NEURAL INFORMATION PROCESSING SYSTEMS

QUEST: Quadruple Multimodal Contrastive Learning with Constraints and Self-Penalization

Qi Song[1]*, Tianxiang Gong[2]*, Shiqi Gao[2],
Haoyi Zhou[1,3]†, Jianxin Li[2,3]
[1]School of Software, Beihang University
[2]School of Computer Science and Engineering, Beihang University
[3]Zhongguancun Laboratory, Beijing
{songqi23, gongtx, gaoshiqi, haoyi, lijx}@buaa.edu

*Equal contribution
†Corresponding author.

Background & Motivation

- Contrastive learning treats all negative samples equally, ignoring the potential semantic relationships between negative samples and the anchor.
- Contrastive learning often neglects significant portions of input information, leading to feature suppression and shortcut learning.

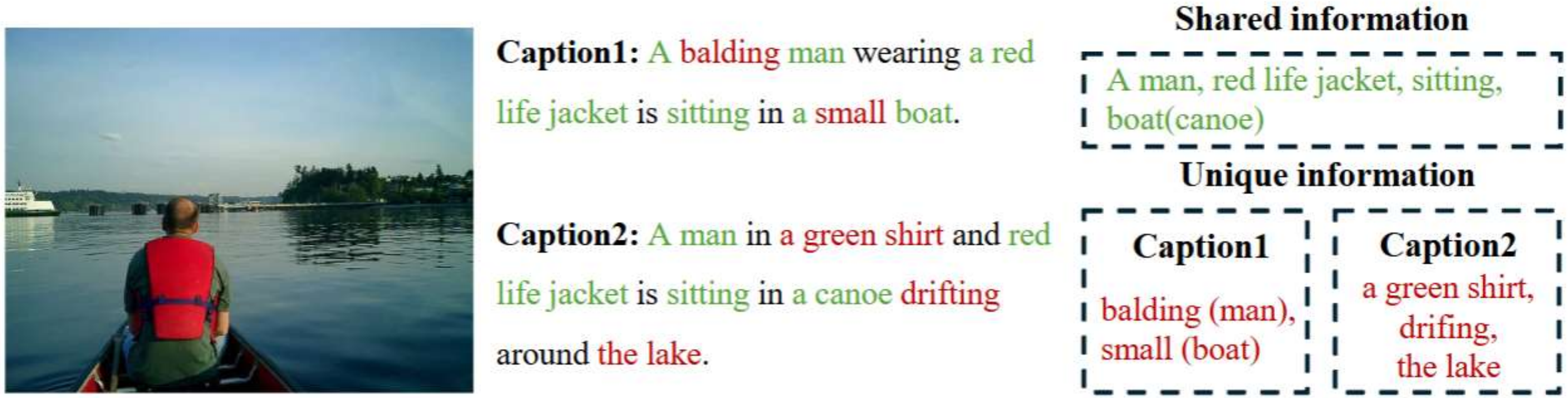


Figure1. Shared and unique Information in Multimodal Multi-view Scenario.

Background & Motivation

- Current contrastive learning methods focus on maximizing mutual information between two views while ignoring unique information.

