energy in a single shared direction, and the same backbone finetuned without and with STR (the unregularized baseline and the regularized model) responds very differently to removing it. Deleting the direction produces a large drop in the baseline's embedding-space retrieval, while under STR residual rescaling alone is disruptive and deletion plus rescaling approximately preserves the aggregate score. Energy inflation and the combined-removal response also recur with pairwise cosine attraction on MSR-VTT and DiDeMo, with three seeds per dataset (Sec. 4.4). The full deletion/rescaling decomposition is evaluated for the baseline and canonical STR; factorial and checkpoint-conditional gradient measurements provide complementary evidence about the coupled terms. The contributions are threefold:

(i) A shared-direction audit that separates deletion from residual-norm rescaling and connects embedding geometry to retrieval through controlled interventions, factorial responses, and gradient measurements.

(ii) Distinct baseline–canonical STR removal responses across four VAST benchmarks and three seeds, with energy inflation and combined-removal behavior also repeated under pairwise attraction on two datasets with three seeds each. Similar final pipeline scores coexist with substantially different direct-embedding retrieval.

(iii) STR as a measurement instrument: separately weighted but coupled agreement, separation, and diversity terms, each motivated by a stated failure mode and invariance requirement, with gradient-matched counterparts to test dependence on functional form.

2 Related Work

Aggregate alignment objectives. Dual-encoder contrastive pretraining (Radford et al., 2021; Jia et al., 2021) has been extended to video-text retrieval (Miech et al., 2020; Bain et al., 2021; Luo et al., 2022; Lin et al., 2022; Ibrahimi et al., 2023) and unified multi-modality architectures (Wang et al., 2024; Yan et al., 2022; Chen et al., 2023b; Ye et al., 2026), where a modality gap (Liang et al., 2022; Yaras et al., 2024) makes alignment trade off against collapse and separability (Yang et al., 2022; Chen et al., 2024; Zhang et al., 2026). GRAM (Cicchetti et al., 2025b), TRIANGLE (Cicchetti et al., 2025a), and Symile (Saporta et al., 2024) swap the measure inside the contrastive loss for higher-order geometric or information-theoretic quantities, but keep agreement, separation, and diversity fused in one score. STR instead keeps them as separately weighted terms over an unchanged contrastive loss and compares against all three under one protocol (Sec. 4.2).

Collapse and anisotropy in embedding spaces. Representation collapse has been addressed through negative sampling (Chen et al., 2020; He et al., 2020), architectural asymmetry (Grill et al., 2020; Chen & He, 2021), and cross-correlation or variance constraints (Zbontar et al., 2021; Bardes et al., 2022); the alignment-uniformity framework (Wang & Isola, 2020) balances alignment against coverage, with later work on dimensional collapse (Jing et al., 2021; Liu et al., 2023; Chen et al., 2024; He et al., 2024) and its multimodal forms (Yang et al., 2022; Lin et al., 2023). The uniformity potential (Wang & Isola, 2020) and the VICReg variance hinge (Bardes et al., 2022) serve as gradient-matched counterparts of L_{div} (Sec. 4.4). All-but-the-Top (ABTT) (Mu et al., 2017) removes the mean and top components from word embeddings as post-processing (Gao et al., 2019); Sec. 4.3 uses a related direction-removal diagnostic in the opposite role, measuring how retrieval responds to removing an inflated shared direction with and without alignment regularization.

Gradient interference and physics-inspired regularization. PCGrad (Yu et al., 2020), GradNorm (Chen et al., 2018), and CAGrad (Liu et al., 2021a) modify gradient conflict between tasks in multi-task learning; here the gradient relations among auxiliary terms inside one objective are read as a diagnostic and never modified. Physics-inspired regularization treats embeddings as interacting particles, from hyperspherical-energy spreading (Liu et al., 2018; Lin et al., 2020; Liu et al., 2021b) to potential-field and Lennard-Jones objectives (Bhatnagar & Ahuja, 2025; Kwan et al., 2024); STR instead applies conditional within-modality repulsion only to near-collapsed pairs (Sec. 3.5).

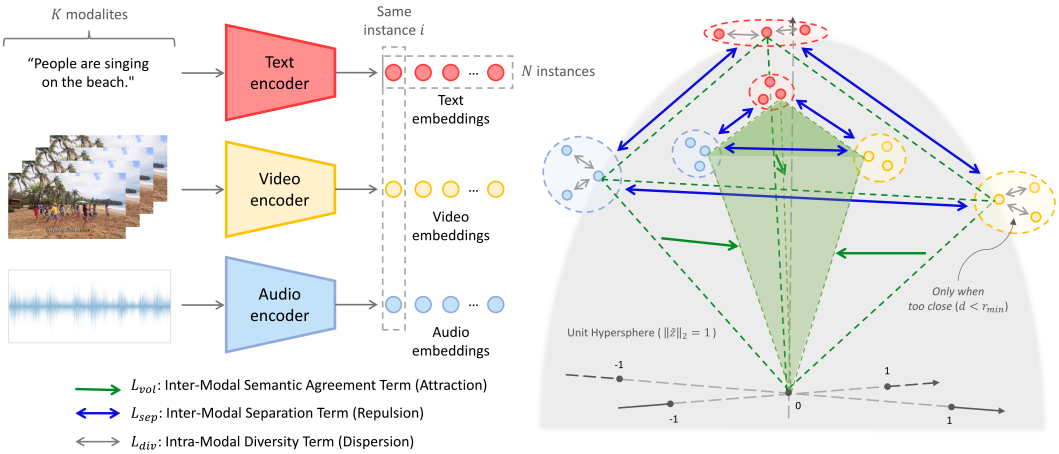

Figure 1: Overview of Surface-Tension Regularization (STR) with K=3 modalities (text, video, audio). Each instance's embeddings span a parallelotope: L_{vol} penalizes its Gramian volume (attraction), L_{sep} resists over-alignment (repulsion), and L_{div} spreads instances within each modality (dispersion). L_{vol} and L_{sep} are drawn at the cluster level but computed per instance.

3 Geometric Surface-Tension Regularization

3.1 Problem Setup and Notation

A dataset contains N instances with observations from K modalities, $\{(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(K)})\}_{i=1}^{N}$. For modality k, an encoder f_k maps $x_i^{(k)}$ to a feature vector, followed by a projection head g_k producing an embedding $z_i^{(k)} \in \mathbb{R}^d$. To prevent trivial minimization of geometric quantities by norm shrinkage, ℓ_2-normalization is applied: $\tilde{z}_i^{(k)} = z_i^{(k)} / \|z_i^{(k)}\|_2$. A mini-batch of size B is denoted by $\{(\tilde{z}_i^{(1)}, \ldots, \tilde{z}_i^{(K)})\}_{i=1}^{B}$. The main VAST experiments use K=3 modalities (text, video, audio).

3.2 Overview

Figure 1 illustrates STR. All terms are computed from ℓ_2-normalized embeddings and introduce no additional learnable parameters. Let L_{task} denote a primary training objective; the overall optimization target is

$$L_{total} = L_{task} + R_{STR}, \qquad R_{STR} = \lambda_{vol} L_{vol} + \lambda_{sep} L_{sep} + \lambda_{div} L_{div}, \tag{1}$$

where in the main experiments $L_{task} = L_{ITC} + L_{ITM}$, the standard ITC and ITM losses (App. L).

Each term below is an operational surrogate, not a unique measure, with residual constants fixed by limited sweeps (App. M). The three terms are not an exhaustive taxonomy: a fourth axis, intra-modal compactness, was evaluated and excluded because its contraction compounds with L_{vol}'s (App. I). Missing modalities are handled by per-sample presence masks (App. L.1).

3.3 Inter-Modal Semantic Agreement Term

Failure mode and requirements. The targeted failure is joint misalignment: the K modality embeddings of an instance failing to occupy a common low-volume configuration. A measure of it should be symmetric across modalities, invariant to global rotations, and vanish exactly when the embeddings are linearly dependent.

For instance i, let $Z_i = [\tilde{z}_i^{(1)}, \ldots, \tilde{z}_i^{(K)}] \in \mathbb{R}^{d \times K}$ denote the stacked embedding matrix and $G_i = Z_i^\top Z_i \in \mathbb{R}^{K \times K}$ the corresponding Gram matrix, with $G_{i,ab} = \tilde{z}_i^{(a)\top} \tilde{z}_i^{(b)}$ under ℓ_2-normalization. For $K \leq d$, the squared K-volume of the parallelotope spanned by $\{\tilde{z}_i^{(k)}\}_{k=1}^{K}$ equals $\det(G_i)$ (Gantmacher, 1959): it satisfies all three requirements and reduces to $\sin^2(\theta_i)$ for K=2, where

minimizing volume favors collinearity ($|\cos\theta_i| \to 1$) and the task loss supplies the positive-alignment direction. In practice, L_{vol} takes the stabilized form

$$L_{vol} = \frac{1}{B}\sum_{i=1}^{B}\sqrt{\det(G_i + \varepsilon I_K)}, \tag{2}$$

where εI_K is a ridge added for the Cholesky-based computation; the numerical behavior of that path is reported in App. L.1. The square root is preferred to the logarithm of the determinant for numerical reasons only: its gradient stays bounded near degeneracy. For a fixed positive ridge both are increasing functions of the per-instance determinant, which does not imply identical optimization dynamics or identical optima after batch aggregation and combination with the task loss.

Why not a pairwise sum. A pairwise-cosine sum $\sum_{a<b}\cos(\tilde{z}_i^{(a)}, \tilde{z}_i^{(b)})$ depends on the off-diagonals of G_i only through their sum and is therefore blind to their joint structure; App. J makes this precise (its derivative couples each pair to the others, a coupling that first appears for $K \geq 3$), and Sec. 4.4 evaluates a gradient-matched pairwise-cosine counterpart directly. For $K>2$, low volume is weaker than enforcing a common direction; the positive-alignment direction is supplied jointly by the task loss. Gramian volume also appears in GRAM (Cicchetti et al., 2025b), as a cross-sample similarity inside a contrastive objective; L_{vol} applies it within each instance as an auxiliary regularizer, without redefining the contrastive similarity.

3.4 Inter-Modal Separation Term

Failure mode and requirements. The targeted failure is over-alignment of the most aligned pairs. The counter-pressure must not oppose the alignment that ITC/ITM and L_{vol} already enforce on typical matched pairs, and its penalty should be invariant to a uniform shift of the similarity distribution, which drifts during finetuning.

Let $P = \{(a,b) \mid 1 \leq a < b \leq K\}$ be the modality pairs and $s_i^{(a,b)} = \cos(\tilde{z}_i^{(a)}, \tilde{z}_i^{(b)})$ the within-instance pairwise cosine. A plain cosine penalty pulls every pair toward -1, opposing matched alignment, so the penalty must be one-sided; and a fixed activation threshold would change its active fraction as the similarity distribution drifts. A batch-quantile threshold meets both requirements: $\tau_q = \mathrm{Quantile}_q\big(\{s_i^{(a,b)}\}_{i\in[B],\,(a,b)\in P}\big)$, $q \in (0,1)$, penalizes the top $(1-q)$ fraction of pairs and, moving with any uniform shift, keeps the penalty and its active set shift-invariant. The inter-modal separation objective is

$$L_{sep} = \frac{1}{|P|}\sum_{(a,b)\in P}\frac{1}{B}\sum_{i=1}^{B}\Big[\max\big(0,\ s_i^{(a,b)} - \tau_q\big)\Big]^2, \tag{3}$$

with τ_q pooled over all batch pairs and instances and recomputed each step. L_{sep} is a guard: it penalizes the top $(1-q)$ fraction by construction, but at the reference configuration the excess above τ_q is small, so its effective pressure is low (Sec. 4.6).

3.5 Intra-Modal Diversity Term

Failure mode and requirements. The targeted failure is intra-modal collapse: embeddings from different instances concentrating at extremely small pairwise distances. The counter-pressure should act on near-collapsed pairs while leaving longer-range semantic structure untouched, which motivates short-range repulsion over global dispersion; global uniformity potentials apply pressure to all pairs regardless of distance, and this design difference is evaluated directly against the uniformity and variance counterparts in Sec. 4.4.

Inspired by the repulsive regime of the Lennard-Jones potential, soft repulsion is applied when distances fall below a threshold r_{min}; the smooth barrier in Eq. 4 keeps gradients bounded, unlike a full Lennard-Jones interaction (App. L.3).

For modality k, define batch pairwise distances $d_{ij}^{(k)} = \|\tilde{z}_i^{(k)} - \tilde{z}_j^{(k)}\|_2$ for $1 \leq i,j \leq B$, and the pairwise penalty

$$\ell(d_{ij}^{(k)}) = \mathbf{1}\Big[\,i \neq j,\ d_{ij}^{(k)} < r_{min}\Big]\left(1 - \exp\big(-(d_{ij}^{(k)} - r_{min})^2/(2\sigma^2)\big)\right). \tag{4}$$

Table 1: Text-to-video retrieval (ITM) of a pretrained VAST finetuned with different alignment objectives under an identical protocol (t-va, held-out selection, one fixed configuration per method). Entries are R1 and R_{avg} as mean ±SD over three seeds $(0, 1, 2)$. Full R1/R5/R10 in Table 8. STR is the fixed three-term configuration L_{vol}+L_{sep}+L_{div}. Bold marks the best mean per dataset and metric; ties are all marked.

