Rebalancing Contrastive Alignment with Bottlenecked Semantic Increments in Text-Video Retrieval

Jian Xiao[1], Zijie Song[2], Jialong Hu[1], Hao Cheng[1], Jia Li[1]*, Zhenzhen Hu[1]*, Richang Hong[1]

[1]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
[2]School of Big Data and Statistics, Anhui University, Hefei, China

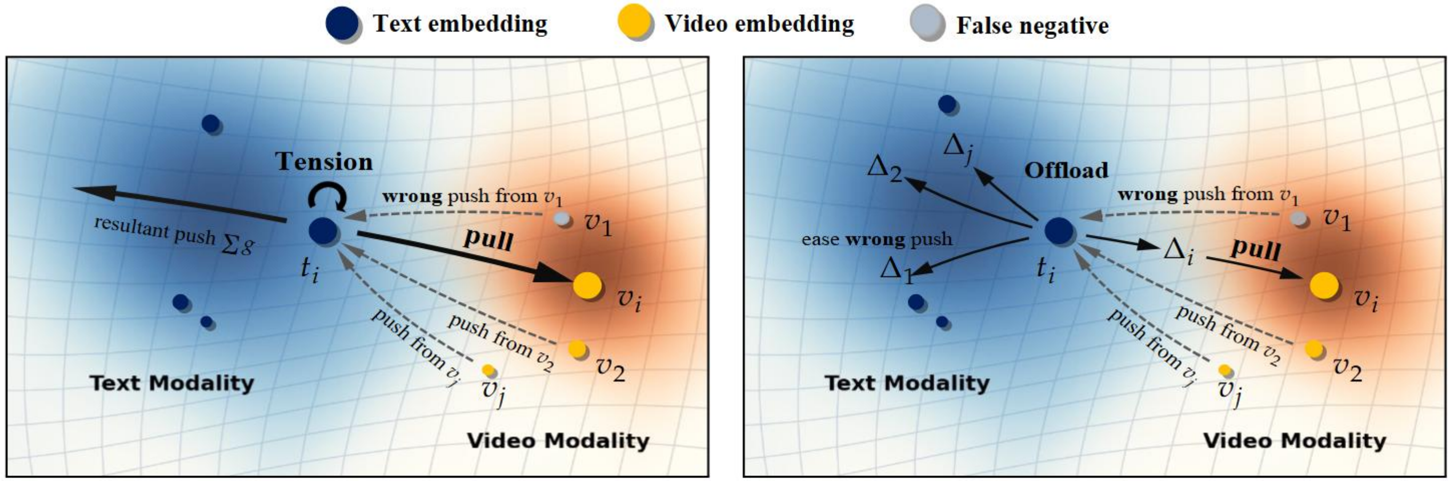
{j.xiao_hfut, chenghao}@mail.hfut.edu.cn, zjsong@ahu.edu.cn
zdszds534@gmail.com, {lijia, zzhu}@hfut.edu.cn, hongrc.hfut@gmail.com

Outline

Content

- Motivation
- Contribution
- Method
- Experiment
- Qualitative Analysis

Motivation



(a) (b)

- Text-video retrieval aims to find relevant videos given a text query. Current contrastive models (e.g., CLIP) face two major issues (see Fig. (a)): **1)** Optimization tension: caused by the modality gap, where gradients from positives and negatives cancel out, leaving the anchor nearly unchanged. **2)** Hard negative noise: semantically similar negatives push the anchor in the wrong direction. These issues limit the upper bound of the modal alignment capability.
- We redistribute gradients by introducing a pair-specific increment Δ_{ij} that linearly perturbs each text anchor t_i. This also offloads noisy gradients Δ_{ij}, stabilizing t_i's semantics (see Fig. (b)).
- Treating InfoNCE loss $\mathcal{L}_i$ for t_i as a multivariate function over $\{\Delta_{ij}\}_{j=1}^{B}$, we derive the gradient update of Δ_{ij} via a multivariate first-order Taylar Expansion under a ℓ_2 trust region constraint and interpret it as an *Information Bottleneck* to prevent trivial solutions.

