TOWARD ROBUST MULTIMODAL RETRIEVAL FOR MLLMS WITH TRANSPORT ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal retrieval ranks image–text candidates according to semantic relevance and supports applications such as image-to-text retrieval, text-to-image retrieval, and composed retrieval. However, existing multimodal retrieval models are highly sensitive to adversarial perturbations on image queries. Even when perturbations are nearly imperceptible, they can shift the query representation away from the original cross-modal semantic structure, distort matching between the query and candidates, and sharply degrade ranking performance. To address this problem, we propose ROTA, a training-free framework for robust multimodal retrieval with transport alignment. ROTA constructs transformed views of the attacked query and uses retrieval entropy to select reliable views, rather than relying on a single corrupted query representation. It further enriches candidate texts with descriptive views and aligns image-view and text-view distributions using optimal transport. The resulting retrieval score combines reconstructed query evidence with distribution-level image–text alignment, reducing reliance on brittle single-view matching. Furthermore, we establish OT margin stability under bounded cost perturbations with fixed transport marginals, and conditionally relate this bound to image-space perturbations. Experiments across retrieval benchmarks, adversarial attacks, and retrieval backbones show that ROTA consistently improves performance over direct retrieval with attacked queries.

1 INTRODUCTION

Multimodal retrieval (Wang et al., 2024; Cao et al., 2022) ranks cross-modal candidates according to their semantic relevance to a query, covering tasks such as image-to-text retrieval (Cao et al., 2022), text-to-image retrieval (Zhan et al., 2025; Yu et al., 2022), and composed retrieval (Zhang et al., 2025). In recent years, vision-language embedding models (Jiang et al., 2024; Li et al., 2026; Günther et al., 2025) have become a common paradigm for multimodal retrieval by mapping images and texts into a shared representation space and matching them with cosine similarity. However, this embedding-similarity-based retrieval process is highly sensitive to adversarial perturbations (Zhang et al., 2024a; Zhou et al., 2024a; Szegedy et al., 2014; Goodfellow et al., 2015) on image queries: even when perturbations are nearly imperceptible, they can shift the query representation away from the original cross-modal semantic structure and sharply degrade the final retrieval ranking.

Figure 1(a) shows this ranking failure: the clean query retrieves the correct caption at rank 1, while an imperceptible perturbation demotes the true match to rank 3 and ranks a mismatched caption first. Figure 1(b) further shows this representation shift in the embedding distribution: as indicated by the red points, adversarial perturbations shift image features away from the clean cross-modal semantic region represented by the blue and green points. Since retrieval ranking depends on cross-modal similarity between the query and candidates, this semantic shift can distort relative matching scores and cause ranking errors (Zhang et al., 2024b; Fang et al., 2025; Xie et al., 2025).

To address this problem, we propose ROTA, a training-free framework for robust multimodal retrieval with transport alignment. Rather than relying on a single corrupted query representation, ROTA constructs transformed views of the attacked query, selects reliable ones using retrieval entropy, and uses the selected views to reconstruct more stable query evidence. Beyond query reconstruction, ROTA enriches candidate texts with descriptive views that preserve their original meanings while providing additional semantic context (Li et al., 2024; Gao et al., 2023). It then aligns image-view

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

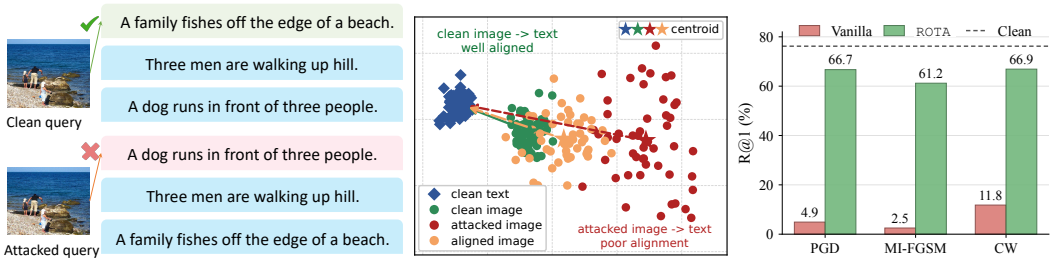


Figure 1: Effect of adversarial perturbations and ROTA on multimodal retrieval. **(a)** A small image perturbation changes the retrieval order and demotes the ground-truth caption from rank 1 to rank 3. **(b)** Attacked queries move away from the clean embedding region, while ROTA reconstructs them toward it. **(c)** R@1 comparison of direct attacked retrieval and ROTA on Flickr30K I2T retrieval with jina under PGD, MI-FGSM, and CW. Dashed lines show clean-query performance.