	DiDeMo		**MSR-VTT**		**VATEX**		**ActivityNet**	
Method	R1	R_{avg}	R1	R_{avg}	R1	R_{avg}	R1	R_{avg}
Baseline	66.2 ±0.92	80.8 ±0.65	60.7 ±0.35	77.1 ±0.00	**69.1** ±0.25	**85.8** ±0.06	66.7 ±0.80	83.8 ±0.46
GRAM	65.9 ±1.13	80.8 ±0.36	59.8 ±0.78	76.9 ±0.23	**69.1** ±0.15	**85.8** ±0.06	**67.3** ±0.67	**84.0** ±0.58
Symile	64.1 ±1.31	79.1 ±0.55	57.6 ±0.85	75.6 ±0.21	66.6 ±0.15	84.0 ±0.00	63.5 ±0.75	81.5 ±0.26
TRIANGLE	**66.5** ±0.85	**80.9** ±0.35	60.0 ±0.29	77.3 ±0.26	**69.1** ±0.20	**85.8** ±0.10	66.6 ±0.53	83.8 ±0.15
STR	66.1 ±0.23	**80.9** ±0.15	**61.0** ±1.06	**77.4** ±0.35	**69.1** ±0.17	**85.8** ±0.12	67.1 ±0.55	**84.0** ±0.23

The intra-modal diversity objective is

$$L_{div} = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \ell(d_{ij}^{(k)}) \right), \tag{5}$$

where $\mathbf{1}[\cdot]$ is the indicator function and $\sigma > 0$ controls the smoothness of repulsion around r_{min}; the penalty is bounded with finite gradients, and the cost is $O(KB^2)$ per step (App. L.1).

4 Empirical Evaluation and Analysis

4.1 Setup

A pretrained VAST (Chen et al., 2023b) model is independently finetuned on each dataset, with text, audio, and video encoders instantiated as BERT-B (Devlin et al., 2019), BEATs (Chen et al., 2022), and EVA-CLIP ViT-G (Sun et al., 2023), respectively. STR hyperparameters are $(\lambda_{vol}, \lambda_{sep}, \lambda_{div}) = (0.1, 0.1, 0.9)$, $\varepsilon = 10^{-8}$ in L_{vol}, $q = 0.8$ in L_{sep} with τ_q capped at 1.0, and $(r_{min}, \sigma) = (1.0, 0.5)$ in L_{div}. This single configuration is applied to all datasets and backbones without per-dataset tuning. Baseline denotes the same finetuning with the task loss $L_{ITC} + L_{ITM}$ alone. Finetuning uses a global batch size of 32 (16 per GPU on two GPUs), learning rate 2×10^{-5}, and gradient checkpointing; remaining hyperparameters follow the original VAST configuration.

Datasets, modality setting, and evaluation. Experiments use four video-text retrieval benchmarks, VATEX (Wang et al., 2019), MSR-VTT (Xu et al., 2016), DiDeMo (Hendricks et al., 2017), and ActivityNet (Heilbron et al., 2015), in the three-modality setting t-va (text, video, audio; K=3, no subtitles; App. A). Every configuration in Table 1 is trained with three seeds, and tables report mean and standard deviation (SD) over seeds unless stated otherwise. Checkpoints are selected on a held-out 15% split of the training data (video-level, seed-specific), except VATEX, which uses its native validation split; the test set is used once, for final evaluation. Metrics are text-to-video Recall at rank k (R1/R5/R10) and its mean (R_{avg}) under the ITM score (cross-encoder reranking of fused-similarity candidates), the standard VAST evaluation mode (video-to-text in App. H).

4.2 Comparison with Alignment Objectives

Table 1 provides final-pipeline context for the geometric audit. Relative to the baseline, STR changes mean ITM R1 by $-0.1/+0.3/0.0/+0.4$ percentage points on DiDeMo/MSR-VTT/VATEX/ActivityNet. These are descriptive differences, not evidence of uniform superiority or statistical equivalence. Symile is lower on all four datasets; the other compared objectives have closer final scores. Similar final scores need not imply similar embeddings: Table 22 reports differing seed-0 ITC retrieval beneath the reported VATEX ITM tie, motivating the separate embedding audit of Secs. 4.3 and 4.4.

Table 2: Common-mode measurement and removal intervention. Left: video-modality common-mode energy $\|\mu_v\|^2$ (0 = zero mean vector, 1 = all unit embeddings collapsed to one direction; App. L.2) for term subsets, where Full STR is the fixed three-term configuration of Table 1; text moves the same way with smaller magnitude. Middle: cos(u, SVD1) between the shared centroid direction u and the top uncentered singular direction (SVD1), baseline → STR. Right: change in text-to-video embedding-cosine R1 when u is removed and embeddings renormalized; negative = removal lowers retrieval. Left and Middle are seed 0; Right is mean ±SD over three seeds $(0, 1, 2)$.

	$\|\mu_v\|^2$					cos(u, SVD1)	ΔR1 after removal	
Dataset	Baseline	L_{vol}	L_{div}	$L_{vol}+L_{div}$	Full STR	Baseline → STR	Baseline	STR
DiDeMo	0.189	0.999	0.014	0.475	0.475	0.57 → 1.00	−5.2 ±1.50	+2.8 ±2.80
MSR-VTT	0.124	1.000	0.007	0.455	0.448	0.83 → 1.00	−10.4 ±0.32	+1.0 ±1.50
VATEX	0.230	1.000	0.006	0.379	0.381	0.63 → 1.00	−6.6 ±0.48	−0.1 ±0.64
ActivityNet	0.317	0.999	0.011	0.458	0.420	0.79 → 1.00	−15.5 ±7.22	+0.3 ±0.30

Table 3: Left: removal decomposition on MSR-VTT (embedding-cosine R1, mean ±SD over three seeds $0, 1, 2$) for each model. O: original scores; D: the shared direction's score contribution deleted only; N: residual-norm rescaling only; DN: deletion followed by rescaling. Right: paired removal response $\Gamma = \delta^{\mathrm{STR}} - \delta^{\mathrm{base}}$, $\delta = \mathrm{R1(DN)} - \mathrm{R1(O)}$, per dataset (mean ±SD over three seeds $0, 1, 2$). All four datasets and per-query transitions are in Table 14 (App. D).

Model	O	D	N	DN
Baseline	41.8 ±0.8	30.9 ±0.9	41.8 ±0.6	31.4 ±0.5
STR	40.4 ±0.8	38.7 ±2.8	15.2 ±1.7	41.4 ±1.6

Dataset	Γ	**Dataset**	Γ
DiDeMo	+8.0 ±1.3	MSR-VTT	+11.4 ±1.6
VATEX	+6.5 ±0.5	ActivityNet	+15.8 ±7.2

4.3 The Shared-Direction Failure Mode

This section reports the central diagnostic finding, post hoc on stored checkpoints and ℓ_2-normalized embeddings. Because the intervention edits embeddings directly, retrieval here is text-video embedding-cosine retrieval, not the ITM metric of Table 1 (full tables and controls in App. D).

Alignment pressure concentrates energy in one shared direction. Table 2 (left) reports the common-mode energy $\|\mu_v\|^2$ (μ_v, μ_t: video and text centroids; App. L.2). On its own, L_{vol} collapses the video modality onto essentially one direction on all four datasets and all but eliminates the mean positive-negative margin (*pos-neg margin* in the tables; App. L.2 defines it and the hard-negative margin that decides rank 1); L_{div} acts oppositely, driving $\|\mu_v\|^2$ below the baseline. Combining the two lands in an intermediate range where the mean margin is largest, so the full configuration (all three terms at the fixed weights; Full STR in the tables) does not collapse – the near-total collapse is an L_{vol}-only property. Under the full configuration the baseline's near-orthogonal video and text centroids merge into one direction, also the top singular direction of the uncentered embeddings (Table 2, middle). The energy inflation (1.7 to 3.8× the baseline) reproduces across seeds (Table 13).

Shared-direction removal has different retrieval effects in the baseline and STR. Table 2 (right) removes the shared centroid direction $u \propto \mu_t + \mu_v$ and renormalizes (an ABTT-style removal (Mu et al., 2017)), for both models. This lowers baseline embedding-cosine R1 in 12 of 12 dataset-seed cells while nearly preserving STR's; the same-seed paired removal response Γ (Table 3) is $+6.5$ to $+15.8$ R1 across datasets. Splitting the removal into deletion (D) and residual-norm rescaling (N) separates the score-level effects: deletion drives the baseline loss and rescaling alone is inert there, whereas under STR rescaling alone is disruptive and D followed by N approximately preserves aggregate R1 (definitions in App. D). The combined intervention does not by itself establish that the direction is uninformative, and query-level transitions separate aggregate-R1 preservation from ranking preservation: under DN the baseline loses top-1 on 3–7 times as many queries as it gains, while under STR gains and losses are small and comparable (Table 14). At MSR-VTT seed 0, the controls yield different retrieval responses for centroid, SVD1, and energy-matched random-subspace removal (Table 15). The same intervention restores most of the mean margin under STR but little in the baseline; the recovered margin does not translate into retrieval ($+1.0$ R1 on MSR-VTT).

Table 4: Common-mode geometry and retrieval for gradient-matched counterparts on DiDeMo and MSR-VTT, seed 0. Each variant changes one term of the full configuration; vol and div rows replace L_{vol} and L_{div}. Emb. R1: text-to-video embedding-cosine R1; ITM R1 and R_{avg}: the main-table metrics; Cent. Δ: change in Emb. R1 under the removal intervention of Table 2. Per-seed results for all four counterparts at seeds 0–2, including selected checkpoint steps, are in App. G (Table 24).

	DiDeMo					MSR-VTT				
Configuration	$\|\mu_v\|^2$	Emb. R1	Cent. Δ	ITM R1	R_{avg}	$\|\mu_v\|^2$	Emb. R1	Cent. Δ	ITM R1	R_{avg}
Baseline	0.189	50.6	-5.1	65.7	80.8	0.124	41.0	-10.0	60.7	77.1
STR ($\sqrt{\det}$)	0.475	43.4	$+3.0$	66.0	80.7	0.448	39.5	$+1.1$	61.8	77.8
vol $\rightarrow$ pairwise cosine	0.463	44.2	$+2.7$	66.1	80.8	0.449	41.7	$+1.0$	61.1	77.5
vol $\rightarrow$ spectral	0.281	38.7	-2.1	64.7	80.6	0.177	9.7	-6.2	60.4	77.6
div $\rightarrow$ uniformity	0.008	38.2	$+0.9$	66.2	81.2	0.005	33.2	-0.5	61.3	77.7
div $\rightarrow$ VICReg variance	0.007	14.2	$+1.9$	66.7	81.3	0.001	2.1	$+0.4$	61.4	77.5

4.4 Form-Generality: Gradient-Matched Standard Counterparts

Is the shared-direction inflation a defect of the Gramian-volume form, or a property of matched-alignment pressure in general? To separate the two, the two canonical terms are replaced by standard counterparts and retrained under the identical protocol: L_{vol} by a *pairwise cosine* and a *spectral* objective ($-\log \lambda_{min}$ of the Gram matrix), L_{div} by the *uniformity* potential (Wang & Isola, 2020) and the *VICReg variance* hinge (Bardes et al., 2022); L_{sep}, nearly inactive at the fixed weights, is unchanged. Counterparts are matched to the canonical terms by embedding-gradient magnitude at a trained reference checkpoint, which controls local gradient scale, not accumulated optimization pressure over training (procedure, magnitudes, and a sensitivity run in App. G). Table 4 reports the seed-0 comparisons; App. G (Table 24) reports the three-seed results and selected checkpoints, distinguishing repeated patterns from seed-dependent variation.

The inflation is not specific to the Gramian volume. Pairwise cosine, which contains no determinant or rank term, produces similar common-mode energy and combined-removal responses to the canonical form in Table 4. Across seeds 0–2, its energy ranges from 0.447 to 0.449 on MSR-VTT and from 0.460 to 0.464 on DiDeMo, with Cent. Δ of $+0.9$ to $+1.0$ and $+1.9$ to $+2.7$ R1, respectively (Table 24). The observed energy inflation and combined-removal response are thus not exclusive to the Gram-volume form in these two VAST settings. The orthogonality-favoring spectral counterpart is a different control: its DN-minus-O response is negative in every dataset–seed cell (Table 24).

Similar ITM scores, different embedding retrieval. Table 4 illustrates large differences in embedding retrieval despite similar final ITM scores at seed 0. The variance counterpart repeats this contrast on MSR-VTT: embedding R1 is 2.1/2.9/2.7 at seeds 0/1/2, versus ITM R1 of $61.4/60.7/61.4$. On DiDeMo, diversity-counterpart geometry and embedding retrieval vary across seeds and selected checkpoints (Table 24), so the near-zero-energy observation does not extend to every selected model. Table 21 separately characterizes fused candidate coverage and conditional reranking success for the baseline and canonical STR; these score paths are distinct from the diagnostic text–video embedding path (App. F), and the measurements do not isolate the causal contributions of fusion and reranking. Final pipeline performance thus does not substitute for evaluating the embedding path used directly.

Retrieval performance does not select a form. In the seed-0 comparison of Table 4, the gradient-matched counterparts stay within 1.4 R1 of the canonical form on both datasets, and on DiDeMo three of the four exceed it: performance alone does not privilege any form. The forms differ sharply in geometry, and that difference is what motivates canonical choices.

4.5 Marginal Responses and Coupling: 2^3 Factorial Decomposition

Table 5 decomposes each observable over the 2^3 term subsets into per-term marginal effects. Each term's dominant axis is the same across all four datasets: L_{vol} dominates matched similarity (*matched cos. sim.*, the alignment observable) and the symmetric nearest-neighbor distance (*Sym. NN*), L_{div} dominates the mean margin (*pos-neg margin*) and maximum mean discrepancy (MMD), and L_{sep}

Table 5: Standardized average marginal effects (AME, normalized by across-combination SD) of each term on each observable, from the 2^3 factorial over all term subsets (seed 0). Each italic-headed block is one observable, one row per term; Sym. NN denotes the symmetric nearest-neighbor distance; formulations in App. L.2, effect computation in App. B. Bold marks the dominant term where one dominates on all four datasets.