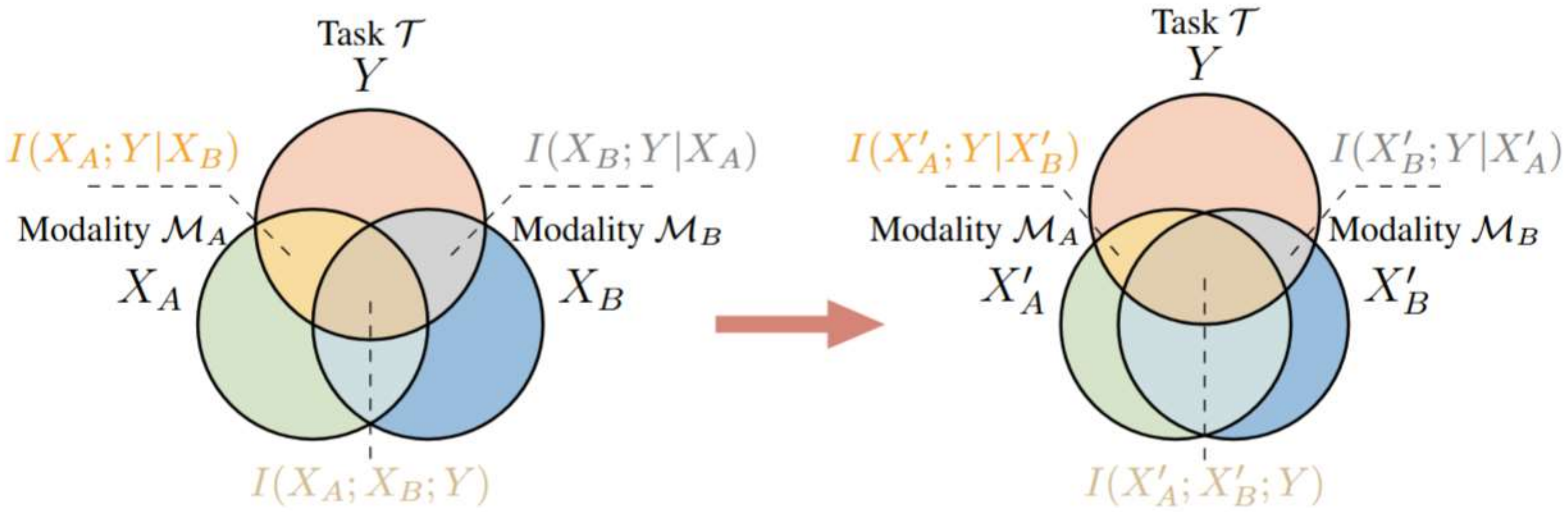


Figure 2: Feature suppression in multi-view contrastive learning. We define $I(X_A;X_B;Y)$ as task-related shared information, $I(X_A;Y|X_B)$ and $I(X_B;Y|X_A)$ as task-related unique information related to task Y in modalities X_A and X_B,respectively. Contrastive losses, such as InfoNCE, tend to maximize the task-related shared information while suppressing the task-related unique information in each modality. Left: before training with InfoNCE. Right: after training with InfoNCE.

Architecture Overview

- Use extra unique decoder to capture unique information simultaneously with constraints and self-penalization.

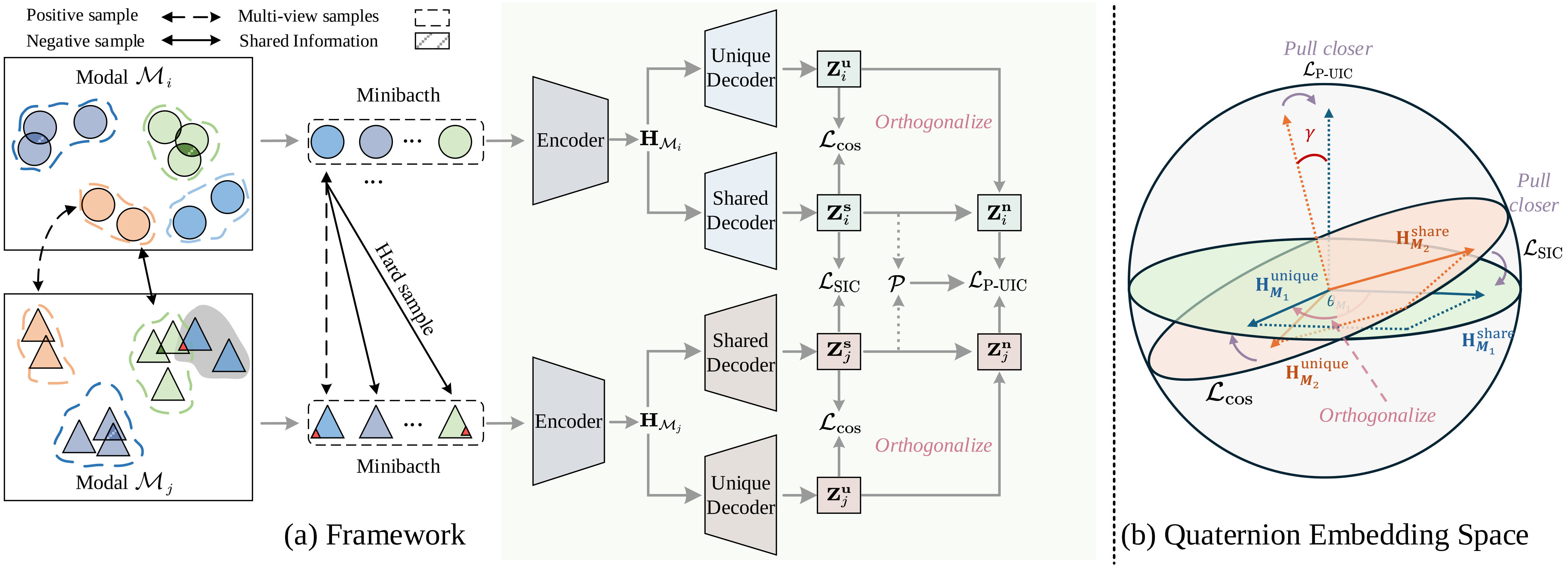


Figure3. Left: An overview of our method's framework. Right: The quaternion embedding space, in which we orthogonalize the representation of high-dimensional data to extend the solution space.