and text-view distributions with optimal transport (OT) (Villani, 2009; Pan et al., 2026), producing a distribution-level retrieval score rather than a cosine similarity between a single pair of views. As indicated by the orange points in Figure 1(b), the reconstructed representations are closer to the clean cross-modal semantic region, reducing the matching shift caused by adversarial perturbations.

Empirically, Figure 1(c) summarizes representative results and shows that, on Flickr30K I2T retrieval with jina under three attacks, direct retrieval suffers a large R@1 drop, while ROTA recovers much of the lost performance and approaches the clean-query baseline. Broader experiments across retrieval benchmarks, attack types, and backbone models confirm consistent robust retrieval gains and show that robustness varies across datasets and model families. Our theoretical analysis shows that, with view indices and transport marginals held fixed, the correct candidate remains top-ranked when the clean retrieval margin is sufficiently large relative to changes in the matching costs (Theorem 1).

Our main contributions are:

- We propose a training-free framework for multimodal retrieval with adversarially perturbed image queries, which stabilizes attacked representations and improves cross-modal matching.
- We derive reliable query representations from transformed views of attacked images, enrich candidate texts with semantic descriptions, and use optimal transport to align the resulting visual and textual distributions for robust matching.
- We provide a robustness analysis for the OT-based retrieval score under bounded cost perturbations, and extensive experiments across retrieval benchmarks, adversarial attacks, and backbone models show that ROTA consistently improves robust retrieval performance.

2 Related Work

Multimodal retrieval. Contrastive vision-language pre-training has become a core approach for multimodal retrieval. CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) learn shared image-text spaces from large-scale paired data, enabling retrieval by feature similarity. SigLIP (Zhai et al., 2023) replaces the softmax-based contrastive loss with a pairwise sigmoid loss, improving training efficiency and scalability. LLM2CLIP (Huang et al., 2024) enhances CLIP's representation space by leveraging the rich textual knowledge encoded in large language models, improving both text-image alignment and retrieval accuracy. BLIP (Li et al., 2022) further improves vision-language pre-training with bootstrapped captions and supports image-text retrieval together with other vision-language tasks. More recent embedding models, such as E5-V (Jiang et al., 2024), mmE5 (Chen et al., 2025), and jina-embeddings-v4 (Günther et al., 2025), extend this paradigm toward universal multimodal embeddings. These backbones typically rank candidates using a single query–candidate embedding pair. We study the robustness of this matching scheme to adversarial image queries.

Adversarial robustness in vision-language models. Adversarial visual perturbations can disrupt vision-language representations (Tian et al., 2025; Lu et al., 2023; Zhou et al., 2023) and cross-modal

retrieval rankings (Fu et al., 2024). Defenses such as TeCoA (Mao et al., 2022), FARE (Schlarmann et al., 2024), and AdvCLIP-LoRA (Ghiasvand et al., 2025) use adversarial fine-tuning, which is costly and can overfit to the fine-tuning distribution (Zhou et al., 2024b). Test-time or training-free approaches include TPT (Shu et al., 2022), which adapts prompts from augmented views; TAPT (Wang et al., 2025), which learns defensive bimodal prompts via multi-view entropy minimization and adversarial-clean alignment; PDA (Xu et al., 2026), which augments texts through paraphrasing and decomposition; MTA (Zanella & Ben Ayed, 2024), which seeks feature-space density modes with MeanShift-based augmentation; and AWT (Zhu et al., 2024), which combines augmentation, confidence weighting, and transport alignment. These approaches motivate view-based evidence and structured matching but do not directly address adversarial image-query retrieval. ROTA refines attacked queries with reliable transformed views, enriches candidate texts with descriptive views, and uses entropy-weighted optimal transport for more stable cross-modal retrieval scores.