Contribution

- **1)** We analyze the gradient structure of InfoNCE and reveal its inherent multi-variable coupling by introducing pairwise increments Δ_{ij}. A multivariate first-order Taylor expansion within a trust region yields a update rule for each Δ_{ij} consistent with the InfoNCE descent direction.
- **2)** We propose a Gap-Aware Retrieval (**GARE**) framework, where a learnable network ψ predicts pair-specific increments Δ_{ij} and integrates them into the forward pass to offload optimization tension while mitigating noise from false negatives. We also introduce a **relaxed** *Variational Information Bottleneck* (**VIB**) objective that regularizes Δ_{ij}, balancing informativeness and compression.
- **3)** Experiments on four text–video retrieval benchmarks, i.e., MSR-VTT, DiDeMo, ActivityNet Captions and MSVD, showing consistent improvements, and further analyses confirm that the learned increments Δ_{ij} are semantically meaningful and geometrically structured.

Method

Observation

- For batch size B, the gradient of $\mathcal{L}_i$ on an anchor t_i is the sum of B pairwise gradients:

$$\nabla_{t_i}\mathcal{L}_i = \frac{1}{\tau}\sum_{j}^{B}(p_{ij} - y_{ij}) \cdot \left(\frac{v_j}{|t_i|_2 |v_j|_2} - \cos(t_i,\, v_j) \cdot \frac{t_i}{|t_i|_2^2}\right), \quad \mathcal{L}_i = -\log\frac{e^{\cos(t_i, v_i)/\tau}}{\sum_j^B e^{\cos(t_i, v_j)/\tau}}$$

where $p_{ij} = \frac{e^{\cos(t_i, v_i)/\tau}}{\sum_j^B e^{\cos(t_i, v_j)/\tau}}$ and $y_{ij} \in \{0,1\}$ is match label.

- Empirical results on 512 dimensions show severe cancellation among them.
  - Gradients from most negative pairs $(t_i,\ v_j)$: magnitude ≈ 40 to 60 (bottom of right-side figure).
  - Adding the positive pair $(t_i,\ v_j)$ shrinks it to 2 to 4 (top of right-side figure).
  - $\longrightarrow$ The anchor t_i barely moves during training.

Problem

- t_i stays trapped in a narrow optimization region.
- The modality gap constrains updates and causes in-place optimization.

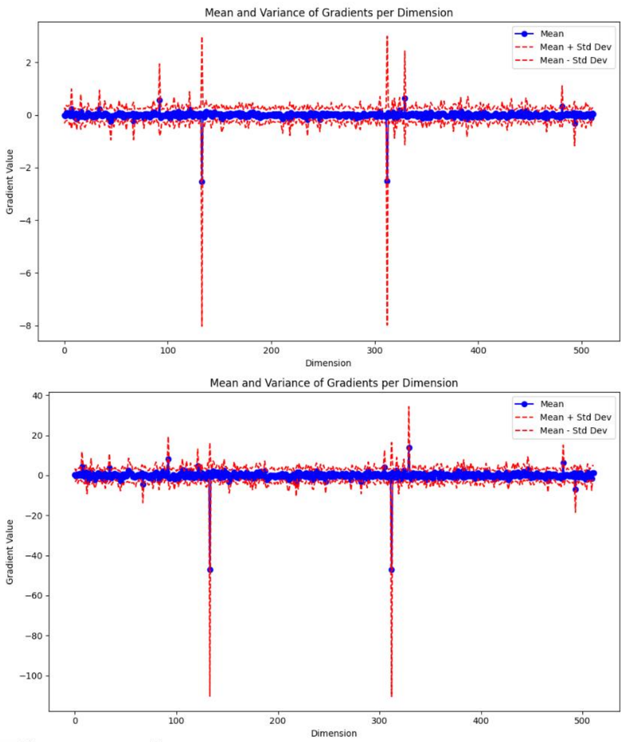


Figure 2: Mean and variance of summed gradient (top) and negative gradients (bottom) across 512 dimensions, showing collinear but opposite forces that largely cancel out.