Shared Decoder and Unique Decoder

- For each modality, input data $\mathbf{X_i}$ is transformed by a modality-specific encoder $\mathcal{F}_{\mathcal{M}_i}(\cdot)$ into a general representation $\mathbf{H}_{\mathcal{M}_i}$. Two decoders $\mathcal{G}^{\mathbf{s}}_{\mathcal{M}_i}(\cdot)$ and $\mathcal{G}^{\mathbf{u}}_{\mathcal{M}_i}(\cdot)$ then separate shared and unique information from $\mathbf{H}_{\mathcal{M}_i}$.

$$
\begin{aligned}
\mathbf{Z}^{\mathbf{u}}_i &= \mathcal{G}^{\mathbf{u}}_{\mathcal{M}_i}(\mathbf{H}_{\mathcal{M}_i}; \Phi_i) = \mathcal{G}^{\mathbf{u}}_{\mathcal{M}_i}(\mathcal{F}_{\mathcal{M}_i}(\mathbf{X}_i; \Theta_i); \Phi_i), \\
\mathbf{Z}^{\mathbf{s}}_i &= \mathcal{G}^{\mathbf{s}}_{\mathcal{M}_i}(\mathbf{H}_{\mathcal{M}_i}; \Psi_i) = \mathcal{G}^{\mathbf{s}}_{\mathcal{M}_i}(\mathcal{F}_{\mathcal{M}_i}(\mathbf{X}_i; \Theta_i); \Psi_i).
\end{aligned}
$$

Shared Information Constraint

- In multimodal and multi-view scenarios, maximizing the lower bound of mutual information (MI) between representations from different views encourages the shared decoder to learn task-related agreement.

$$\mathcal{L}_{\text{SIC}} = \sum_{i,j} \mathbb{1}_{\mathcal{M}_i \neq \mathcal{M}_j} \mathbb{E}_{\mathbf{Z}_i^{\mathbf{s}}} \left[-\log \frac{\exp(s(\mathbf{Z}_i^{\mathbf{s}}, \mathbf{Z}_j^{\mathbf{s}+})/\tau)}{\exp(s(\mathbf{Z}_i^{\mathbf{s}}, \mathbf{Z}_j^{\mathbf{s}+})/\tau) + \sum_{k=1}^{m} \mathbb{1}_{\hat{\mathbf{y}}^-} \exp(s(\mathbf{Z}_i^{\mathbf{s}}, \mathbf{Z}_{jk}^{\mathbf{s}\,-})/\tau)} \right].$$

Unique Information Constraint

- In contrast to shared information, unique information is modality-specific and task-relevant, providing essential insights for downstream tasks.
- Firstly, we derive the representation space of normal vectors for shared and unique embedding spaces through cross-product calculations:

$$\mathbf{Z}_i^{\mathbf{n}} = \mathbf{Z}_i^{\mathbf{s}} \times \mathbf{Z}_i^{\mathbf{u}}$$

- In the newly projected space, our objectives aim to maximize the alignment of unique representation from different modalities within the plane spanned by the shared representation

$$\mathcal{L}_{\text{UIC}} = \sum_{i,j} \mathbb{1}_{\mathcal{M}_i \neq \mathcal{M}_j} \mathbb{E}_{\mathbf{Z}_i^{\mathbf{n}}} \left[-\log \frac{\exp(s(\mathbf{Z}_i^{\mathbf{n}}, \mathbf{Z}_j^{\mathbf{n}+})/\tau)}{\exp(s(\mathbf{Z}_i^{\mathbf{n}}, \mathbf{Z}_j^{\mathbf{n}+})/\tau) + \sum_{k=1}^{m} \mathbb{1}_{\hat{\mathbf{y}}^-} \exp(s(\mathbf{Z}_i^{\mathbf{n}}, \mathbf{Z}_{jk}^{\mathbf{n}-})/\tau)} \right] + \sum_i \sum_j \frac{\mathbf{Z}_{ij}^{\mathrm{s}} \cdot \mathbf{Z}_{ij}^{\mathrm{u}}}{\| \mathbf{Z}_{ij}^{\mathrm{s}} \| \| \mathbf{Z}_{ij}^{\mathrm{u}} \|}$$

Self-Penalization Constraint

- Challenges in practical Contrastive learning implementations
  - Uniform treatment of all other samples within a batch B as negative examples
  - Misclassification of semantically similar samples as negative
- Use the intra-model shared information similar to penalization to term guide the optimization of unique information