Term	DiDeMo	MSR-VTT	VATEX	ActivityNet	**Term**	DiDeMo	MSR-VTT	VATEX	ActivityNet
matched cos. sim.					*MMD*				
L_{vol}	$+\mathbf{1.67}$	$+\mathbf{1.65}$	$+\mathbf{1.61}$	$+\mathbf{1.70}$	L_{vol}	-0.28	-0.31	-0.76	-1.00
L_{sep}	-0.55	-0.50	-0.59	-0.47	L_{sep}	$+0.78$	$+0.68$	$+0.68$	$+0.64$
L_{div}	$+0.41$	$+0.46$	$+0.43$	$+0.31$	L_{div}	$\mathbf{-1.40}$	$\mathbf{-1.42}$	$\mathbf{-1.25}$	$\mathbf{-1.14}$
Sym. NN					*pos-neg margin*				
L_{vol}	$\mathbf{-1.74}$	$\mathbf{-1.76}$	$\mathbf{-1.66}$	$\mathbf{-1.77}$	L_{vol}	-0.09	$+0.54$	$+0.42$	$+0.16$
L_{sep}	$+0.43$	$+0.28$	$+0.35$	$+0.33$	L_{sep}	-0.66	-0.58	-0.66	-0.35
L_{div}	-0.38	-0.43	-0.60	-0.30	L_{div}	$\mathbf{+1.31}$	$\mathbf{+1.11}$	$\mathbf{+1.12}$	$\mathbf{+1.43}$
intra-modal var.					*ITM R1*				
L_{vol}	-1.78	-1.66	-0.87	-1.44	L_{vol}	-0.26	$+1.10$	$+0.12$	$+0.83$
L_{sep}	$+0.09$	-0.10	-0.15	$+0.27$	L_{sep}	-0.52	$+0.08$	$+0.12$	-0.83
L_{div}	-0.14	$+0.14$	$+1.47$	$+1.11$	L_{div}	$+1.57$	-0.17	-0.35	$+1.14$

Table 6: Pairwise gradient cosines among the three terms at the full-configuration checkpoint (seed 0), averaged over 200 batches per dataset, at the embedding level (emb, w.r.t. embeddings) and the parameter level (par, w.r.t. trainable parameters). Entries shown as ~ 0 are of order 10^{-6}.

Dataset	emb L_{vol}/L_{sep}	emb L_{vol}/L_{div}	emb L_{sep}/L_{div}	par L_{vol}/L_{sep}	par L_{vol}/L_{div}	par L_{sep}/L_{div}
DiDeMo	~ 0	-0.489	-0.176	-0.690	-0.670	$+0.344$
MSR-VTT	~ 0	-0.409	-0.157	-0.615	-0.506	$+0.214$
VATEX	~ 0	-0.291	-0.181	-0.654	-0.229	$+0.099$
ActivityNet	~ 0	-0.368	-0.201	-0.708	-0.405	$+0.173$

dominates no axis, with a negative margin effect. The factorial quantifies alignment inflation: L_{vol} raises matched similarity dominantly yet contributes little, and with mixed sign, to the mean margin, which L_{div} supplies. Table 10 indicates the terms are coupled rather than independently controllable, with interaction-to-main-effect ratios reaching one on the margin. The geometric assignments repeat across datasets while the dominant term for ITM R1 varies; the subset variation of Table 9 shows this.

4.6 Gradient Relations and the Geometry-Performance Dissociation

Table 6 measures the pairwise gradient cosines. At the parameter level, the volume term opposes both other terms while the separation and diversity terms align with each other, on all four datasets. This mirrors the common-mode picture of Sec. 4.3 in an independent measurement. In the three-regime frame of App. K.2, (L_{vol}, L_{sep}) and (L_{vol}, L_{div}) are opposing pairs governed by weight ratios; these local checkpoint relations do not establish functional non-redundancy or the necessity of each term, especially given the small magnitude of L_{sep}.

L_{vol} and L_{sep} differ by level, near-orthogonal at the embedding level yet strongly opposed at the parameter level; the mechanism behind the embedding-level value is unverified, and all cosines are checkpoint-conditional first-order statistics over dependent batches. L_{sep} exerts little effective pressure here: the quantile hinge penalizes the top 20% of pairs by construction, but the excess above the threshold is small (loss of order 10^{-4}, gradient norm at most 1/30 of L_{vol}'s; measured over the batches of Table 6, not tabulated), so its cosines are angles between near-zero vectors, in line with the factorial evidence. L_{sep} is retained as a guard for over-alignment regimes.

Geometry-performance dissociation. Table 5 attributes the aggregate changes to individual terms. Under the fixed configuration, matched similarity and margin rise together and the modality gap (MMD and Sym. NN) closes on all four datasets (Baseline and Full rows of Table 11); split by subset,

Table 7: Shared-direction diagnostics on MSR-VTT (seed 0) for each additional backbone and model, computed on each backbone's retrieval-producing pooled embeddings. Columns: common-mode energy $\|\mu_v\|^2$, text–video centroid cosine, cos(u, SVD1), and embedding-cosine R1 before/after the removal intervention of Table 2. VAST rows repeat Sec. 4.3. Per-backbone provenance and the DiDeMo counterpart are in App. E; VALOR rows use the token-mean pooled embeddings the regularizer acts on, not its token-level ranking space.

Backbone	**Model**	$\|\mu_v\|^2$	$\cos(\mu_t, \mu_v)$	$\cos(u, \mathrm{SVD1})$	R1	R1 (removed)	ΔR1
VAST	Baseline	0.124	-0.024	0.832	41.0	31.0	-10.0
	STR	0.448	$+0.998$	1.000	39.5	40.6	$+1.1$
CLIP4Clip	Baseline	0.575	$+0.268$	0.998	43.7	33.7	-10.0
	STR	0.444	$+0.342$	0.998	43.8	37.3	-6.5
AVIGATE	Baseline	0.067	-0.333	0.706	41.7	40.4	-1.3
	STR	0.060	-0.358	0.706	41.4	40.8	-0.6
InternVideo2	Baseline	0.223	$+0.158$	0.956	41.9	39.1	-2.8
	STR	0.156	$+0.262$	0.969	42.6	39.6	-3.0
VALOR	Baseline	0.173	$+0.173$	0.934	37.5	38.0	$+0.5$
	STR	0.510	$+0.996$	1.000	35.1	35.2	$+0.1$

however, the terms that create matched similarity and the margin differ. Improved geometry need not improve retrieval: increasing λ_{vol} on MSR-VTT raises matched similarity and margin monotonically while R1 peaks at λ_{vol}=0.1 then declines (Table 12, App. C), consistent with the common-mode account of Sec. 4.3: increasing the mean margin does not necessarily improve top-1 retrieval, which depends on candidate ordering, not the mean margin alone. The same dissociation appears in Sec. 4.4, where sharply different geometries produce similar ITM scores in the seed-0 comparison of Table 4.

4.7 Diagnosis Transfer to Additional Backbones

Table 7 repeats the shared-direction diagnosis of Sec. 4.3, without re-tuning, on the retrieval embeddings of four additional backbones: CLIP4Clip (Luo et al., 2022), AVIGATE (Jeong et al., 2025), InternVideo2 (Wang et al., 2024), and VALOR (Chen et al., 2023a). Different components of the pattern transfer to different backbones. VALOR shows the common-mode energy inflation in its regularized pooled space, its text and video centroids merging, while CLIP4Clip shows the baseline's removal sensitivity on two datasets (-10.0 R1 as in VAST, -9.7 on DiDeMo; App. E) without the inflation. Neither component appears in AVIGATE or InternVideo2, and no backbone reproduces the full VAST pattern. Encoder update magnitude (the CLIP-based encoders train at an effective 10^{-7}) and the gap between regularized and ranking spaces (AVIGATE and InternVideo2 rank on fusion outputs downstream of the regularized embeddings) are possible explanations, neither tested here: raising the CLIP4Clip encoder learning rate degraded both models under the tested settings, leaving the question open (App. E). For CLIP4Clip's K=2 setting, L_{vol} favors collinearity ($\det G = 1 - \cos^2\theta$, minimized at $|\cos\theta| \to 1$), a K=2 control rather than evidence about higher-order coupling. Table 20 (App. E) lists retrieval numbers for STR on these backbones under each repository's own protocol.

5 Conclusion

This paper uses STR as an instrument for auditing multimodal embedding geometry. Across four VAST benchmarks and three seeds per benchmark, baseline and canonical STR differ in shared-direction removal responses; separating deletion from residual-norm rescaling is necessary to interpret this comparison. On MSR-VTT and DiDeMo, pairwise attraction also yields energy inflation and a similar combined-removal response across three seeds per dataset. Aggregate recall preservation does not establish information absence or ranking preservation. Final fused-and-reranked scores can remain similar despite distinct embedding-space behavior. Additional backbones reproduce components, but not the full VAST pattern, and the conditions governing its occurrence remain unestablished. Identifying those conditions and testing whether the inflation can be suppressed while alignment is retained are direct extensions of these measurements. These findings motivate evaluating embedding retrieval separately from final pipeline performance and interpreting removal experiments through their individual components.

Reproducibility statement

Section 4.1 and App. L.1 specify the training protocol, the fixed hyperparameters, the held-out checkpoint-selection rule, and the missing-modality handling; App. L.1 describes the numerical implementation, the scope of the reported stability checks, and the gradient path of the L_{sep} threshold. Appendix G states the exact counterpart formulas, their implementation sources, and the gradient-matching procedure with measured magnitudes; it also reports the counterpart results by seed together with the selected checkpoint steps. Appendix L.2 gives the diagnostic metric formulations. Dataset splits and statistics are in App. A.

Ethics statement

This work concerns geometric regularization and diagnosis of multimodal embeddings for video-text retrieval on standard public benchmarks and does not introduce new capabilities with foreseeable risks beyond those of standard contrastive learning systems.

AI use statement

In this work, generative AI tools were used for experiment orchestration and analysis code, for aggregating experimental results, and for drafting and revising the manuscript text under author direction. All AI-assisted analyses were re-verified against raw experiment logs, all reported numbers were checked against the result files by the authors, and AI-drafted text was reviewed and edited by the authors. The authors take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1728–1738, 2021.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022.

Shubhang Bhatnagar and Narendra Ahuja. Potential field based deep metric learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25549–25559, 2025.

Huiyuan Chen, Vivian Lai, Hongye Jin, Zhimeng Jiang, Mahashweta Das, and Xia Hu. Towards mitigating dimensional collapse of representations in collaborative filtering. In *Proceedings of the 17th ACM international conference on web search and data mining*, pp. 106–115, 2024.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022.

Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 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. *Advances in Neural Information Processing Systems*, 36:72842–72866, 2023b.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

Giordano Cicchetti, Eleonora Grassucci, and Danilo Comminiello. A triangle enables multimodal alignment beyond cosine similarity. In *Advances in Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=3Hjfzh5Eyk. NeurIPS 2025 Poster.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=ftGnpZrW7P. ICLR 2025 Poster.

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, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Felix R Gantmacher. Matrix theory. *Chelsea, New York*, 21:6, 1959.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Junlin He, Jinxiao Du, and Wei Ma. Preventing dimensional collapse in self-supervised learning via orthogonality regularization. *Advances in Neural Information Processing Systems*, 37:95579–95606, 2024.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Fabian 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*, pp. 961–970, 2015.

Lisa 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*, pp. 5803–5812, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12054–12064, 2023.

Boseung Jeong, Jicheol Park, Sungyeon Kim, and Suha Kwak. Learning audio-guided video representation with gated attention for video-text retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26202–26211, 2025.

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

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

Jianlong Kwan, Jiayu Xiong, Dilong Li, and Jing Wang. Self-supervised feature re-representation via lennard-jones potential loss. OpenReview, 2024. URL https://openreview.net/forum?id=BgvAzuCfHc. ICLR 2025 Conference, withdrawn submission.

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

Rongmei Lin, Weiyang Liu, Zhen Liu, Chen Feng, Zhiding Yu, James M Rehg, Li Xiong, and Le Song. Regularizing neural networks via minimizing hyperspherical energy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6917–6927, 2020.

Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. Eclipse: Efficient long-range video retrieval using sight and sound. In *European Conference on Computer Vision*, pp. 413–430. Springer, 2022.

Zudi Lin, Erhan Bas, Kunwar Yashraj Singh, Gurumurthy Swaminathan, and Rahul Bhotika. Relaxing contrastiveness in multimodal representation learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2227–2236, 2023.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in neural information processing systems*, 34:18878–18890, 2021a.

Weiyang Liu, Rongmei Lin, Zhen Liu, Lixin Liu, Zhiding Yu, Bo Dai, and Le Song. Learning towards minimum hyperspherical energy. *Advances in neural information processing systems*, 31, 2018.

Weiyang Liu, Rongmei Lin, Zhen Liu, Li Xiong, Bernhard Schölkopf, and Adrian Weller. Learning with hyperspherical uniformity. In *International Conference On Artificial Intelligence and Statistics*, pp. 1180–1188. PMLR, 2021b.

Weiyang Liu, Longhui Yu, Adrian Weller, and Bernhard Schölkopf. Generalizing and decoupling neural collapse via hyperspherical uniformity gap. *arXiv preprint arXiv:2303.06484*, 2023.