Method

- **Idea**
  - To relax optimization tension, introduce a pair-specific increment Δ_{ij} for each pair $(t_i,\ v_j)$.
  - Replace the anchor by a linearly perturbed representation: $t_{\Delta_{ij}} = t_i + \Delta_{ij}$, this results a multivariate InfoNCE $\mathcal{L}_i$:

$$\mathcal{L}_i(\Delta_{i1},\ \Delta_{i2},\ \ldots,\ \Delta_{iB}) = -\log \frac{\exp\left(s_{ii}/\tau\right)}{\sum_j^B \exp\left(s_{ij}/\tau\right)},\quad s_{ij} = \cos\left(t_i + \Delta_{ij},\ v_j\right)$$

- **Effects**
  - 1) Gradient redistribution — redirects gradients from t_i to Δ_{ij}.
    - Each Δ_{ij} only receives gradient from its own pair $(t_i,\ v_j)$, where the gradients are

$$\nabla_{t_{\Delta_{ij}}} \mathcal{L}_i(\Delta_{i*}) = \frac{1}{\tau} \sum_j^B \left(p_{ij} - y_{ij}\right) \cdot \left(\frac{v_j}{\left|t_i + \Delta_{ij}\right|_2 \left|v_j\right|_2} - \cos\left(t_i + \Delta_{ij},\ v_j\right) \cdot \frac{t_i + \Delta_{ij}}{\left|t_i + \Delta_{ij}\right|_2^2} \right)$$

$$\nabla_{\Delta_{ij}} \mathcal{L}_i(\Delta_{i*}) = \nabla_{t_{\Delta_{ij}}} \mathcal{L}_i(\Delta_{i*}),\ \ \nabla_{t_i} \mathcal{L}_i(\Delta_{i*}) = \sum_j^B \nabla_{t_{\Delta_{ij}}} \mathcal{L}_i(\Delta_{i*})\,.$$

    - Collectively, $\{\Delta_{ij}\}_j^B$ enlarge the **effective optimization region** of t_i.
  - 2) Δ_{ij} absorbs noisy gradients from hard negatives, reducing semantic interference

Method

- **Multivariate Taylor Expansion**
  - Gradient of $\boldsymbol{\mathcal{L}}_i$ w.r.t. one Δ_{ij} depends on other non-zero Δ_{ik} → capturing inter-pair coupling.
  - Expanding at $\Delta_{ik} = 0$ would break the relative ranking prior among pairs. This results to a multivariate first-order Taylor Expansion:

$$\boldsymbol{\mathcal{L}}_i(\Delta_{i*}) \approx \boldsymbol{\mathcal{L}}_i\left(\Delta_{i*}^{(t)}\right) + \sum_{j}^{B}\left[\nabla_{\Delta_{ij}}\boldsymbol{\mathcal{L}}_i\left(\Delta_{i*}^{(t)}\right)\right]^{\top}\left(\Delta_{ij} - \Delta_{ij}^{(t)}\right)$$

- **ℓ_2 Trust-Region Constraint $\left|\Delta_{ij}\right|_2 \leq \varepsilon_{ij}$ to limit perturbation magnitude.**

- **Derived Iterative Update with Initial Non-Zero State $\Delta_{i*}^{(t)}$** (by steepest descent + Cauchy–Schwarz):

$$\Delta_{ij}^{(t+1)} = \Delta_{ij}^{(t)} - \alpha_{ij}^{(t)} \cdot \frac{\nabla_{\Delta_{ij}}\boldsymbol{\mathcal{L}}_i\left(\Delta_{i*}^{(t)}\right)}{\left|\nabla_{\Delta_{ij}}\boldsymbol{\mathcal{L}}_i\left(\Delta_{i*}^{(t)}\right)\right|_2}, \quad \text{where } \alpha_{ij}^{(t)} \text{ analytically ensures } \left|\Delta_{ij}^{(t+1)}\right|_2 \leq \varepsilon_{ij}.$$

- **Implementation**
  - Each iteration initializes $\Delta_{i*}^{(t)}$ from a neural module $\psi\left(t_i - v_j, \mathbf{V}; \mathbf{\Theta}^{(t)}\right) = q_\psi(\Delta_{ij}^{(t)} | t_i, v_j)$ after **CLIP Encoder**.
  - Back-propagation naturally satisfies this update rule with different learning rate η from optimizer.