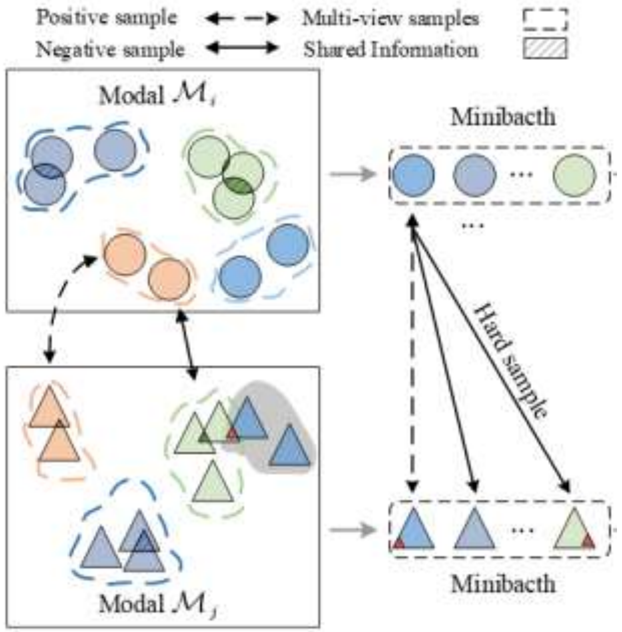


Figure 3(a). Potential correlations among data.

Solution

Self-Penalization Term

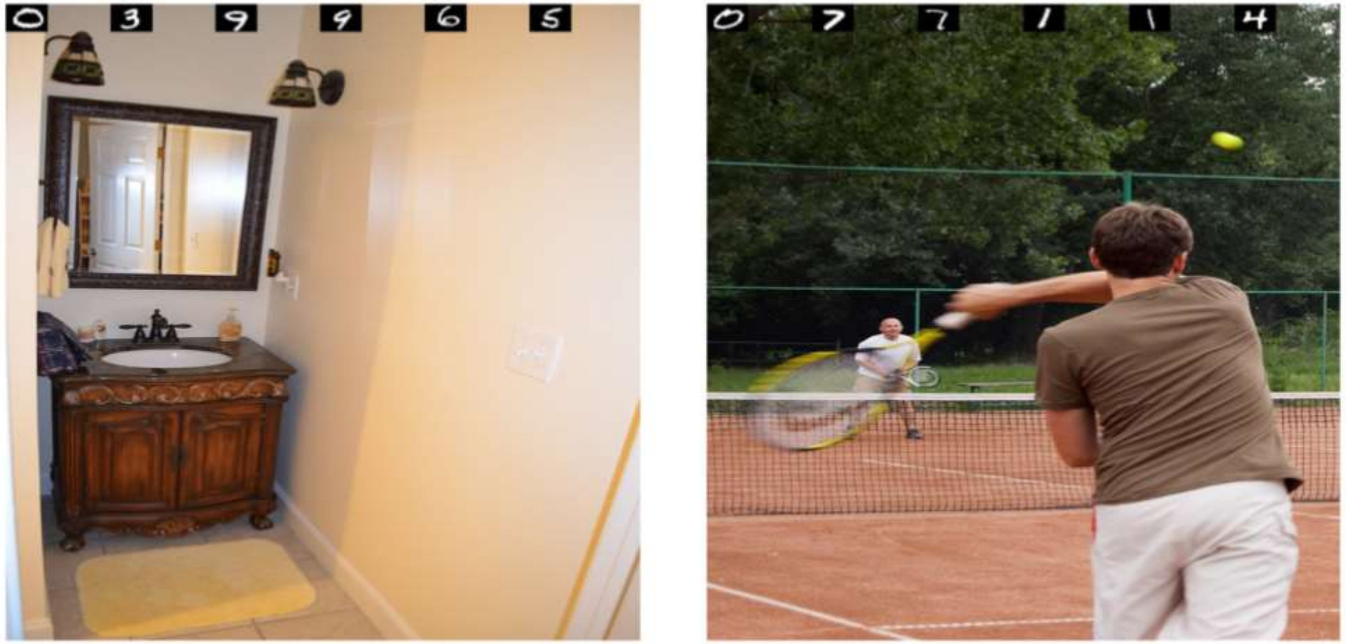
$$\mathcal{P} = \exp(\lambda[\mathbf{S} - diag(\mathbf{S}) + \mathbf{I}]), \qquad where\ \mathbf{S} = \mathbf{Z}_i^s \mathbf{Z}_j^{sT}$$

Penalized UIC

$$\mathcal{L}_{\text{P-UIC}} = \sum_{i,j} \mathbb{1}_{\mathcal{M}_i \neq \mathcal{M}_j} \mathbb{E}_{\mathbf{Z}_i^{\mathbf{n}}} \left[-\log \frac{\exp(\mathcal{P}^+ \cdot s(\mathbf{Z}_i^{\mathbf{n}}, \mathbf{Z}_j^{\mathbf{n}+})/\tau)}{\exp(\mathcal{P}^+ \cdot s(\mathbf{Z}_i^{\mathbf{n}}, \mathbf{Z}_j^{\mathbf{n}+})/\tau) + \sum_{k=1}^m \mathbb{1}_{\hat{\mathbf{y}}^-} \exp(\mathcal{P} \cdot s(\mathbf{Z}_i^{\mathbf{n}}, \mathbf{Z}_{jk}^{\mathbf{n}-})/\tau)} \right] + \sum_i \sum_j \frac{\mathbf{Z}_{ij}^{\mathrm{s}} \cdot \mathbf{Z}_{ij}^{\mathrm{u}}}{\| \mathbf{Z}_{ij}^{\mathrm{s}} \| \| \mathbf{Z}_{ij}^{\mathrm{u}} \|}$$