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.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9879–9889, 2020.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*, 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

Adriel Saporta, Aahlad Manas Puli, Mark Goldstein, and Rajesh Ranganath. Contrasting with symile: Simple model-agnostic representation learning for unlimited modalities. *Advances in Neural Information Processing Systems*, 37:56919–56957, 2024.

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

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Xin Wang, Jiawei 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/CVF international conference on computer vision*, pp. 4581–4591, 2019.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pp. 396–416. Springer, 2024.

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*, pp. 5288–5296, 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.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15671–15680, 2022.

Can Yaras, Siyi Chen, Peng Wang, and Qing Qu. Explaining and mitigating the modality gap in contrastive multimodal learning. *arXiv preprint arXiv:2412.07909*, 2024.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Zhen Wan, Jinchuan Tian, An-Chieh Cheng, Ligeng Zhu, Yuanhang Su, Yuming Lou, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. In *International Conference on Learning Representations*, volume 2026, pp. 56101–56138, 2026.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Ziruo Zhang, JIAWEI DU, Zuozhu Liu, Joey Tianyi Zhou, et al. Beyond modality collapse: Representation blending for multimodal dataset distillation. *Advances in Neural Information Processing Systems*, 38:114809–114829, 2026.

APPENDIX

A DATASET AND EVALUATION DETAILS

The four benchmarks are VATEX (Wang et al., 2019) (27,053 videos, 218K captions), MSR-VTT (Xu et al., 2016) (10,000 videos, 200K captions), DiDeMo (Hendricks et al., 2017) (9,399 videos), and ActivityNet (Heilbron et al., 2015) (14,926 videos). The evaluation sets used are the MSR-VTT 1k-A test set (1,000 videos), DiDeMo (1,003), VATEX (2,500), and ActivityNet (4,483), with training on the corresponding standard training splits (MSR-VTT 9,000 videos).

Subtitle coverage, which determines where a $K{=}4$ setting is measurable, is: MSR-VTT 76.6% of training and 76.2% of test videos; VATEX 49.9% of training and 0% of test videos; DiDeMo and ActivityNet have no subtitle field. A fraction of instances lacks audio (11.9% of MSR-VTT training videos); these are handled by the per-sample presence masks of Sec. 3.

Why K is unified at 3. The $t\text{-}va$ setting is used on all datasets for three reasons, the first two following from the subtitle coverage above: DiDeMo and ActivityNet lack the field; a subtitle-based ($t\text{-}vas$) VATEX model would be evaluated with an empty subtitle stream, so its numbers could not be read as evidence about subtitles; and mixing K across datasets changes the scale of $\det(G_i)$, preventing a single λ_{vol} from transferring across datasets. MSR-VTT is the only dataset where $K{=}4$ is measurable. A single MSR-VTT $K{=}4$ ($t\text{-}vas$) run with λ_{vol} and λ_{div} rematched to their $K{=}3$ gradient magnitudes (0.087 and 1.074; λ_{sep} unchanged) trains stably and reaches 63.3 ITM R1, 1.5 above the $K{=}3$ result on the same dataset; no matched $t\text{-}vas$ baseline was trained, so the number reflects feasibility at $K{=}4$; attributing the difference to subtitles would require that baseline.

Table 8: Full text-to-video recall (ITM) behind Table 1, as mean ±SD over three seeds $(0, 1, 2)$. Bold marks the best mean per dataset and metric; ties are all marked.

	DiDeMo			MSR-VTT		
Method	R1	R5	R10	R1	R5	R10
Baseline	66.2 ±0.92	86.0 ±0.67	90.1 ±0.58	60.7 ±0.35	82.3 ±0.47	88.3 ±0.25
GRAM	65.9 ±1.13	**86.3** ±0.26	**90.2** ±0.29	59.8 ±0.78	82.4 ±0.46	88.7 ±0.50
Symile	64.1 ±1.31	84.3 ±0.51	88.7 ±0.17	57.6 ±0.85	81.9 ±0.12	87.4 ±0.50
TRIANGLE	**66.5** ±0.85	86.1 ±0.35	90.0 ±0.36	60.0 ±0.29	**82.9** ±0.71	**88.9** ±0.47
STR	66.1 ±0.23	**86.3** ±0.60	90.1 ±0.36	**61.0** ±1.06	**82.9** ±0.62	88.4 ±0.17
	VATEX			**ActivityNet**		
Baseline	**69.1** ±0.25	**92.0** ±0.06	**96.4** ±0.00	66.7 ±0.80	89.4 ±0.69	95.4 ±0.06
GRAM	**69.1** ±0.15	91.9 ±0.06	96.3 ±0.06	**67.3** ±0.67	89.5 ±0.75	95.2 ±0.32
Symile	66.6 ±0.15	90.4 ±0.06	95.1 ±0.06	63.5 ±0.75	87.3 ±0.46	93.7 ±0.31
TRIANGLE	**69.1** ±0.20	91.9 ±0.00	**96.4** ±0.10	66.6 ±0.53	89.5 ±0.30	95.2 ±0.25
STR	**69.1** ±0.17	91.9 ±0.00	**96.4** ±0.00	67.1 ±0.55	**89.6** ±0.21	**95.5** ±0.21

Table 9: Text-to-video retrieval for all subsets of $\{L_{vol}, L_{sep}, L_{div}\}$ under the t-va setting, seed 0; each entry is ITM R1 (R_{avg}). Datasets are abbreviated as DiD (DiDeMo), MSR (MSR-VTT), VTX (VATEX), and ACT (ActivityNet). Bold marks the best value per row.

Dataset	Baseline	L_{vol}	L_{sep}	L_{div}	L_{vol}+L_{sep}	L_{vol}+L_{div}	L_{sep}+L_{div}	Full
DiD	65.7 (80.8)	65.7 (80.7)	65.2 (80.8)	66.1 (81.0)	65.2 (80.7)	65.9 (80.8)	**66.2 (81.3)**	66.0 (80.7)
MSR	60.7 (77.1)	61.1 (77.8)	60.3 (77.3)	60.3 (77.4)	60.6 (77.5)	60.3 (77.5)	59.9 (77.0)	**61.8 (77.8)**
VTX	68.9 (85.8)	**69.5 (86.0)**	69.4 (85.9)	69.4 (85.9)	69.4 (86.0)	69.2 (85.9)	69.3 (85.9)	69.0 (85.7)
ACT	65.9 (83.3)	66.6 (83.7)	65.9 (83.4)	**67.1 (83.9)**	66.2 (83.6)	66.9 (83.8)	66.0 (83.4)	66.8 (83.9)

Evaluation details. Table 8 lists the full text-to-video recall (ITM) behind Table 1. Dual-encoder results appear in two distinct score paths: the text–video embedding-cosine diagnostics of Sec. 4.3, where the intervention operates directly on embeddings and cross-encoder scores are structurally unaffected, and the ITC (text–fused) retrieval of App. F; the paths score different representations and are defined in App. F.

B Component Combinations and Factorial Analysis

Table 9 reports retrieval for all subsets of the three terms. The best subset differs on every dataset and no subset dominates (L_{sep}+L_{div} on DiDeMo, the full triple on MSR-VTT, L_{vol} on VATEX, L_{div} on ActivityNet); on MSR-VTT the full configuration is the best of the tested subsets, though L_{vol} alone also exceeds the baseline, while on VATEX every subset is at or above the full triple. These subsets use test-oracle access and are a diagnostic upper envelope, not a deployable rule, and no principled subset-selection rule emerges from the variation; the fixed three-term configuration is nonetheless within 0.5 R1 of the per-dataset best on every dataset, so the reported behavior is obtained without per-dataset subset search. Each cell is a single run at seed 0, so the per-dataset subset rankings are single-run observations that may shift across seeds.

Effect computation. The factorial of Table 5 decomposes each geometric observable measured on the eight term subsets of Table 9; the per-subset raw values are in Table 11. For each observable, the AME of a term is the average change in the observable when that term is switched on, taken over the four on/off settings of the other two terms (equivalently, the mean over the four subsets containing the term minus the mean over the four without it). Interaction effects I use product contrasts on the same difference-of-means scale, so their magnitudes are directly comparable to the main effects. Standardized values divide each effect by the observable's sample SD across the eight configurations. The eight effects reconstruct the eight configuration values exactly, and Table 10 reports the resulting interaction-to-main-effect ratios.

Table 10: Interaction-to-main-effect ratio $\max|I|/\max|\mathrm{AME}|$ per observable from the 2^3 factorial (seed 0), with the dominant interaction. Bold marks ratios near or above one.

Observable	DiDeMo	MSR-VTT	VATEX	ActivityNet	Dominant interaction
matched cos. sim.	0.26	0.30	0.36	0.29	$I_{vol,sep}$
MMD	0.45	0.46	0.54	0.57	$I_{vol,sep}$
Sym. NN	0.17	0.16	0.21	0.21	$I_{vol,sep}$ / $I_{vol,div}$
intra-modal var.	0.28	0.48	0.50	0.19	$I_{vol,div}$
pos-neg margin	0.73	**1.00**	**0.98**	0.70	$I_{vol,div}$
ITM R1	0.50	0.77	**3.67**	0.64	$I_{sep,div}$ / $I_{vol,div}$

Table 11: Per-subset geometric diagnostics (text-video), seed 0. Columns per dataset: matched similarity (matched), maximum mean discrepancy (MMD), symmetric nearest-neighbor distance (Sym. NN), intra-modal variance (var), positive-negative margin (margin), and ITM R1; formulations in App. L.2.

	DiDeMo						**MSR-VTT**					
Subset	matched	MMD	Sym. NN	var	margin	R1	matched	MMD	Sym. NN	var	margin	R1
Baseline	0.209	0.185	1.252	0.773	0.212	65.7	0.171	0.218	1.280	0.665	0.175	60.7
L_{vol}	0.582	0.462	0.843	0.572	0.004	65.7	0.497	0.568	0.868	0.591	0.001	61.1
L_{sep}	−0.225	0.870	1.533	0.776	0.037	65.2	−0.306	0.935	1.535	0.632	−0.003	60.3
L_{div}	0.231	0.099	1.232	0.833	0.232	66.1	0.174	0.152	1.276	0.716	0.174	60.3
L_{vol}+L_{sep}	0.504	0.548	0.930	0.635	0.002	65.2	0.488	0.583	0.881	0.588	0.001	60.6
L_{vol}+L_{div}	0.737	0.029	0.700	0.518	0.269	65.9	0.771	0.024	0.635	0.536	0.323	60.3
L_{sep}+L_{div}	−0.044	0.251	1.424	0.813	0.101	66.2	−0.136	0.288	1.417	0.709	−0.003	59.9
Full	0.736	0.030	0.702	0.518	0.268	66.0	0.773	0.023	0.629	0.552	0.336	61.8
	VATEX						**ActivityNet**					
Baseline	0.298	0.296	1.117	0.487	0.321	68.9	0.240	0.362	1.232	0.532	0.235	65.9
L_{vol}	0.670	0.400	0.725	0.363	0.001	69.5	0.733	0.278	0.687	0.423	0.008	66.6
L_{sep}	−0.330	1.065	1.479	0.460	0.019	69.4	−0.295	0.999	1.601	0.611	0.063	65.9
L_{div}	0.350	0.169	1.069	0.529	0.352	69.4	0.279	0.158	1.199	0.691	0.282	67.1
L_{vol}+L_{sep}	0.669	0.401	0.725	0.363	0.001	69.4	0.734	0.277	0.687	0.423	0.008	66.2
L_{vol}+L_{div}	0.949	0.009	0.304	0.517	0.555	69.2	0.845	0.019	0.549	0.518	0.386	66.9
L_{sep}+L_{div}	−0.126	0.304	1.307	0.519	0.001	69.3	−0.036	0.315	1.436	0.691	0.161	66.0
Full	0.948	0.009	0.308	0.513	0.550	69.0	0.864	0.016	0.512	0.555	0.442	66.8

Table 11 gives the per-subset diagnostics that the factorial above decomposes. Four patterns hold on all datasets: L_{vol} alone raises matched similarity sharply while collapsing the margin to near zero (0.212 to 0.004, 0.175 to 0.001, 0.321 to 0.001, 0.235 to 0.008), the most direct evidence of alignment inflation; L_{vol}+L_{sep} behaves the same, so L_{sep} does not prevent the collapse; the margin is restored by L_{div}, with L_{vol}+L_{div} combining top-level matched similarity (0.74 to 0.95) with above-baseline margin; and L_{sep} alone drives matched similarity negative (-0.23 to -0.33) and widens the modality gap, acting as designed but unhelpfully for retrieval.

C Hyperparameter Sweeps

Table 12 reports retrieval under the λ_{vol}/λ_{sep}/τ-cap sweep together with the geometric diagnostics of the same checkpoints. Two observations follow. First, raising λ_{sep} to 10 drives matched similarity negative on DiDeMo (-0.320 and -0.358, margin also negative): L_{sep} acts as designed when given sufficient weight, but in a direction opposed to the mean margin; at the default weight its contribution is negligible (Sec. 4.6). Second, on MSR-VTT the geometric indicators improve monotonically in λ_{vol} while R1 falls past λ_{vol}=0.1. The internal constants q and (r_{min}, σ) were fixed from limited exploration and are treated as fixed design constants of the reference configuration; the functional-form question that hyperparameter sweeps cannot answer is addressed by the counterpart study of Sec. 4.4 and App. G.

Table 12: λ_{vol} / λ_{sep} / τ-cap sweep, reusing the main protocol (held-out selection, seed 0): retrieval (ITM, left block) alongside the geometry of the same checkpoints (right block; Sym. NN denotes the symmetric nearest-neighbor distance) (formulations in App. L.2). The per-dataset optimum differs (DiDeMo prefers λ_{vol}=0.5, MSR-VTT prefers 0.1). Bold marks the best R1 and R_{avg} per dataset within the sweep.