Method

- ***Variation Information Bottleneck* Regularization for Δ_{ij}**
  - Δ_{ij} only receives gradients from its own pair $(t_i,\ v_j)$, lacking contrastive interaction with other pairs.
    - $\rightarrow$ Direct optimization easily leads to **trivial or collapsed** Δ_{ij}.
  - Treat Δ_{ij} as an *information bottleneck variable* that captures only essential **alignment information** between t_i and v_j. This results a *Variation Information Bottleneck* objective:

$$\mathcal{L}_{\mathrm{VIB}} \coloneqq \underbrace{-\ \mathbb{E}_{(t,v,y)}\mathbb{E}_{\Delta\sim q_\psi(\Delta|t,v)}[\log q_\theta(y|\Delta)]}_{\text{multivariate InfoNCE loss}} + \beta\cdot\underbrace{\mathbb{E}_{(t,v)}\left[\mathrm{KL}\left(q_\psi(\Delta|t,v)||\mathcal{N}(0,\ \boldsymbol{I})\right)\right]}_{\text{compression term } \mathcal{L}_{\mathrm{IB}}}.$$

  - The $\psi(\cdot)$ serves as a *deterministic posterior*, each Δ_{ij} is viewed as a **Dirac delta** centered at a fixed value.
    - Since the Dirac posterior is *singular* w.r.t. the Gaussian prior $\mathcal{N}(0,\ \boldsymbol{I})$, we relax the compression term $\mathcal{L}_{\mathrm{IB}}$ **on the text side**, leveraging the *one-to-many* nature of video–text pairs.
    - $\rightarrow$ This overly penalizes video-side information and circumvents the singularity between the deterministic and stochastic distributions. By the convexity of $\mathrm{KL}\left(\ \cdot\ ||\ \mathcal{N}(0,\boldsymbol{I})\right)$ and Jensen's inequality, this yields a relaxation:

$$\begin{aligned}\mathbb{E}_{(t,v)}\left[\mathrm{KL}\left(q_\psi(\Delta|t,v)||\mathcal{N}(0,\ \boldsymbol{I})\right)\right] &= \mathbb{E}_v\mathbb{E}_{t|v}\left[\mathrm{KL}\left(q_\psi(\Delta|t,v)||\mathcal{N}(0,\ \boldsymbol{I})\right)\right]\\ &\geq \mathbb{E}_v\left[\mathrm{KL}\left(\overline{q_\psi}(\Delta|v)||\mathcal{N}(0,\ \boldsymbol{I})\right)\right].\end{aligned}$$

Method

- **Extra Regularization: Radii Prior & Direction Diversity**
  - **Motivation**
    Δ_{ij} from $\psi(\cdot)$ often lie on the trust-region boundary. We regularize them to (1) enlarge their **magnitude diversity**, and (2) increase **directional variety** across pairs.
  - **Trust-Region Radii Prior**
    - Encourage heterogeneous radii for each anchor t_i :

$$\mathcal{L}_{\varepsilon} = -\max\left(\mathbb{E}_t\left[\mathrm{Var}\left(\{\varepsilon_{ij}\}_j^B\right)\right],\ \lambda\right),\quad \lambda > 0$$

    - Larger variance → richer optimization radii.
    - Prevents all Δ_{ij} collapsing to similar magnitudes.
  - **Direction Diversity**
    - Promote angular diversity among normalized increments:

$$\mathcal{L}_{\mathrm{dir}} = \mathbb{E}_t\left[\log \mathbb{E}_{j,k}\left[\exp\left(-\alpha \cdot \left(1 - \langle z_{ij}, z_{ik}\rangle\right)\right)\right]\right],\qquad z_{ij} = \frac{\Delta_{ij}}{\left|\Delta_{ij}\right|_2}$$

    - Reduces directional redundancy.
    - Expands geometric coverage of Δ_{ij} around t_i.

Experiment

Table 1: Comparison results on MSR-VTT dataset on Text-to-Video Retrieval and Video-to-Text Retrieval. DiCoSA [24] utilizes QB-Norm [6] for inference and is grayed out for a fair comparison. Note that T2VLA [45] is a non-CLIP method.