Experiment Setting

- Datasets: Flickr30k, MS-COCO, Flickr30k-shortcuts, MS-COCO-shortcuts
- Baseline: Vanilla InfoNCE, Latent Target Decoding (LTD), Implicit Feature Modification (IFM)



(a) **Caption**: "A bathroom sink with wood finish cabinets. 0 3 9 9 6 5."

(b) **Caption**: "A guy in a brown shirt has just hit a tennis ball. 0 7 7 1 1 4."

Figure 4 [1]. Two andom samples from the MS-COCO dataset including shortcuts added on both the image andcaption.

[1]Demonstrating and Reducing Shortcuts in Vision-Language Representation Learning

■ Results on public benchmark

Table1:Result on Flickr30k and MS-COCO with varied method. sc denotes shortcut, we evaluate CLIP and VSE++ w/wo shortcut on i2t and i2i task. QUEST outperforms InfoNCE and achieve superior performance compare with other baselines in most cases. †denote use of ltd.

Method	sc	Flickr30k *i2t* R@1	Flickr30k *i2t* R@5	Flickr30k *i2t* R@10	Flickr30k *t2i* R@1	Flickr30k *t2i* R@5	Flickr30k *t2i* R@10	Flickr30k RSUM	MS-COCO *i2t* R@1	MS-COCO *i2t* R@5	MS-COCO *i2t* R@10	MS-COCO *t2i* R@1	MS-COCO *t2i* R@5	MS-COCO *t2i* R@10	MS-COCO RSUM
CLIP															
$\mathcal{L}_{\text{InfoNCE}}$	✗	$86.9_{\pm0.1}$	$97.4_{\pm0.1}$	$99.0_{\pm0.0}$	$72.4_{\pm0.1}$	$92.1_{\pm0.0}$	$95.8_{\pm0.0}$	$543.5_{\pm1.1}$	$63.8_{\pm0.3}$	$86.1_{\pm0.2}$	$92.3_{\pm0.0}$	$46.3_{\pm0.3}$	$74.8_{\pm0.1}$	$84.1_{\pm0.2}$	$447.5_{\pm0.5}$
$\mathcal{L}_{\text{InfoNCE+LTD}}$	✗	$86.5_{\pm0.6}$	$97.1_{\pm0.0}$	$98.5_{\pm0.0}$	$72.4_{\pm0.0}$	$\mathbf{92.3}_{\pm0.0}$	$\mathbf{95.9}_{\pm0.0}$	$542.8_{\pm0.8}$	$63.8_{\pm0.0}$	$86.1_{\pm0.0}$	$92.3_{\pm0.0}$	$46.3_{\pm0.0}$	$74.7_{\pm0.0}$	$84.1_{\pm0.0}$	$447.4_{\pm0.0}$
$\mathcal{L}_{\text{InfoNCE+IFM}}$	✗	$87.4_{\pm0.1}$	$97.4_{\pm0.2}$	$99.1_{\pm0.0}$	$73.2_{\pm0.0}$	$92.2_{\pm0.0}$	$95.6_{\pm0.0}$	$544.9_{\pm0.2}$	$63.0_{\pm0.1}$	$86.6_{\pm0.1}$	$92.6_{\pm0.2}$	$47.2_{\pm0.0}$	$75.6_{\pm0.0}$	$84.5_{\pm0.0}$	$449.5_{\pm1.7}$