	Retrieval					Geometry				
λ_{vol} / λ_{sep} / τ cap	R1	R5	R10	R_{avg}	ΔR1	matched	margin	MMD	Sym. NN	intra var.
DiDeMo										
0.1 / 0.01 / 1.0	65.9	86.2	90.0	80.7	+0.2	0.736	0.269	0.029	0.700	0.518
0.1 / 0.1 / 1.0 (default)	66.0	86.2	89.8	80.7	+0.3	0.736	0.268	0.030	0.702	0.518
0.1 / 1.0 / 1.0	66.3	85.9	90.1	80.8	+0.6	0.726	0.259	0.031	0.717	0.518
0.1 / 10 / 0.70	66.4	86.5	90.4	81.1	+0.7	−0.320	−0.130	0.510	1.443	0.522
0.1 / 10 / 1.0	67.2	87.0	90.3	81.5	+1.5	−0.358	−0.261	0.385	1.347	0.583
0.5 / 0.1 / 1.0	**68.0**	86.2	90.4	**81.6**	+2.3	0.825	0.264	0.024	0.560	0.430
0.5 / 10 / 0.85	66.2	85.9	90.3	80.8	+0.5	0.830	0.405	0.020	0.538	0.552
0.5 / 10 / 1.0	65.8	85.9	89.9	80.6	+0.1	0.861	0.449	0.016	0.463	0.566
MSR-VTT										
0.025 / 0.1 / 1.0	60.4	83.1	88.8	77.4	−0.3	0.673	0.279	0.032	0.772	0.603
0.1 / 0.1 / 1.0 (default)	**61.8**	83.4	88.3	**77.8**	+1.1	0.773	0.336	0.023	0.629	0.552
0.5 / 0.1 / 1.0	60.2	82.4	88.2	76.9	−0.5	0.857	0.627	0.015	0.420	0.741
1.0 / 0.1 / 1.0	59.9	83.0	88.4	77.1	−0.8	0.878	0.687	0.012	0.332	0.777

Table 13: Common-mode energy $\|\mu_v\|^2$ across three seeds $(0, 1, 2$; mean $\pm$SD).

	$\|\mu_v\|^2$ (3 seeds)	
Dataset	Baseline	STR
DiDeMo	0.181 ±0.007	0.446 ±0.061
MSR-VTT	0.121 ±0.002	0.456 ±0.012
VATEX	0.230 ±0.001	0.381 ±0.001
ActivityNet	0.259 ±0.050	0.435 ±0.014

D Common-Mode Analysis: Details and Controls

Table 13 reports the common-mode energy across three seeds. Table 16 quantifies the centroid overlap of Sec. 4.3: the video and text centroids are nearly orthogonal in the baseline ($|\cos| \leq 0.08$) and overlap under the full configuration ($\cos \geq 0.99$). Table 17 gives the per-dataset numbers behind the 66-to-89-percent margin recovery. Table 15 details the control runs for both models.

In Table 15, random single-direction controls produce small changes in aggregate R1 (41.2 ±0.25 for the baseline and 39.5 ±0.12 for STR over 20 draws, each direction's energy 0.001 to 0.002). Energy-matched random-subspace removal gives R1 of 38.6 for the baseline and 38.9 for STR, compared with 31.0 and 40.6 under centroid DN. These observations distinguish the tested removals, but the energy-matched subspaces remove 77 and 231 dimensions, respectively, rather than one, so they do not isolate direction from all other geometric changes. The near-preservation under centroid DN concerns deletion followed by residual-norm rescaling (DN), not deletion alone: at seed 0, D changes STR R1 from 39.5 to 36.8, whereas DN changes it to 40.6. The HM column of the same table shows why the mean margin overstates ranking change: under rescaling alone, the STR mean margin rises to 0.607 while the hard-negative margin falls to −0.181 and R1 to 13.9. The two models also differ in how the centroid relates to the top singular direction: in the baseline the two are distinct ($\cos = 0.83$, Table 2) and their removals move R1 in opposite directions (−10.0 vs. +4.4, Table 15), whereas under STR they coincide ($\cos = 1.00$) and both removals leave aggregate R1 nearly unchanged.

Table 14: Removal decomposition on text-video embedding R1, mean ±SD over three seeds $(0, 1, 2)$, for each model. O: original scores; D: the shared direction's score contribution deleted, no rescaling; N: residual-norm rescaling only, no deletion; DN: deletion followed by rescaling (the intervention of Table 2). Gain and Loss are the percent of queries whose top-1 result becomes correct or incorrect under DN.

Dataset	**Model**	O	D	N	DN	Gain	Loss
DiDeMo	Baseline	50.9 ±0.6	44.8 ±0.7	50.7 ±0.3	45.7 ±1.0	2.8	8.0
	STR	42.6 ±2.9	43.9 ±7.8	7.7 ±8.5	45.4 ±5.6	4.9	2.1
MSR-VTT	Baseline	41.8 ±0.8	30.9 ±0.9	41.8 ±0.6	31.4 ±0.5	4.2	14.6
	STR	40.4 ±0.8	38.7 ±2.8	15.2 ±1.7	41.4 ±1.6	3.9	2.8
VATEX	Baseline	88.0 ±0.4	79.9 ±0.2	88.0 ±0.4	81.4 ±0.5	1.4	8.0
	STR	87.2 ±0.4	76.8 ±0.2	62.2 ±0.4	87.1 ±0.4	1.3	1.4
ActivityNet	Baseline	49.8 ±0.7	33.1 ±7.5	50.7 ±0.4	34.4 ±7.8	2.8	18.3
	STR	46.9 ±0.5	42.3 ±1.1	17.4 ±0.5	47.2 ±0.7	3.0	2.7

Table 15: Intervention controls on MSR-VTT (seed 0), text-video embedding retrieval, for the baseline and the full STR configuration. MM and HM denote the mean margin and the hard-negative margin of App. L.2; the random-direction row reports R1 as mean ±SD over 20 draws, with margins from the first draw. The centroid direction carries 0.142 of the embedding energy in the baseline and 0.444 under STR; the energy-matched random subspace has 77 and 231 dimensions, respectively. The top singular direction is that of the uncentered embeddings (SVD1 of Table 2).

	Baseline			STR		
Intervention	R1	MM	HM	R1	MM	HM
None (original)	41.0	0.176	−0.017	39.5	0.335	−0.029
Centroid deletion + rescaling (DN)	31.0	0.207	−0.042	40.6	0.601	−0.047
Centroid deletion only (D)	30.2	0.174	−0.038	36.8	0.334	−0.036
Rescaling only (N)	41.2	0.210	−0.018	13.9	0.607	−0.181
Random single direction (20 draws)	41.2 ±0.25	0.176	−0.017	39.5 ±0.12	0.335	−0.029
Random subspace, energy-matched	38.6	0.174	−0.022	38.9	0.333	−0.032
Top singular direction (SVD1)	45.4	0.227	−0.008	40.4	0.601	−0.047

Score-level definitions of the removal operations. Decompose the unit text and video embeddings along the unit shared centroid direction u and its orthogonal complement,

$$\tilde{z}_i^{(t)} = a_i u + r_i, \qquad \tilde{z}_j^{(v)} = b_j u + v_j, \qquad u^\top r_i = u^\top v_j = 0, \tag{6}$$

where i indexes queries and j candidates; u is extracted by the same rule per model and seed, so it is numerically different for the baseline and STR. With residual norms $p_i = \|r_i\|_2$ and $q_j = \|v_j\|_2$ and residual cosine $h_{ij} = r_i^\top v_j/(p_i q_j)$, the four score matrices are

$$O_{ij} = a_i b_j + p_i q_j h_{ij}, \qquad D_{ij} = p_i q_j h_{ij}, \qquad N_{ij} = h_{ij} + \frac{a_i}{p_i}\frac{b_j}{q_j}, \qquad DN_{ij} = h_{ij}, \tag{7}$$

so $N_{ij} - DN_{ij} = (a_i/p_i)(b_j/q_j)$. Three consequences follow. First, N is not a generic ℓ_2 normalization: it rescales the direction-retaining scores by the residual norms that DN uses, a contrast operation. Second, within one query a_i/p_i is constant but b_j/q_j varies across candidates, so rescaling with the direction retained adds a candidate-dependent score term that can reorder candidates. Third, a per-query positive scale or offset changes margin magnitudes while preserving candidate order, so a mean-margin change alone cannot establish a ranking change; the candidate-dependent term above is the part that can. In the implementation the residual norms are clamped at 10^{-8}; the logged fraction of near-zero residuals is zero in all reported runs.

Table 14 gives the deletion/rescaling decomposition of Sec. 4.3 on all four datasets. The paired removal response Γ is the difference between the two models' DN−O changes; computed from the table's mean columns it is $+6.5$ to $+15.8$ R1 across the four datasets, and it is positive for each seed individually.

Table 16: Cosine between the video and text centroids, seed 0. Bold marks $\cos \geq 0.99$.

Dataset	Baseline	L_{vol}	L_{div}	$L_{vol}+L_{div}$	Full STR
DiDeMo	-0.019	0.902	-0.009	**0.993**	**0.993**
MSR-VTT	-0.024	0.794	0.004	**0.997**	**0.998**
VATEX	-0.076	0.906	-0.030	**0.999**	**0.999**
ActivityNet	0.016	0.966	-0.034	**0.997**	**0.997**

Table 17: Mean positive-negative margin before $\rightarrow$ after removing the shared centroid direction (with renormalization), seed 0. Percentages are relative recovery, omitted where the pre-removal margin is near zero.

Config.	DiDeMo	MSR-VTT	VATEX	ActivityNet
Baseline	0.213 $\rightarrow$ 0.237 (+11%)	0.176 $\rightarrow$ 0.207 (+18%)	0.321 $\rightarrow$ 0.376 (+17%)	0.234 $\rightarrow$ 0.295 (+26%)
L_{vol}	0.004 $\rightarrow$ 0.030	0.001 $\rightarrow$ 0.003	0.001 $\rightarrow$ 0.007	0.008 $\rightarrow$ 0.107
L_{div}	0.232 $\rightarrow$ 0.253 (+9%)	0.174 $\rightarrow$ 0.214 (+23%)	0.352 $\rightarrow$ 0.416 (+18%)	0.282 $\rightarrow$ 0.334 (+18%)
$L_{vol}+L_{div}$	0.270 $\rightarrow$ 0.510 (+89%)	0.322 $\rightarrow$ 0.590 (+83%)	0.555 $\rightarrow$ 0.916 (+65%)	0.385 $\rightarrow$ 0.714 (+85%)
Full STR	0.269 $\rightarrow$ 0.508 (+89%)	0.336 $\rightarrow$ 0.602 (+79%)	0.550 $\rightarrow$ 0.914 (+66%)	0.441 $\rightarrow$ 0.764 (+73%)

Two observations go beyond Sec. 4.3. First, on DiDeMo, the reported removal of the top principal component of the centered embeddings changes the full-configuration mean margin from 0.269 to 0.239 (not tabulated), unlike the increase under shared-centroid removal in Table 17. This compares mean-margin responses; it does not establish a unique causal direction for retrieval behavior. Second, the margin rises under removal in every configuration of Table 17, so the margin alone hides the baseline-STR dissociation that the retrieval numbers of Table 2 expose. The shared direction's direct contribution to the margin is near zero in all configurations (at most 0.003 in absolute value, or 1.1% of the margin; Table 18): the common component raises positive and negative similarities together, cancelling in their difference.

E Shared-Direction Diagnostics on the Additional Backbones

The pooled embeddings behind Table 7 are each backbone's retrieval-producing embeddings, except VALOR, whose regularized token-mean pooled space differs from its token-level ranking space: the end-of-text token embedding and the frame-mean video embedding for CLIP4Clip, the audio-gated fused video embedding for AVIGATE (whose full score adds a frame-level term, so its single-vector R1 of 41.7 is below the reported 44.5), the audio-visual fusion output for InternVideo2, and the token-mean pooled embeddings for VALOR. In InternVideo2 the regularizer acts on the vision and audio projections before fusion; their energies move from 0.35 to 0.26 and from 0.46 to 0.28, so the inflation is absent in the regularized space as well. The AVIGATE and CLIP4Clip runs use the CLIP4Clip optimizer convention of a 10^{-3} learning-rate coefficient on all CLIP parameters. Table 19 repeats the diagnostics for these two backbones on DiDeMo: CLIP4Clip reproduces the baseline's large removal sensitivity (ΔR1 -9.7 vs. -10.0 on MSR-VTT, with the baseline more removal-sensitive than STR, -9.7 vs. -7.2), and AVIGATE stays near-null on both datasets (ΔR1 -2.0 vs. -1.3), so the component-wise pattern of these two backbones holds on a second dataset.

Learning-rate probe. To test the frozen-encoder explanation causally, CLIP4Clip was retrained on MSR-VTT (seed 0, 5 epochs, baseline and STR) with the learning-rate coefficient raised from 10^{-3} to 10^{-2} and 10^{-1}, giving effective encoder rates of 10^{-6} and 10^{-5} in place of the repository's 10^{-7}. Both settings degrade both models: best selected R1 is 29.1/29.6 (baseline/STR) at 10^{-6} and 10.0/10.4 at 10^{-5}, against 43.6/43.9 at the repository rate (Table 20). Under the tested optimizer, schedule, and five-epoch setting, both tested increases degraded both models, so these runs leave the proposed explanation untested and say nothing about geometry under a trainable encoder.