Methods	Text-to-Video Retrieval					Video-to-Text Retrieval				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
T2VLA [45] CVPR21	29.5	59.0	70.1	4.0	-	31.8	60.0	71.1	3.0	-
CLIP4Clip [33] Neurocomputing22	44.5	71.4	81.6	**2.0**	15.3	42.7	70.9	80.6	**2.0**	11.6
X-Pool [17] CVPR22	46.9	72.8	82.2	**2.0**	14.3	44.4	73.3	84.0	**2.0**	9.0
TS2-Net [32] ECCV22	47.0	74.5	83.8	**2.0**	13.0	45.3	74.1	83.7	**2.0**	9.2
EMCL-Net [22] NeurIPS22	46.8	73.1	83.1	**2.0**	12.8	46.5	73.5	83.5	**2.0**	8.8
UATVR [16] ICCV23	47.5	73.9	83.5	**2.0**	12.3	46.9	73.8	83.8	**2.0**	8.6
DiCoSA [24] IJCAI23	47.5	74.7	83.8	2.0	13.2	46.7	75.2	84.3	2.0	8.9
ProST [29] ICCV23	48.2	74.6	83.4	**2.0**	12.4	46.3	74.2	83.2	**2.0**	8.7
HBI [23] CVPR23	48.6	74.6	83.4	**2.0**	**12.0**	46.8	74.3	84.3	**2.0**	8.9
DiffusionRet [25] ICCV23	49.0	**75.2**	82.7	**2.0**	12.1	47.7	73.8	84.5	**2.0**	8.8
EERCF [38] AAAI24	47.8	74.1	**84.1**	-	-	44.7	74.2	83.9	-	-
MPT [54] ACM MM24	48.3	72.0	81.7	-	14.9	46.5	74.1	82.6	-	11.8
Baseline	46.6	73.4	82.2	**2.0**	12.6	45.6	73.4	82.4	**2.0**	9.6
GARE (Ours)	**49.1**	74.7	83.6	**2.0**	**12.0**	**48.6**	**75.3**	**85.3**	**2.0**	**8.5**

Table 2: Comparison results on DiDeMo, ActivityNet Captions, and MSVD datasets on Text-to-Video Retrieval. Note that FROZEN [3] is a non-CLIP method.

DiDeMo				
Methods	R@1	R@5	R@10	MnR
TS2-Net	41.8	71.6	82.0	14.8
CLIP4Clip	42.8	68.5	79.2	18.9
DiCoSA	45.7	74.6	83.5	**11.7**
DiffusionRet	46.7	74.7	82.7	14.3
HBI	46.9	74.9	82.7	12.1
Baseline	45.4	74.3	82.0	12.3
GARE (Ours)	**47.6**	**75.4**	**83.1**	**12.0**

ActivityNet Captions				
Methods	R@1	R@5	R@10	MnR
CLIP4Clip	40.5	72.4	83.6	7.5
TS2-Net	41.0	**73.6**	84.5	8.4
DiCoSA	42.1	73.6	84.6	6.8
MPT	41.4	70.9	82.9	7.8
HBI	42.2	73.0	84.6	**6.6**
Baseline	40.2	72.5	83.6	7.5
GARE (Ours)	**42.6**	73.2	**84.8**	**6.6**

MSVD				
Methods	R@1	R@5	R@10	MnR
FROZEN [3]	33.7	64.7	76.3	-
CLIP4Clip	45.2	75.5	84.3	**10.3**
EMCL-Net	42.1	71.3	81.1	17.6
UATVR	46.0	**76.3**	**85.1**	10.4
Diffusion	**46.6**	75.9	84.1	15.7
Baseline	45.0	75.5	84.5	10.7
GARE (Ours)	46.4	76.1	84.5	10.6

Experiment

Table 3: Ablation on losses combination on Text-to-Video Retrieval results on MSR-VTT 1k-A. First row denotes the baseline.