$\mathcal{L}_{\text{QUEST(Ours)}}$	✗	$\mathbf{89.3}_{\pm0.3}$	$\mathbf{97.8}_{\pm0.2}$	$\mathbf{99.2}_{\pm0.3}$	$\mathbf{73.9}_{\pm0.5}$	$91.5_{\pm0.3}$	$95.0_{\pm0.3}$	$\mathbf{546.7}_{\pm1.9}$	$\mathbf{65.4}_{\pm0.5}$	$\mathbf{87.7}_{\pm0.2}$	$\mathbf{93.6}_{\pm0.4}$	$\mathbf{48.5}_{\pm0.2}$	$\mathbf{75.7}_{\pm0.5}$	$\mathbf{84.7}_{\pm0.6}$	$\mathbf{455.6}_{\pm2.4}$
$\mathcal{L}_{\text{InfoNCE}}$	✓	$57.2_{\pm8.3}$	$84.0_{\pm4.8}$	$91.0_{\pm1.9}$	$44.9_{\pm4.5}$	$74.9_{\pm6.0}$	$84.2_{\pm2.5}$	$436.2_{\pm145.0}$	$13.6_{\pm0.9}$	$31.5_{\pm2.4}$	$42.2_{\pm3.7}$	$7.3_{\pm0.6}$	$22.1_{\pm1.0}$	$32.7_{\pm1.7}$	$149.4_{\pm32.7}$
$\mathcal{L}_{\text{InfoNCE+LTD}}$	✓	$64.0_{\pm1.3}$	$87.8_{\pm0.9}$	$93.2_{\pm0.8}$	$50.7_{\pm0.6}$	$79.8_{\pm0.7}$	$88.1_{\pm0.5}$	$463.6_{\pm17.3}$	$18.9_{\pm0.1}$	$41.8_{\pm0.1}$	$54.1_{\pm0.1}$	$16.5_{\pm0.0}$	$39.4_{\pm0.0}$	$52.6_{\pm0.1}$	$223.4_{\pm0.2}$
$\mathcal{L}_{\text{InfoNCE+IFM}}$	✓	$73.8_{\pm0.8}$	$91.5_{\pm0.5}$	$95.6_{\pm0.0}$	$58.9_{\pm0.1}$	$84.4_{\pm0.1}$	$91.1_{\pm0.2}$	$495.2_{\pm5.7}$	$23.4_{\pm1.5}$	$46.5_{\pm2.7}$	$58.2_{\pm2.5}$	$17.1_{\pm0.3}$	$38.9_{\pm0.9}$	$51.3_{\pm1.0}$	$235.5_{\pm43.8}$
$\mathcal{L}_{\text{QUEST(Ours)}}$	✓	$\mathbf{84.2}_{\pm0.3}$	$\mathbf{96.0}_{\pm0.1}$	$\mathbf{97.7}_{\pm0.2}$	$\mathbf{67.6}_{\pm0.5}$	$\mathbf{88.9}_{\pm0.2}$	$\mathbf{93.4}_{\pm0.1}$	$\mathbf{527.8}_{\pm1.4}$	$\mathbf{50.8}_{\pm0.3}$	$\mathbf{75.4}_{\pm0.4}$	$\mathbf{84.1}_{\pm0.4}$	$\mathbf{37.9}_{\pm0.3}$	$\mathbf{65.1}_{\pm0.3}$	$\mathbf{76.1}_{\pm0.4}$	$\mathbf{389.4}_{\pm2.1}$
VSE++															
$\mathcal{L}_{\text{InfoNCE}}$	✗	$52.6_{\pm1.1}$	$79.8_{\pm0.1}$	$87.8_{\pm0.1}$	$39.5_{\pm0.3}$	$69.8_{\pm0.0}$	$79.4_{\pm0.1}$	$409.0_{\pm4.0}$	$42.2_{\pm0.1}$	$72.7_{\pm0.1}$	$83.2_{\pm0.1}$	$30.9_{\pm0.0}$	$61.2_{\pm0.1}$	$73.5_{\pm0.1}$	$363.8_{\pm2.3}$