3 METHODOLOGY

We present ROTA in the image-to-text setting; the text-to-image formulation follows analogously. Section 3.1 introduces the problem setting, Section 3.2 presents robust query reconstruction, Section 3.3 describes cross-modal transport matching, and Section 3.4 presents the theoretical analysis.

3.1 PRELIMINARIES

Multimodal retrieval. Given a query q and a gallery $\mathcal{G} = \{g_i\}$, multimodal retrieval ranks cross-modal gallery items according to their semantic relevance to the query. In the image-to-text setting, the query is an image and each gallery item is a text candidate; in the text-to-image setting, the query is a text description and each gallery item is an image candidate. The retrieval backbone contains an image encoder $f_I(\cdot)$ and a text encoder $f_T(\cdot)$, which project the two modalities into a shared embedding space. The standard matching score is computed by cosine similarity:

$$s(q, g_i) = \langle f_I(q), f_T(g_i) \rangle \,, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product between normalized embeddings. The gallery is ranked according to these scores.

Adversarial attack. For a matched pair (x, g_y) with $g_y \in \mathcal{G}$, a common loss is cross-entropy:

$$\mathcal{L}_{\text{ret}}(x, \mathcal{G}; y) = -\log \frac{\exp(s(x, g_y))}{\sum_{g_j \in \mathcal{G}} \exp(s(x, g_j))}. \tag{2}$$

With the encoders and gallery fixed, an untargeted attack finds a perturbed query $x^{\text{adv}} = x + \delta$ by maximizing this loss (Madry et al., 2018; Dong et al., 2018):

$$\delta^* = \arg\max_{\|\delta\|_p \le \epsilon} \mathcal{L}_{\text{ret}}(x + \delta, \mathcal{G}; y), \tag{3}$$

where ϵ is the perturbation budget. Minimizing the loss favors the matched text over competing candidates; maximizing it lowers its relative matching probability. Our goal is to preserve reliable cross-modal ranking under such perturbations.

3.2 ROBUST QUERY REPRESENTATION

To reduce the sensitivity of adversarial retrieval to local feature perturbations, we construct transformed image views (Zhu et al., 2024) around the attacked query, denoted by $\mathcal{X}(x^{\text{adv}}) = \{\mathbf{x}^{(n)}\}_{n=0}^{N}$, where $\mathbf{x}^{(0)}$ is the attacked-image feature and $\mathbf{x}^{(n)}$ for $n \ge 1$ are features produced by a diverse transformation pool (Zhu et al., 2024; Zeng et al., 2020; Pérez et al., 2021). We select reliable views using retrieval entropy: for each $\mathbf{x}^{(n)}$, we compute $p_i^{(n)} \propto \exp(\langle \mathbf{x}^{(n)}, \mathbf{t}_i \rangle)$over the gallery, where $\mathbf{t}_i$ is the base text representation of candidate g_i, and measure $H_I^{(n)} = -\sum_i p_i^{(n)} \log p_i^{(n)}$. Low-entropy views (Wang et al., 2021; Sheng et al., 2025) induce more concentrated retrieval responses; we retain their indices as $\mathcal{K} \subseteq \{1, \ldots, N\}$. Motivated by test-time defense observations that high-quality transformed views move coherently in feature space, we compute their mean displacement:

$$\bar{\Delta} = \frac{1}{|\mathcal{K}|} \sum_{n \in \mathcal{K}} \left(\mathbf{x}^{(n)} - \mathbf{x}^{(0)} \right), \tag{4}$$

define the normalized direction as $\mathbf{d} = \bar{\Delta}/\|\bar{\Delta}\|_2$, and write the normalized displacement of each selected view as $\mathbf{r}^{(n)} = (\mathbf{x}^{(n)} - \mathbf{x}^{(0)})/\|\mathbf{x}^{(n)} - \mathbf{x}^{(0)}\|_2$. We quantify directional coherence by the average projection onto $\mathbf{d}$:

$$\eta = \frac{1}{|\mathcal{K}|} \sum_{n \in \mathcal{K}} \left\langle \mathbf{r}^{(n)}, \mathbf{d} \right\rangle. \tag{5}$$