Table 18: Direct contribution of the shared centroid direction u to the mean margin, seed 0. With a_i and b_j the components of the text and video embeddings along u, the direction's score contribution is the rank-one term $a_i b_j$; the table reports the $\langle$positive$\rangle - \langle$negative$\rangle$ margin share carried by this term. All entries are near zero although the margins of Table 17 range from 0.001 to 0.55.

Dataset	Baseline	L_{vol}	L_{div}	$L_{vol}+L_{div}$	Full STR
DiDeMo	$+0.002$	0.000	$+0.001$	$+0.001$	$+0.001$
MSR-VTT	$+0.002$	0.000	0.000	$+0.001$	$+0.001$
VATEX	0.000	0.000	0.000	$+0.003$	$+0.003$
ActivityNet	$+0.001$	0.000	0.000	$+0.001$	$+0.001$

Table 19: Shared-direction diagnostics on DiDeMo (seed 0) for each model of the two CLIP-family backbones, the second-dataset counterpart of the MSR-VTT rows in Table 7 (columns as there; R1 is single-vector embedding-cosine R1, which for AVIGATE is below its reported score as noted above). InternVideo2 and VALOR are not included: their DiDeMo retrieval vectors were not dumped.

Backbone	**Model**	$\|\mu_v\|^2$	$\cos(\mu_t, \mu_v)$	$\cos(u, \mathrm{SVD1})$	R1	R1 (removed)	ΔR1
CLIP4Clip	Baseline	0.560	0.285	0.980	40.2	30.5	-9.7
	STR	0.434	0.322	0.991	40.2	33.0	-7.2
AVIGATE	Baseline	0.089	0.128	0.925	35.0	33.0	-2.0
	STR	0.084	0.156	0.932	35.6	33.7	-1.9

Retrieval performance under each backbone's own protocol. Table 20 reports what STR does to retrieval when added to these backbones on MSR-VTT and DiDeMo, each under its own repository's evaluation protocol and hard-coded model-selection rule (R1 for AVIGATE, CLIP4Clip, and InternVideo2; R_{avg} for VALOR); these protocols differ from Table 1 and are not directly comparable to it. On MSR-VTT the selected R1 is higher with STR in all four single runs, by margins ($+0.3$ to $+1.1$) on the scale of the seed SDs of Table 1, and R_{avg} improves on one of four, so the direction is read as repeated rather than as per-backbone gains. On DiDeMo the direction is mixed: InternVideo2 and AVIGATE rise ($+0.7$, $+0.1$), CLIP4Clip is flat (-0.1), and VALOR falls 2.8; the repeated MSR-VTT direction is therefore a dataset-level observation rather than a backbone property.

F Pipeline Stages and Dual-Encoder Retrieval

Score paths. Four retrieval scores appear in the paper; all use the same finetuned checkpoint per run (the held-out-selected step) and each dataset's standard test split, and where a video has multiple captions the caption embeddings are averaged into one query and renormalized. (i) The diagnostic text–video cosine (Sec. 4.3, Sec. 4.4, App. D) scores the ℓ_2-normalized text and video contrastive embeddings over the full gallery with one ground truth per query; the removal operations and controls change only this scoring operation, holding the query set, gallery, and ground-truth mapping fixed. (ii) The ITC retrieval of Table 22 scores the text embedding against a separate fused embedding, built by concatenating the pooled vision and audio outputs and passing them through their own contrastive head; the MSR-VTT seed-0 gap between the two paths (R1 41.0 vs. 47.0 for the baseline, 39.5 vs. 47.5 for STR) reflects this difference in scored representations, not a re-evaluation of the same score. (iii) Candidate generation in Table 21 takes the top-50 of the same text–fused matrix, so C_1 equals the ITC R1. (iv) The final metric of Tables 1 and 21 rescores those 50 candidates with the ITM head and reads recall from the rescored matrix.

Table 21 separates the evaluation pipeline into fused candidate generation and ITM reranking. Across three seeds the two models' coverage (C_{50}) and conditional reranking success (S_{50}) agree within one seed SD on both datasets, while their embedding spaces respond very differently to the shared-direction removal (Table 14). These measurements characterize the stages of the observed pipelines; they do not isolate the causal contributions of fusion and reranking.

Table 22 records a mixed ITC direction at seed 0, two datasets rising and two falling. VATEX is negative on all three seeds (mean -3.6), so its direction is established; the other datasets are single-

Table 20: Text-to-video retrieval on DiDeMo and MSR-VTT with STR ported to four additional backbones without re-tuning (single run each). Each backbone follows its own repository's evaluation protocol and model-selection rule (R1 for AVIGATE, CLIP4Clip, and InternVideo2; R_{avg} for VALOR); this differs from Table 1 and the tables are not directly comparable. CLIP4Clip supports no audio and runs with K=2 (text, video).

	DiDeMo				**MSR-VTT**			
	R1		R_{avg}		R1		R_{avg}	
Backbone	Baseline	STR	Baseline	STR	Baseline	STR	Baseline	STR
VALOR	46.7	43.9	69.8	66.6	42.5	43.6	64.4	63.8
AVIGATE	38.1	38.2	59.3	59.5	44.5	44.8	67.0	67.3
CLIP4Clip	40.2	40.1	60.5	59.7	43.6	43.9	64.4	64.1
InternVideo2	23.4	24.1	50.1	50.2	42.0	42.6	64.9	64.8

Table 21: Pipeline-stage decomposition of text-to-video retrieval on MSR-VTT and DiDeMo over three seeds ($0, 1, 2$; mean and SD) for each model. C_k: percent of queries whose ground-truth video is among the fused-similarity top-k candidates; S_{50}: percent of those queries, at k=50, for which ITM reranking of the candidates places the ground truth first; ITM R1 $= C_{50}\, S_{50}/100$ per seed, recomputed from the dumped score matrices.

Dataset	**Model**	C_1	C_{10}	C_{50}	S_{50}	ITM R1
DiDeMo	Baseline	47.5 $\pm$1.10	83.2 $\pm$0.76	93.8 $\pm$0.58	70.6 $\pm$0.90	66.2 $\pm$0.93
	STR	46.1 $\pm$1.02	82.1 $\pm$0.45	93.3 $\pm$0.90	70.9 $\pm$0.48	66.1 $\pm$0.23
MSR-VTT	Baseline	47.2 $\pm$0.68	84.7 $\pm$0.46	95.3 $\pm$0.30	63.7 $\pm$0.49	60.7 $\pm$0.35
	STR	47.0 $\pm$0.50	85.3 $\pm$0.40	95.6 $\pm$0.50	63.8 $\pm$1.18	61.0 $\pm$1.06

seed observations. The mixed ITC picture alongside unchanged or improved ITM is consistent with the pipeline account of Sec. 4.4: the embedding-level differences are not reflected to the same extent in the final fused-and-reranked metric, which does not by itself identify the separate contributions of fusion and reranking.

G Standard Counterparts: Formulas, Weight Matching, and Retrieval

Counterpart formulas and provenance. For L_{vol}: (a) the pairwise cosine objective, which involves no determinant or rank structure,

$$L_{pair} = 1 - \frac{1}{|P|} \sum_{(a,b)\in P} \frac{1}{B} \sum_{i=1}^{B} s_i^{(a,b)}; \tag{8}$$

(b) the spectral objective, rank-based but with a different minimizer structure ($\det$ is the product of eigenvalues and vanishes when any one vanishes; the spectral form penalizes only the weakest direction),

$$L_{spec} = \frac{1}{B} \sum_{i=1}^{B} \left(-\log \lambda_{min}(G_i + \varepsilon I_K)\right). \tag{9}$$

The log-determinant is not a separate counterpart: for a fixed ridge it is a monotone transform of the per-instance $\sqrt{\det}$ (Sec. 3.3). For L_{div}: (c) the uniformity potential of Wang & Isola (2020) with t=2,

$$L_{unif} = \log \mathbb{E}_{i\neq j} \left[\exp\left(-t\|\tilde{z}_i^{(k)} - \tilde{z}_j^{(k)}\|^2\right)\right]; \tag{10}$$

Table 22: Dual-encoder (ITC) text-to-video retrieval for the baseline and STR, seed 0.

	Baseline			STR			
Dataset	R1	R5	R10	R1	R5	R10	ΔR1
DiDeMo	48.0	75.8	84.0	47.1	75.4	81.7	−0.9
MSR-VTT	47.0	76.8	84.6	47.5	76.2	84.9	+0.5
VATEX	54.2	84.5	92.7	50.8	82.5	91.8	−3.4
ActivityNet	48.0	78.9	89.2	49.7	80.8	90.7	+1.7

Table 23: Measured embedding-gradient magnitudes $\|\partial L/\partial z\|$ and matched weights, at initialization and at the trained full-configuration checkpoint (seed 0; 50 batches each). Trained-checkpoint values (bold) are used.

Variant	DiDeMo init	DiDeMo trained	MSR-VTT init	MSR-VTT trained
vol: $\sqrt{\det}$ (reference)	0.4274 / 0.100	0.3052 / **0.100**	0.4271 / 0.100	0.2885 / **0.100**
vol: pairwise cosine	0.2169 / 0.197	0.2663 / **0.115**	0.2159 / 0.198	0.2685 / **0.107**
vol: spectral	0.5730 / 0.075	1.1735 / **0.026**	0.5827 / 0.073	1.5021 / **0.019**
div: local Gaussian (reference)	0.0039 / 0.900	0.0193 / **0.900**	0.0084 / 0.900	0.0193 / **0.900**
div: uniformity	1.0297 / 0.003	0.8325 / **0.021**	0.9201 / 0.008	0.8093 / **0.022**
div: VICReg variance	0.0064 / 0.551	0.0063 / **2.783**	0.0063 / 1.187	0.0063 / **2.777**

(d) the VICReg variance hinge (Bardes et al., 2022) with γ=1 and its own constant ϵ=10^{-4}, where $Z \in \mathbb{R}^{B\times d}$ stacks one modality's embeddings and $\mathrm{Var}(z_j)$ is the batch variance of its j-th dimension,

$$v(Z) = \frac{1}{d}\sum_{j=1}^{d} \max\left(0,\, \gamma - \sqrt{\mathrm{Var}(z_j) + \epsilon}\right). \tag{11}$$

Both counterparts are applied per modality and averaged, matching the support of L_{div}; the implementations follow the official code of each method and agree with it to a numerical difference of 0.00. One caution for (d): on unit vectors the per-dimension standard deviation ($\approx 1/\sqrt{d} = 0.044$) keeps the hinge in its linear regime, so gradients flow but the loss cannot reach zero, unlike on the unnormalized projector outputs it was designed for.

Gradient matching. Weights are set by

$$\lambda_{variant} = \lambda_{canonical} \cdot \frac{\|\partial L_{canonical}/\partial z\|}{\|\partial L_{variant}/\partial z\|}, \tag{12}$$

measured over 50 batches in fp32 at the paper's fixed setting; Table 23 reports the measured magnitudes and matched weights. Unmatched, the uniformity gradient is 110 times that of L_{div} at initialization on MSR-VTT and 264 times on DiDeMo, falling to 42 and 43 times at the trained checkpoint, so without matching the comparison in Sec. 4.4 would measure strength rather than form. Loss-value matching is not suitable: the counterparts differ by constant offsets that contribute nothing to gradients. Matching at the trained (full-configuration) checkpoint is adopted because L_{div} is nearly silent at initialization (its gradient grows fivefold during training as L_{vol} pulls embeddings together), so no single weight fits both time points, and the trained-checkpoint weights are nearly identical across the two datasets while initialization-time weights differ by more than a factor of two (Table 23). A sensitivity run at the initialization-matched uniformity weight tests the matching-point choice itself. Matching equalizes the local gradient magnitude at one reference point, not the integrated pressure over training. In that sensitivity run (DiDeMo, uniformity at the initialization-matched weight), the video and audio embeddings collapse onto one direction ($\|\mu_v\|^2 = 0.9999$) while text stays dispersed, embedding retrieval drops to 29.6 R1, and ITM moves only 0.3 below the baseline, the most extreme instance of the insensitivity of Sec. 4.4.

Per-seed replication. Table 24 reports all four counterparts at seeds 0–2 with their selected checkpoints. The DiDeMo diversity counterparts vary across seeds at their held-out-selected checkpoints: for VICReg variance, common-mode energy is $0.007/0.006/0.124$ and embedding R1

Table 24: Per-seed counterpart diagnostics on MSR-VTT and DiDeMo (seeds 0, 1, 2). Each row uses the checkpoint selected by the held-out rule for that run, at the reported step; gradient-matched weights are fixed across seeds within each dataset. Emb. R1 is unmodified text–video embedding-cosine retrieval, Cent. $\Delta = \mathrm{R1(DN)} - \mathrm{R1(O)}$ measures deletion followed by residual rescaling, and ITM R1 uses the final fused-and-reranked score path (App. F). The combined removal is the only intervention reported; counterpart D-only and N-only decompositions are not included.