$\mathcal{L}_{\text{InfoNCE+LTD}}$	✗	$54.1_{\pm0.1}$	$81.1_{\pm0.8}$	$88.6_{\pm0.1}$	$42.5_{\pm0.0}$	$71.9_{\pm0.1}$	$81.3_{\pm0.0}$	$419.6_{\pm0.1}$	$43.6_{\pm0.1}$	$73.5_{\pm0.0}$	$83.7_{\pm0.0}$	$32.4_{\pm0.1}$	$62.5_{\pm0.0}$	$74.7_{\pm0.0}$	$370.5_{\pm0.1}$
$\mathcal{L}_{\text{InfoNCE+IFM}}$	✗	$52.4_{\pm0.2}$	$76.9_{\pm0.1}$	$85.3_{\pm0.0}$	$39.1_{\pm0.0}$	$68.8_{\pm0.1}$	$78.2_{\pm0.1}$	$400.7_{\pm0.0}$	$40.2_{\pm0.0}$	$70.8_{\pm0.1}$	$81.6_{\pm0.1}$	$30.8_{\pm0.0}$	$61.5_{\pm0.0}$	$74.3_{\pm0.0}$	$359.3_{\pm1.1}$
$\mathcal{L}_{\text{QUEST(Ours)}}$	✗	$\mathbf{54.7}_{\pm0.2}$	$\mathbf{81.3}_{\pm0.4}$	$\mathbf{88.8}_{\pm0.3}$	$\mathbf{42.9}_{\pm0.1}$	$\mathbf{72.3}_{\pm0.4}$	$\mathbf{81.6}_{\pm1.1}$	$\mathbf{421.6}_{\pm2.5}$	$\mathbf{45.3}_{\pm0.1}$	$\mathbf{75.5}_{\pm0.5}$	$\mathbf{85.4}_{\pm0.4}$	$\mathbf{34.1}_{\pm0.1}$	$\mathbf{64.5}_{\pm0.2}$	$\mathbf{76.3}_{\pm0.2}$	$\mathbf{381.1}_{\pm1.5}$
$\mathcal{L}_{\text{InfoNCE}}$	✓	$0.1_{\pm0.0}$	$0.4_{\pm0.0}$	$0.8_{\pm0.0}$	$0.1_{\pm0.0}$	$0.4_{\pm0.0}$	$1.0_{\pm0.0}$	$2.9_{\pm0.0}$	$0.0_{\pm0.0}$	$0.1_{\pm0.0}$	$0.2_{\pm0.0}$	$0.0_{\pm0.0}$	$0.1_{\pm0.0}$	$0.2_{\pm0.0}$	$0.6_{\pm0.0}$
$\mathcal{L}_{\text{InfoNCE+LTD}}$	✓	$24.7_{\pm0.5}$	$\mathbf{51.8}_{\pm0.7}$	$\mathbf{65.6}_{\pm1.4}$	$\mathbf{20.7}_{\pm1.0}$	$\mathbf{49.2}_{\pm0.6}$	$\mathbf{62.6}_{\pm1.2}$	$\mathbf{274.6}_{\pm4.6}$	$3.9_{\pm0.0}$	$13.7_{\pm0.6}$	$21.6_{\pm0.9}$	$3.1_{\pm0.2}$	$11.0_{\pm1.6}$	$18.1_{\pm3.0}$	$71.4_{\pm3.6}$
$\mathcal{L}_{\text{InfoNCE+IFM}}$	✓	$0.0_{\pm0.0}$	$0.6_{\pm0.1}$	$0.9_{\pm0.2}$	$0.1_{\pm0.0}$	$0.5_{\pm0.0}$	$1.0_{\pm0.0}$	$3.2_{\pm0.8}$	$0.0_{\pm0.0}$	$0.1_{\pm0.0}$	$0.2_{\pm0.0}$	$0.0_{\pm0.0}$	$0.1_{\pm0.0}$	$0.2_{\pm0.0}$	$0.7_{\pm0.0}$
$\mathcal{L}_{\text{QUEST(Ours)}^\dagger}$	✓	$\mathbf{24.9}_{\pm0.4}$	$48.4_{\pm0.3}$	$61.1_{\pm0.5}$	$17.5_{\pm0.3}$	$43.4_{\pm0.6}$	$56.5_{\pm0.8}$	$251.8_{\pm2.9}$	$\mathbf{10.5}_{\pm0.6}$	$\mathbf{27.9}_{\pm0.3}$	$\mathbf{40.6}_{\pm0.9}$	$\mathbf{9.4}_{\pm0.5}$	$\mathbf{29.0}_{\pm1.4}$	$\mathbf{42.6}_{\pm2.1}$	$\mathbf{160.0}_{\pm5.8}$