Here η measures the concentration of the transformed features around the dominant direction. To account for different levels of directional consistency, we use the average selected view $\bar{\mathbf{x}}_{\mathcal{K}} = |\mathcal{K}|^{-1} \sum_{n \in \mathcal{K}} \mathbf{x}^{(n)}$ for adaptive reconstruction:

$$\hat{\mathbf{x}}^{(0)} = \begin{cases} \mathbf{x}^{(0)} + \lambda \bar{\Delta}, & \eta > \tau, \\ \bar{\mathbf{x}}_{\mathcal{K}}, & \eta \le \tau, \end{cases} \tag{6}$$

where τ is a consistency threshold and λ controls the shift magnitude. At $\lambda = 1$, both branches equal $\bar{\mathbf{x}}_{\mathcal{K}}$. Otherwise, coherent views guide a directional shift, while less coherent views are averaged. Writing $\hat{\mathbf{x}}^{(n)} = \mathbf{x}^{(n)}$ for $n \in \mathcal{K}$, the matching set is:

$$\hat{\mathcal{X}}(x^{\text{adv}}) = \left\{ \hat{\mathbf{x}}^{(n)} \mid n \in \{0\} \cup \mathcal{K} \right\}. \tag{7}$$

3.3 Cross-Modal Transport Matching

Stabilizing the visual query is not sufficient for robust retrieval, because candidate texts may be semantically ambiguous. Following transport-based adaptation strategies (Zhu et al., 2024), we construct multiple textual views for each candidate, denoted by $\mathcal{T}(g_i) = \{\mathbf{t}_i^{(m)}\}_{m=0}^{M}$, where $m = 0$ is the original text representation and $m \ge 1$ are additional semantic descriptions. Given the refined image-view set $\hat{\mathcal{X}}(x^{\text{adv}})$ and the candidate-specific text-view set $\mathcal{T}(g_i)$, we model their alignment through optimal transport (Villani, 2009; Cuturi, 2013). The distributions are:

$$\mathbb{P}(x) = \sum_{n \in \{0\} \cup \mathcal{K}} a_n \, \delta_{\hat{\mathbf{x}}^{(n)}}, \qquad \mathbb{Q}_i(t) = \sum_{m=0}^{M} b_{i,m} \, \delta_{\mathbf{t}_i^{(m)}}, \tag{8}$$

where $\delta_{\mathbf{z}}$ denotes a unit point mass at $\mathbf{z}$, and a_n, $b_{i,m}$ are nonnegative view weights satisfying $\sum_n a_n = 1$ and $\sum_m b_{i,m} = 1$. The alignment between the attacked query and candidate g_i is then measured by the OT distance:

$$d_{\text{OT}}(\mathbb{P}(x), \mathbb{Q}_i(t); C_i) = \min_{\mathbf{T}_i \ge 0} \langle \mathbf{T}_i, C_i \rangle - \rho h(\mathbf{T}_i), \quad \text{s.t. } \mathbf{T}_i \mathbf{1} = \mathbf{a}, \ \mathbf{T}_i^{\top} \mathbf{1} = \mathbf{b}_i. \tag{9}$$

where $\rho > 0$ is the entropic regularization coefficient, $h(\mathbf{T}_i) = -\sum_{n,m} \mathbf{T}_i(n,m) \log \mathbf{T}_i(n,m)$ is the transport-plan entropy, $\mathbf{T}_i \in \mathbb{R}_+^{(|\mathcal{K}|+1)\times(M+1)}$, and $\langle \mathbf{T}_i, C_i \rangle$ denotes the Frobenius inner product. The vectors $\mathbf{a}$ and $\mathbf{b}_i$ collect the view weights in the same order as the rows and columns of $\mathbf{T}_i$, and $\mathbf{1}$ is an all-ones vector of compatible size. The transport cost between image view $n \in \{0\} \cup \mathcal{K}$ and text view $m = 0, \ldots, M$ is:

$$C_i(n,m) = 1 - \cos\left(\hat{\mathbf{x}}^{(n)}, \mathbf{t}_i^{(m)} \right). \tag{10}$$