Δ	$\mathcal{L}_{\text{IB}}$	$\mathcal{L}_{\varepsilon}$	$\mathcal{L}_{\text{dir}}$	R@1↑	R@5↑	R@10↑	MnR↓
Baseline				46.6	73.4	82.2	12.6
✓				47.4	73.8	82.8	12.4
✓		✓		47.2	73.3	82.2	12.4
✓			✓	47.0	73.1	82.3	12.6
✓		✓	✓	47.4	73.7	82.8	12.3
✓	✓			48.3	74.2	83.2	12.4
✓	✓	✓	✓	**49.1**	**74.7**	**83.6**	**12.0**

Table 4: Ablation on Context Modality Choice of ψ. Text-to-video retrieval results on three datasets under different context modalities.

Dataset	Context $\mathbf{C}$	R@1↑	R@5↑	R@10↑	MnR↓
MSR-VTT	$\mathbf{T}_{\text{word}}$	47.4	**73.5**	82.1	12.9
	$\mathbf{V}_{\text{frame}}$	**49.1**	73.3	**82.2**	**12.4**
ActivityNet	$\mathbf{T}_{\text{word}}$	**42.6**	**73.6**	**84.4**	**6.8**
	$\mathbf{V}_{\text{frame}}$	40.2	72.2	83.6	8.1
DiDeMo	$\mathbf{T}_{\text{word}}$	46.5	74.3	82.6	12.3
	$\mathbf{V}_{\text{frame}}$	**47.6**	**75.4**	**83.1**	**12.0**

Table 5: Ablation on the interaction mode of ψ on Text-to-Video Retrieval results on MSR-VTT 1k-A. The variant removes the relative gap modeling by using t_i as the query and $\mathbf{V}_{\text{frame}}$ as the key–value, producing t'_{ij} and $\Delta_{ij} = v_j - t'_{ij}$. Our gap-aware design preserves pair-specific structure and yields superior alignment.

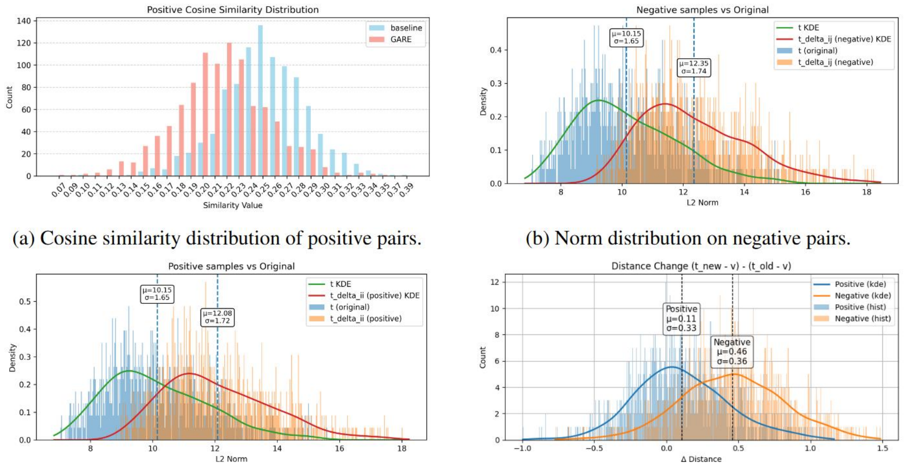
Interaction Mode of ψ	R@1↑	R@5↑	R@10↑	MnR↓
Query $= t_i$ (no gap)	46.1	73.2	81.9	13.7
Query $= v_j - t_i$	**49.1**	**74.7**	**83.6**	**12.0**

Table 6: Ablation on the IB prior $r(\Delta)$ on MSR-VTT 1k-A. Comparison between normalized and unnormalized Δ_{ij} distributions with different Gaussian priors.

σ	R@1↑	R@5↑	R@10↑	MnR↓
Normalized Δ				
1.0	47.8	74.5	82.1	12.9
Unnormalized Δ				
0.1	47.7	73.4	82.2	12.9
1.0	**49.1**	**74.7**	**83.6**	12.0
10.0	48.1	74.6	83.5	12.0
100.0	48.6	**74.7**	83.2	**11.8**

Qualitative Analysis

- **Lower Cosine Similarity**
  - → better **uniformity** on unit hypersphere
  - also can be seen as **lowering model's confidence** (belief mass)
  - see Fig.6 for hard negative comparison with baseline
    - GARE produces smoother logits than baseline
    - → semantic similar samples with similar logits
- **Larger $t_{\Delta_{ij}}$ Norm Magnitude on both positive and negative**
  - → expanding representation to a broader space region for **better fine-grained alignment**
- **Larger ℓ_2 distance between $t_{\Delta_{ij}}$ and v_j** compared to the pair of (t_i, v_j)
  - → also can be seen as promoting **uniformity**



(a) Cosine similarity distribution of positive pairs. (b) Norm distribution on negative pairs.