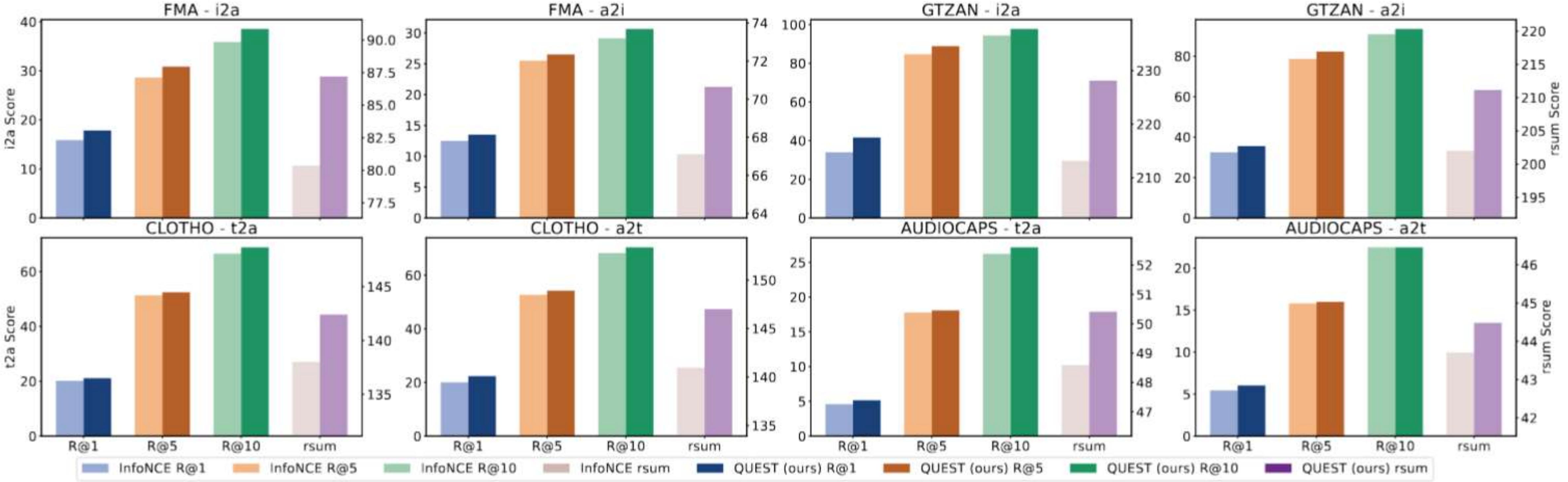
Ablation Study

Table2: ablation study on image caption retrieval task with different training objectives. D1 and D2denote decoders in the architecture. Decoder with all × beneath are omitted, while those with √ indicate optimization with corresponding objective functions. Bold and underlined numbers indicate the best and second-best results, respectively.

Methods				Flickr30k							MS-COCO						
D1	D2			i2t			t2i			RSUM	i2t			t2i			RSUM
$\mathcal{L}_{\text{SIC}}$	$\mathcal{L}_{\text{SIC}}$	$\mathcal{L}_{\text{UIC}}$	$\mathcal{L}_{\text{P-UIC}}$	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
				CLIP							CLIP						
✓	✗	✗	✗	86.9	97.4	98.8	72.4	**92.1**	**95.8**	543.4	63.8	86.1	92.3	46.3	74.8	84.1	447.4
✗	✗	✓	✗	80.8	94.4	96.5	66.9	88.4	92.9	519.9	55.1	81.6	89.4	43.3	72.5	82.9	424.8
✗	✗	✗	✓	85.5	96.6	98.1	70.0	87.3	90.7	528.2	63.0	85.9	91.7	45.2	70.4	79.2	435.4
✓	✗	✓	✗	81.9	96.2	98.0	69.5	90.0	94.2	529.8	64.9	86.6	92.8	47.3	75.1	83.9	450.6
✓	✓	✗	✗	76.6	92.3	95.9	59.0	84.4	90.9	499.1	54.4	79.2	86.7	39.9	68.4	78.9	407.5
✓	✗	✗	✓	**89.3**	**97.8**	**99.2**	**73.9**	91.5	95.0	**546.7**	**65.4**	**87.7**	**93.6**	**48.5**	**75.7**	**84.7**	**455.6**
				VSE++							VSE++						
✓	✗	✗	✗	52.6	79.8	87.8	39.5	69.8	79.4	408.9	42.2	72.7	83.2	30.9	61.2	73.5	363.7
✗	✗	✓	✗	47.8	72.8	80.8	36.7	63.2	73.3	374.6	40.7	71.2	82.1	30.3	60.5	73.0	357.8
✗	✗	✗	✓	49.0	74.1	81.3	36.4	64.1	73.1	378.0	40.9	71.4	82.4	30.8	60.6	73.2	359.3
✓	✗	✓	✗	53.3	79.8	87.6	40.5	68.1	78.0	407.3	44.9	74.1	84.4	32.3	62.8	74.7	373.2
✓	✗	✗	✓	**54.7**	**80.3**	**88.2**	**42.0**	**70.3**	**79.6**	**415.1**	**45.3**	**75.5**	**85.4**	**34.1**	**64.5**	**76.3**	**381.1**

■ Extend To More Modalities

Figure 5: Performance comparison of InfoNCE and QUEST methods with additional audio modality on image-to-audio (i2a) and audio-to-image (a2i) retrieval tasks across FMA, GTZAN, CLOTHO, and AUDIOCAPS datasets.



■ Visualization

Figure 6: Case Study: (a) Image-to-text retrieval , where the results of L_{QUEST} and $L_{InfoNCE}$ are denoted by italics and underlines, respectively. (b) Text-to-image retrieval, where red and green borders indicate the top-5 retrievals using L_{QUEST}, while blue borders represent those using $L_{InfoNCE}$.

Conclusions

- We propose QUEST, a novel multimodal contrastive learning method that effectively captures and preserves more task-relevant unique information from individual modalities.

- Quaternions indirectly optimize unique and shared information simultaneously, expanding the unique representational space.

- Unique decoders extract unique and shared information through constraints and self-penalization, outperforming SOTA methods.

NEURAL INFORMATION PROCESSING SYSTEMS

QUEST: Quadruple Multimodal Contrastive Learning with Constraints and Self-Penalization

Qi Song[1]*, Tianxiang Gong[2]*, Shiqi Gao[2], Haoyi Zhou[1,3]†, Jianxin Li[2,3]

[1]School of Software, Beihang University
[2]School of Computer Science and Engineering, Beihang University
[3]Zhongguancun Laboratory, Beijing
{songqi23, gongtx, gaoshiqi, haoyi, lijx}@buaa.edu

*Equal contribution
†Corresponding author.