Variant	Seed	Step	$\|\mu_v\|^2$	Emb. R1	Cent. Δ	ITM R1
MSR-VTT						
vol → pairwise cos.	0	27,527	0.449	41.7	+1.0	61.1
	1	30,968	0.447	40.6	+0.9	58.9
	2	34,409	0.448	41.1	+0.9	62.1
vol → spectral	0	24,086	0.177	9.7	−6.2	60.4
	1	34,409	0.171	10.0	−7.3	60.2
	2	34,409	0.173	10.4	−7.3	61.3
div → uniformity	0	27,527	0.005	33.2	−0.5	61.3
	1	13,763	0.009	37.1	−0.2	58.7
	2	27,527	0.005	34.4	+0.0	60.7
div → VICReg var.	0	34,409	0.001	2.1	+0.4	61.4
	1	27,527	0.014	2.9	+0.7	60.7
	2	27,527	0.005	2.7	+0.4	61.4
DiDeMo						
vol → pairwise cos.	0	3,549	0.463	44.2	+2.7	66.1
	1	3,549	0.460	45.3	+2.4	66.3
	2	3,549	0.464	46.8	+1.9	66.1
vol → spectral	0	5,324	0.281	38.7	−2.1	64.7
	1	3,549	0.289	42.6	−3.8	66.5
	2	3,549	0.304	44.4	−2.5	66.0
div → uniformity	0	5,324	0.008	38.2	+0.9	66.2
	1	3,549	0.013	42.2	−0.1	66.5
	2	1,774	0.025	48.2	+0.4	67.4
div → VICReg var.	0	3,549	0.007	14.2	+1.9	66.7
	1	3,549	0.006	12.9	+2.6	65.2
	2	1,774	0.124	41.9	+3.9	66.2

Table 25: Both retrieval directions (ITM) for the baseline and the fixed configuration, mean ±SD over three seeds $(0, 1, 2)$: t→v retrieves videos from a text query (the direction of Table 1); v→t retrieves captions from a video query.

		t→v			v→t		
Dataset		R1	R5	R10	R1	R5	R10
DiDeMo	Baseline	66.2 ±0.92	86.0 ±0.67	90.1 ±0.58	64.7 ±0.12	86.6 ±0.44	90.9 ±0.32
	STR	66.1 ±0.23	86.3 ±0.60	90.1 ±0.36	63.8 ±0.76	86.5 ±1.29	90.8 ±0.49
MSR-VTT	Baseline	60.7 ±0.35	82.3 ±0.47	88.3 ±0.25	58.7 ±0.99	83.8 ±1.01	90.3 ±0.44
	STR	61.0 ±1.06	82.9 ±0.62	88.4 ±0.17	58.3 ±0.87	83.5 ±0.15	90.1 ±0.35
VATEX	Baseline	69.1 ±0.25	92.0 ±0.06	96.4 ±0.00	87.1 ±0.15	97.6 ±0.15	99.0 ±0.06
	STR	69.1 ±0.17	91.9 ±0.00	96.4 ±0.00	86.8 ±0.25	97.5 ±0.06	99.0 ±0.06
ActivityNet	Baseline	66.7 ±0.80	89.4 ±0.69	95.4 ±0.06	64.0 ±1.53	88.5 ±1.12	94.3 ±0.38
	STR	67.1 ±0.55	89.6 ±0.21	95.5 ±0.21	64.5 ±0.64	88.9 ±0.56	94.7 ±0.40

is $14.2/12.9/41.9$ at seeds 0/1/2, with selected steps 3,549/3,549/1,774; uniformity selects steps 5,324/3,549/1,774. The different seed-2 geometries coincide with earlier checkpoint selection, but these measurements do not isolate selection time from other seed-dependent variation. All selected checkpoints are retained under the unchanged selection rule.

Table 26: Text-to-video retrieval (ITM R1, with R_{avg} in parentheses) for the eight term subsets containing L_{comp}, on MSR-VTT at seed 0 with the gradient-matched weight $\lambda_{comp} = 0.167$. Together with the eight L_{comp}-free subsets of Table 9 these form the full 2^4 factorial. Column headers give the terms added to L_{comp} (+all = all three).

	L_{comp}	$+L_{div}$	$+L_{sep}$	$+L_{sep,div}$	$+L_{vol}$	$+L_{vol,div}$	$+L_{vol,sep}$	+all
MSR-VTT	60.6 (77.4)	60.4 (77.3)	60.4 (77.4)	61.2 (77.4)	60.6 (77.5)	62.1 (78.1)	60.1 (77.3)	61.3 (77.7)

H Video-to-Text Retrieval

Table 25 shows the two directions do not move together. Every t→v change is within one seed SD. In the v→t direction STR is lower on DiDeMo, MSR-VTT, and VATEX and higher on ActivityNet, and the DiDeMo and VATEX drops (-1.0 and -0.3 R1) are the only changes in either direction that exceed the larger of the two models' seed SDs. ActivityNet is the only dataset where both directions rise, and both of its changes are within seed SD, so no bidirectional improvement is established; the t→v direction of Table 1 should not be read as evidence about v→t.

I Intra-Modal Compactness: The Fourth Axis

Viewing the design space as inter/intra-modal crossed with similarity/diversity yields four combinations; the fourth, intra-modal similarity, was evaluated in an early variant and excluded. For modality k, with the normalized batch mean

$$\bar{z}^{(k)} = \frac{\frac{1}{B}\sum_{i=1}^{B} \tilde{z}_i^{(k)}}{\left\| \frac{1}{B}\sum_{i=1}^{B} \tilde{z}_i^{(k)} \right\|_2}, \tag{13}$$

the compactness loss is

$$L_{comp} = \frac{1}{K}\sum_{k=1}^{K}\left(1 - \frac{1}{B}\sum_{i=1}^{B} \tilde{z}_i^{(k)\top}\bar{z}^{(k)}\right), \tag{14}$$

which pulls embeddings within each modality toward a shared centroid. Its exclusion is consistent with compounding contraction: L_{vol} pulls the K modality embeddings of each instance toward a low-volume configuration and L_{comp} pulls different instances within each modality toward a shared centroid, and combined these pressures plausibly reduce the cross-instance variability needed for retrieval; in the common-mode terms of Sec. 4.3, L_{comp} pushes in the same direction as the shared-component inflation that L_{div} counteracts. The three retained terms are not an exhaustive taxonomy.

Table 26 completes the 2^4 factorial on MSR-VTT under the corrected implementation. Adding L_{comp} to a subset changes R1 by -0.5 to $+1.8$ depending on the partner terms, its average marginal effect on R1 over the sixteen cells is $+0.2$, and the best four-term subset (L_{vol}+L_{comp}+L_{div}, 62.1) exceeds the full three-term configuration (61.8) by 0.3, within the seed variability of Table 1. These are single-run observations at seed 0; the exclusion of L_{comp} follows the compounding-contraction argument above, with the four-term subsets performing at the level of the three-term ones.

J Cross-Pair Coupling for $K \geq 3$

A pairwise-cosine sum sees the off-diagonals of G_i only through their sum, so it cannot distinguish configurations with the same mean cosine but different joint structure: the triples $(0.9, 0.3, 0.3)$ and $(0.5, 0.5, 0.5)$ share mean cosine 0.5 yet differ in $\det(G_i)$ by a factor of 2.9 (0.172 vs. 0.500). The determinant instead couples the pairs, as the derivative below shows.

For K=3, the ridge-stabilized Gram matrix of the three unit embeddings, with s_{ab} the within-instance cosine between modalities a and b, has determinant

$$D = (1+\varepsilon)^3 - (1+\varepsilon)\left(s_{12}^2 + s_{13}^2 + s_{23}^2\right) + 2\, s_{12}\, s_{13}\, s_{23}, \tag{15}$$

whose third term couples every pair to the product of the other two. The derivative with respect to one pair is

$$\frac{\partial D}{\partial s_{12}} = 2\,[\,s_{13}\,s_{23} - (1+\varepsilon)\,s_{12}\,], \tag{16}$$

so the pressure on pair $(1,2)$ depends on the other two pairwise cosines through $s_{13}s_{23}$; when the other pairs are strongly aligned ($s_{13}s_{23} > (1+\varepsilon)s_{12}$) the sign flips. By contrast, for a pairwise-sum objective $L = -\sum_{a<b} s_{ab}$, the derivative with respect to any pair is the constant -1: the pressure on a pair is independent of the state of the others. For K=2, $D = (1+\varepsilon)^2 - s^2$ and $\partial D/\partial s = -2s$: the volume reduces to $|\sin\theta|$ (at ε=0), minimized at $|\cos\theta| \to 1$ (collinearity), with the positive-alignment direction supplied by the task loss, so the coupling term has no analogue.

This algebra does not by itself show practical superiority; the gradient-matched pairwise-cosine counterpart of Sec. 4.4 tests the forms empirically.

K Structural Analysis of the Regularization Terms

K.1 Active-Regime Agreement-Separation Interaction

L_{vol} and L_{sep} exert opposing pressures on within-instance cross-modal similarity. For K=2, a single pairwise cosine remains per instance, written s; for an instance with $s > \tau_q$, the regularization energy reduces to (derivation in the subsections below)

$$e_{reg}(s) = \lambda_{vol}\sqrt{(1+\varepsilon)^2 - s^2} + \lambda_{sep}(s-\tau_q)^2, \tag{17}$$

whose first term attracts toward higher similarity and whose second resists over-alignment, with relative strength governed by $\lambda_{vol}/\lambda_{sep}$. In the small-$\tau_q$ limit the interior stationary point is a local maximum, so the regularizer alone does not pin an operating point along this axis; away from that limit an interior local minimum can also appear (see below), and in either case the realized matched similarities reflect the balance with L_{task}. This predicts a negative gradient cosine between the two terms, not independent control; the prediction is confirmed at the parameter level in Sec. 4.6, and the trade-off motivates the λ_{vol} sweep in App. C.

K.2 Gradient Supports and the Three-Regime Frame

The gradients of L_{vol} and L_{sep} with respect to $\tilde{z}_i^{(k)}$ are linear combinations of the same instance's other-modality embeddings, while the gradient of L_{div} is a linear combination of the same modality's other-instance embeddings; for $d \gg K{+}B$ these supports are generically transverse. This concerns embedding-level supports only: shared encoders couple the terms at the parameter level, both levels are measured in Sec. 4.6, and neither level by itself establishes independence. The reason to measure these relations is that a pair of regularization terms falls into one of three regimes, opposing (a weight-ratio trade-off governs the pair), transverse (adjustable with little interference), or redundant (one term removable), and pairwise gradient cosines are a first-order diagnostic distinguishing the three at a given checkpoint. The measured assignment shapes tuning: the opposing pair is governed by one ratio, making a one-dimensional sweep informative (App. C).

The remaining subsections derive Eq. 17 for the two-modality case with ℓ_2-normalized embeddings, where the STR energy reduces to a scalar function of the inter-modal cosine similarity $s = \tilde{z}_i^{(1)\top}\tilde{z}_i^{(2)}$.

K.3 Reduction of L_{vol} to a Function of s

For instance i, the stacked embedding matrix is $Z_i = [\tilde{z}_i^{(1)}, \tilde{z}_i^{(2)}] \in \mathbb{R}^{d\times 2}$, with $s = \cos(\tilde{z}_i^{(1)}, \tilde{z}_i^{(2)})$ the within-instance cosine and Gram matrix

$$G_i = \begin{pmatrix} 1 & s \\ s & 1 \end{pmatrix}, \qquad \det(G_i) = 1 - s^2, \tag{18}$$

using ℓ_2-normalization for the unit diagonal. Note $\det(G_i) \geq 0$ for $s \in [-1,1]$, with equality when $s = \pm 1$ (collinear embeddings). Substituting into the stabilized volume objective (Eq. 2), whose per-instance summand is written $L_{vol}^{(i)}$, yields

$$L_{vol}^{(i)} = \sqrt{\det(G_i + \varepsilon I_2)} = \sqrt{(1+\varepsilon)^2 - s^2}, \tag{19}$$

where in exact arithmetic $\varepsilon > 0$ keeps the argument positive and the gradient defined at $s = \pm 1$; the fp32 behavior is reported in App. L.1.

K.4 Reduction of L_{sep} and the Role of L_{div}

For $K{=}2$ there is a single modality pair, and Eq. 3 reduces to $L_{sep}^{(i)} = [\max(0,\ s - \tau_q)]^2$ per instance, active only when $s > \tau_q$. L_{div} involves pairs of different instances within a modality and does not depend on the within-instance similarity s, so it does not appear in the per-instance energy.

K.5 Per-Instance Energy and Its Stationary Point

For $K{=}2$, a single instance, and the active hinge regime $s > \tau_q$, the energy is Eq. 17, with derivative (taken at fixed τ_q, matching the implementation, which detaches the batch threshold and recomputes it each step; App. L.1)

$$\frac{de_{reg}}{ds} = -\frac{\lambda_{vol}\, s}{\sqrt{(1+\varepsilon)^2 - s^2}} + 2\lambda_{sep}(s - \tau_q). \tag{20}$$

Setting the derivative to zero gives the stationary condition

$$\frac{\lambda_{vol}\, s}{\sqrt{(1+\varepsilon)^2 - s^2}} = 2\lambda_{sep}(s - \tau_q), \tag{21}$$

whose left-hand side is the marginal alignment pressure from L_{vol} (approaching $\lambda_{vol}/\sqrt{2\varepsilon + \varepsilon^2}$ as $s \to 1$) and whose right-hand side is the marginal separation pressure, linear in s for fixed τ_q.

For qualitative interpretation, taking $\varepsilon \to 0$ and the limit $\tau_q \to 0$ and dividing by $s \neq 0$ gives $\lambda_{vol}/\sqrt{1 - s^2} = 2\lambda_{sep}$, which admits an interior solution only when $\lambda_{vol} < 2\lambda_{sep}$; in this limit the second derivative shows the stationary point is a local maximum of e_{reg}, so the regularizer-only energy is driven by the volume term as $|s| \to 1$. This conclusion is specific to the limit: for a general threshold an interior local minimum can coexist with the maximum (at $\lambda_{vol}{=}\lambda_{sep}{=}0.1$, $\tau_q{=}0.2$: a minimum near $s{=}0.457$ and a maximum near $s{=}0.722$), so the scalar reduction does not capture the full batch-quantile dynamics. The realized operating point under joint optimization is therefore set by the balance with L_{task}, not by the regularizer alone. The pressure-ratio role of $\lambda_{vol}/\lambda_{sep}$ remains informative and underlies the sweep of App. C. The threshold τ_q controls when separation activates: for $s \leq \tau_q$ the per-instance energy is governed by L_{vol} alone.