(c) Norms distribution on positive pairs. (d) Distance shift: t_Δ vs. t.

Figure 3: Qualitative analysis on the MSR-VTT 1k-A validation set. t_{delta} denotes t_Δ. Our method induces greater angular separation between positive pairs (a), redistributes t_Δ norms to release gradient tension (b, c), and pushes t_Δ outward from v_j (d), promoting uniformity.

(a) (b)

Figure 6: Comparison of hard negative alignment before and after applying Δ_{ij} optimization. Compared with the baseline, GARE produces smaller similarity gaps among semantically related videos v_j. This indicates that GARE effectively mitigates the noise from hard negatives and reduces the semantic deviation of the anchor t_i, leading to more stable and consistent alignment across similar samples.

Qualitative Analysis

- **Gradient Analysis: How Δ Redistributes Optimization Tension**
  - **Observation of Gradients on t_i**
    In dimensions with strong optimization activity, both positive and negative gradients reach similar magnitudes ($g \approx 2.5$) and appear as near opposites (Figure. 4).
  - **Gradient Redistribution**
    When aggregated across all pairs, opposite gradients cancel in the anchor update $\nabla_{t_i}\mathcal{L}_i(\Delta_{i*}) \rightarrow$ near zero (Figure. 7).
    Each Δ_{ij}, however, receives gradients **only from its own pair** $(t_i,\ v_j)$:
    • positive $\Delta_{ij} \approx + g$ • negative $\Delta_{ij} \approx - g/B$
    Thus, the total effective optimization strength per anchor $\approx |+g| + B \cdot |-g/B| \approx 2|g|$.
  - **Insight**
    Δ_{ij} components remain **actively optimized** and trace how t_i explores the representation space. By distributing gradient flow across Δ, the framework **offloads optimization tension** from anchors and **expands their reachable region**, breaking the **locality constraint** imposed by the modality gap.

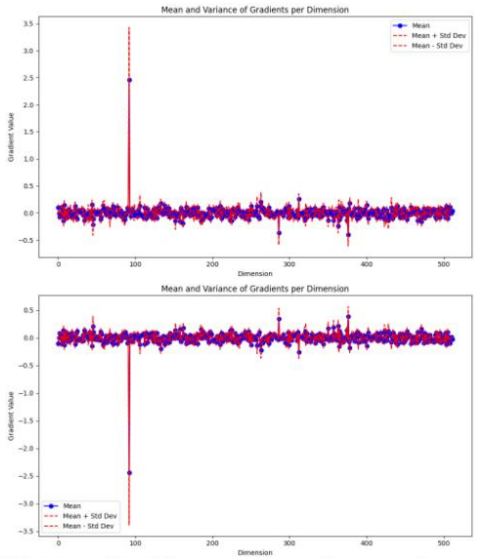


Figure 4: Mean and variance of per-dimension gradients, indicating the positive gradients (top) acting on $t_{\Delta_{ii}}$ and Δ_{ii} and the sum of all negative gradients (bottom) for $t_{\Delta_{ij}}$ and Δ_{ij}.

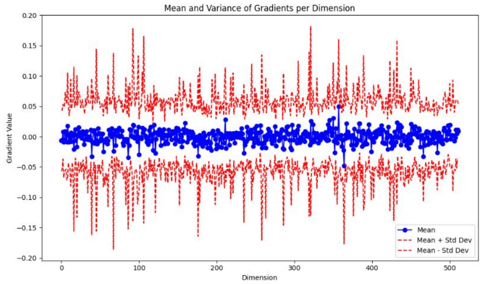


Figure 7: Mean and variance of total gradients acting on t_i on each dimension.

Thanks!