L Additional Formulations

L.1 Numerical and Computational Details

Stable computation of L_{vol}. Direct evaluation of $\sqrt{\det(G_i + \varepsilon I_K)}$ can be numerically unstable when modality embeddings approach linear dependence. The determinant is therefore evaluated in the log domain via Cholesky factorization: since in exact arithmetic $G_i + \varepsilon I_K$ is positive definite for any $\varepsilon > 0$, the upper-triangular Cholesky factor R_i satisfies $G_i + \varepsilon I_K = R_i^\top R_i$, giving

$$\log\det(G_i + \varepsilon I_K) = 2\sum_{j=1}^{K} \log(R_{i,jj}), \quad \sqrt{\det(G_i + \varepsilon I_K)} = \exp\big(\tfrac{1}{2}\log\det(G_i + \varepsilon I_K)\big). \tag{22}$$

The Gram path is computed in fp32 with automatic mixed precision disabled; samples whose Cholesky factorization fails fall back to a sign-log-determinant routine. The numerical checks cover the 92 STR-instrumented finetuning runs in the results archive – the main STR runs at seeds 0–2 on all four datasets, the term-subset, L_{comp}, $K{=}4$, and sweep runs, and all counterpart runs – whose diagnostics are logged every 100 optimizer steps: the fallback was invoked in none of the 21,567 logged steps. The logged similarity statistic is the batch mean of within-instance cosines pooled over the three modality pairs on training batches; its maximum over all logged steps is 0.896. This is a different statistic from the matched similarity of Tables 11 and 12, which is the text–video pair mean on the evaluation split at the selected checkpoint and reaches up to 0.948. An fp64 reference that rebuilds the Gram matrix in fp64 through the same routine was compared on controlled batches spanning

pooled-mean cosines 0.50 to 1.0, with relative errors $|\Delta L|/|L|$ for the loss and $\|\Delta g\|/\|g\|$ for the gradient: both stay below 10^{-6} through pooled mean 0.94, covering the logged training range, grow to about 2×10^{-5} at 0.997, and break down only when the cosine reaches 1 at fp32 precision, where the fallback yields non-finite gradients. The nominal ridge $\varepsilon{=}10^{-8}$ lies below the fp32 spacing at 1.0 (1.2×10^{-7}), so the effective diagonal increment in fp32 is zero. At the logged diagnostic points the fp32 Cholesky path completed without invoking the fallback despite this zero increment – an empirical observation about the recorded matrices, not a positive-definiteness guarantee provided by the fallback or the ineffective ridge; the nominal ridge takes effect in the fp64 reference. These checks characterize the logged runs and the controlled comparison; they do not establish numerical behavior outside this scope.

Computational cost of L_{div}. Eq. 5 sums a soft repulsion over all instance pairs within each modality, giving $O(KB^2)$ pair interactions per step. All pairwise distances are computed in vectorized form per modality; self-pairs ($i{=}j$) are excluded by masking the diagonal, implementing the indicator of Eq. 4 (the $1/B^2$ normalization is retained for notational simplicity, and the resulting constant scaling is absorbed into λ_{div}). For the configurations used here ($B{=}16$ per GPU, $K{=}3$), the cost is negligible relative to the encoder passes.

Batch support and pooling. τ_q in Eq. 3 is a single quantile pooled over all modality pairs and all instances of the global batch, gathered across GPUs before the quantile is taken, with the cap $\tau = \min(\tau_q, 1.0)$; it is not computed per pair. STR terms see the gathered global batch (32) for the quantile and the per-GPU batch (16) for the L_{div} distances. The gathered cosines used for τ_q are detached, so τ_q carries no gradient and none crosses GPUs; the gradient of L_{sep} flows only through the local hinge, which is what the fixed-threshold derivative of App. K describes. Missing modalities are handled by per-sample presence masks: L_{vol} drops instances with any missing modality from its average (reducing K per instance would mix Gram scales), L_{div} removes missing rows before the distance matrix, so they are absent from the normalizing denominator, and L_{sep} excludes a pair whose either side is missing from both the hinge and the quantile pool. The four backbone experiments of Sec. 4.7 use one shared implementation of the three terms.

L.2 Geometric Diagnostic Metrics

Let $\{(\tilde{z}_i^{(t)}, \tilde{z}_i^{(v)})\}_{i=1}^{B}$ denote a mini-batch of ℓ_2-normalized text-video embedding pairs and $S_{ij} = \tilde{z}_i^{(t)\top} \tilde{z}_j^{(v)}$ the similarity matrix; matched pairs lie on its diagonal. Metrics defined on a mini-batch are averaged across batches sampled from the evaluation split.

Matched Similarity. $\mathrm{MatchedSim} = \frac{1}{B} \sum_i S_{ii}$.

Common-Mode Energy. $\|\mu_k\|^2$ with $\mu_k = \frac{1}{n} \sum_i \tilde{z}_i^{(k)}$, the squared norm of the mean of the ℓ_2-normalized embeddings of modality k over the n instances of the full evaluation split. Because the embeddings are unit vectors, it ranges from 0 (zero mean vector) to 1 (all embeddings collapsed onto one direction); a zero mean does not imply the absence of shared low-dimensional structure. Sec. 4.3 reports the video modality, $\|\mu_v\|^2$.

Maximum Mean Discrepancy (MMD). Squared MMD between the text and video samples under an RBF kernel $\kappa(a, b) = \exp(-\gamma \|a - b\|_2^2)$ with $\gamma{=}1$:

$$\mathrm{MMD}^2 = \frac{1}{B^2} \sum_{i,j} \kappa(\tilde{z}_i^{(t)}, \tilde{z}_j^{(t)}) + \frac{1}{B^2} \sum_{i,j} \kappa(\tilde{z}_i^{(v)}, \tilde{z}_j^{(v)}) - \frac{2}{B^2} \sum_{i,j} \kappa(\tilde{z}_i^{(t)}, \tilde{z}_j^{(v)}). \tag{23}$$

Symmetric Nearest-Neighbor Distance. The symmetric nearest-neighbor distance, averaging nearest-neighbor distances in both directions (not an optimal-transport Wasserstein distance):

$$\mathrm{Sym.NN} = \tfrac{1}{2} \left(\tfrac{1}{B} \sum_i \min_j \|\tilde{z}_i^{(t)} - \tilde{z}_j^{(v)}\|_2 + \tfrac{1}{B} \sum_j \min_i \|\tilde{z}_j^{(v)} - \tilde{z}_i^{(t)}\|_2 \right). \tag{24}$$

Intra-modal Variance. $\mathrm{IntraVar} = \frac{1}{B}\sum_i \|\tilde{z}_i^{(t)} - \mu_t\|_2^2$ with the batch mean $\mu_t = \frac{1}{B}\sum_i \tilde{z}_i^{(t)}$ (not normalized).

Pos-Neg Similarity Margin. $\mathrm{Margin} = \frac{1}{B}\sum_i S_{ii} - \frac{1}{B(B-1)}\sum_{i\neq j} S_{ij}$. This mean margin compares the positive against the average negative and is not the quantity that decides rank 1. That quantity is the per-query hard-negative margin $h_i = S_{ii} - \max_{j\neq i} S_{ij}$, reported as the mean $\mathrm{HM} = \frac{1}{Q}\sum_i h_i$ over the Q queries: with a single ground truth per query, $h_i > 0$ makes query i's top-1 correct (ties are resolved in favor of the ground truth by the rank rule), while the sign of the mean alone decides nothing. In Table 15 both margins are computed per query over the full evaluation gallery rather than a mini-batch.

L.3 Lennard-Jones Diversity: An Unstable Alternative

Early variants of L_{div} used a Lennard-Jones interaction, $\phi_{LJ}(d) = (r_{min}/d)^{12} - 2(r_{min}/d)^6$ with $d \leftarrow \max(d, \epsilon_d)$ for a small floor ϵ_d. Its high-order repulsion yields near-unbounded gradients as embeddings concentrate, which empirically caused unstable optimization and occasional loss explosions, and its attractive regime is unnecessary here. The soft-repulsion form of Eq. 4 keeps the separation effect with bounded gradients.

L.4 Image-Text Contrastive Loss (ITC)

Given a mini-batch of paired normalized embeddings $\{(\tilde{z}_i^{(t)}, \tilde{z}_i^{(k)})\}_{i=1}^B$, ITC is a symmetric InfoNCE objective:

$$L_{ITC} = \frac{1}{2}\left(L_{t\to k} + L_{k\to t}\right), \tag{25}$$

with

$$L_{t\to k} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\tilde{z}_i^{(t)\top}\tilde{z}_i^{(k)}/\tau\right)}{\sum_{j=1}^{B}\exp\left(\tilde{z}_i^{(t)\top}\tilde{z}_j^{(k)}/\tau\right)}, \quad L_{k\to t} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\tilde{z}_i^{(k)\top}\tilde{z}_i^{(t)}/\tau\right)}{\sum_{j=1}^{B}\exp\left(\tilde{z}_i^{(k)\top}\tilde{z}_j^{(t)}/\tau\right)}, \tag{26}$$

where τ is a temperature parameter; L_{ITC} is averaged over modality combinations when multiple pairs are used.

L.5 Image-Text Matching Loss (ITM)

ITM is a binary classification distinguishing matched from mismatched pairs. Given a candidate text-modality pair (t, k), a multimodal encoder produces a scalar matching logit $s_{t,k} \in \mathbb{R}$. With $\mathcal{P}$ the positive and $\mathcal{N}$ the negative pairs sampled from a mini-batch,

$$L_{ITM} = \frac{1}{|\mathcal{P}| + |\mathcal{N}|}\left(\sum_{(t,k)\in\mathcal{P}} \ell_{CE}(s_{t,k}, 1) + \sum_{(t,k)\in\mathcal{N}} \ell_{CE}(s_{t,k}, 0)\right), \tag{27}$$

where ℓ_{CE} is the binary cross-entropy (CE) loss, $\ell_{CE}(s, y) = -y\log\sigma(s) - (1-y)\log(1-\sigma(s))$ with $\sigma(\cdot)$ the sigmoid. ITM complements ITC by enforcing pair-level cross-modal consistency beyond global embedding similarity.

M Limitations

Performance claims. Under three-seed means the fixed configuration ties the baseline on VATEX, sits 0.1 R1 below it on DiDeMo, and exceeds it only on MSR-VTT and ActivityNet (+0.3 and +0.4, both within one seed SD), so these are descriptive changes under the fixed protocol, not evidence of uniform improvement or worst-case robustness, and no per-dataset significance is claimed. The spread among STR, GRAM, and TRIANGLE is below the largest of their seed SDs on VATEX and DiDeMo (0.0 and 0.6 R1) but not on ActivityNet and MSR-VTT (0.7 and 1.2 R1). Higher-rank recall can move against R1: on the additional backbones (single runs) R1 rises on all four MSR-VTT

backbones while R_{avg} improves on only one. On DiDeMo, STR lowers R1 on two of four backbones, one by 2.8. The main retrieval results are text-to-video; in the video-to-text direction STR is lower than the baseline on three of four datasets, and the DiDeMo and VATEX drops exceed the seed SDs there (Table 25). Reduced sensitivity to DN does not imply superior original embedding retrieval: STR has lower unmodified embedding-cosine R1 than the baseline on all four datasets in the reported three-seed means (Table 14).

Configuration. The fixed three-term configuration is a shared reference, not an optimum: per-dataset test-oracle subsets are better by up to 0.5 R1, and no dataset-adaptive selection rule is established. L_{sep} exerts little effective pressure in this setting, and three-term non-redundancy is not claimed; the internal constants (q, r_{min}, σ) were fixed by limited sweeps and are not shown optimal.

Mechanism claims. The common-mode analysis identifies a property of the embedding geometry by post-hoc intervention; it is a major mediator of margin contraction rather than a proven training-time causal pathway, and no derivation is given for why volume pressure inflates the shared direction. The intervention is diagnostic only: removal is not proposed as a method, and STR-side changes under removal are not claimed as recovery. Mean-margin recovery substantially overstates ranking change. Gradient cosines are checkpoint-conditional first-order statistics without cross-seed stability claims, and the mechanism behind the embedding-level near-orthogonality of (L_{vol}, L_{sep}) is unverified.

Scope. All evidence concerns finetuning for video-text retrieval with contrastive-plus-matching training; transfer to other tasks, architectures, or training regimes is not shown. The factorial analysis remains single-run at seed 0 on four datasets. Counterpart common-mode energy, unmodified embedding R1, combined-removal responses (DN minus O), and final ITM R1 are reported over seeds 0–2 on MSR-VTT and DiDeMo, with per-seed values and selected steps in App. G; this replication does not include a full D-only/N-only decomposition. The main comparison, the baseline–STR removal decomposition, and both retrieval directions are three-seed on four datasets; the pipeline-stage measurements are three-seed on two datasets. The ITM insensitivity finding implies the reported final metric hides embedding-space differences that embedding-deploying systems would inherit. Mitigation, suppressing shared-direction inflation while retaining alignment, is not attempted here. Backbone evidence is limited to single runs on two datasets with repository-defined selection rules; different components of the shared-direction diagnosis reproduce on different backbones and none reproduces the full VAST pattern (App. E), the proposed explanations for this heterogeneity (encoder update magnitude, the separation between regularized and ranking spaces) are untested, and the learning-rate probe did not reach a trainable regime under the tested schedule, so the diagnosis's generality beyond the tested settings is not established.