To instantiate the weights, we use entropy-aware confidence estimation. For visual views, the entropy values defined in the previous subsection are converted into normalized weights:

$$a_n = \frac{\exp(-H_I^{(n)})}{\sum_{k \in \{0\} \cup \mathcal{K}} \exp(-H_I^{(k)})}, \qquad n \in \{0\} \cup \mathcal{K}, \tag{11}$$

so that lower-entropy visual views receive larger weights. Each descriptive view is weighted in the same way, based on how concentrated its response is over the refined image views: each $\mathbf{t}_i^{(m)}$ induces a distribution $q_{i,n}^{(m)} \propto \exp(\langle \hat{\mathbf{x}}^{(n)}, \mathbf{t}_i^{(m)} \rangle)$ over $n \in \{0\} \cup \mathcal{K}$, whose entropy is $H_{T,i}^{(m)} = -\sum_n q_{i,n}^{(m)} \log q_{i,n}^{(m)}$. We set $b_{i,m} = \exp(-H_{T,i}^{(m)}) / \sum_{\ell=0}^{M} \exp(-H_{T,i}^{(\ell)})$ for $m = 0, \ldots, M$. Given these view weights, we solve the OT problem with Sinkhorn iterations (Cuturi, 2013) and define the retrieval score as:

$$\gamma(g_i \mid x^{\text{adv}}) = -d_{\text{OT}}(\mathbb{P}(x), \mathbb{Q}_i(t); C_i). \tag{12}$$

The gallery is ranked by $\gamma(g_i \mid x^{\text{adv}})$. This formulation assigns more transport mass to semantically matched image-text views. Thus, ROTA remains stable by jointly matching multiple views when attacks make single-view similarity unreliable. Algorithm 1 in Appendix A.1 summarizes the procedure.

Table 1: Image-to-text retrieval on Flickr30K and COCO under image-query attacks. **Bold** marks the best value and underline the second best.

Attack	Method	Flickr30K						COCO					
		R@1		R@5		R@10		R@1		R@5		R@10	
		Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
PGD	jina	76.2	4.9	94.6	11.0	**98.0**	15.7	**49.7**	3.5	**75.4**	8.5	**84.8**	12.0
	Ensemble	74.1	43.1	94.3	66.9	97.1	75.6	45.0	23.2	72.8	46.6	82.5	58.1
	MTA	65.6	15.5	87.0	27.0	91.7	31.5	40.2	8.4	66.4	17.5	76.9	23.5
	ROTA	**78.5**	**66.7**	**95.1**	**82.7**	97.8	**86.9**	48.5	**42.6**	75.0	**67.3**	84.2	**76.7**
	LLM2CLIP	78.5	0.0	96.0	0.4	98.4	0.8	51.0	0.0	75.9	0.2	84.8	0.2
	Ensemble	79.0	34.8	95.5	59.3	98.3	69.3	50.5	17.1	75.2	35.5	84.3	45.5
	MTA	75.9	36.5	94.2	60.2	97.2	67.1	46.5	19.7	72.0	36.8	81.5	45.4
	ROTA	**90.5**	**64.2**	**98.4**	**80.2**	**99.4**	**84.1**	**58.1**	**40.7**	**80.4**	**62.3**	**88.5**	**70.7**
MI-FGSM	jina	76.2	2.5	94.6	6.0	**98.0**	8.2	**49.7**	1.4	**75.4**	3.8	**84.8**	5.8
	Ensemble	74.1	30.0	94.3	54.4	97.1	63.7	45.0	16.3	72.8	35.0	82.5	45.3
	MTA	65.6	10.0	87.0	17.0	91.7	22.2	40.2	5.5	66.4	11.6	76.9	15.9
	ROTA	**78.5**	**61.2**	**95.1**	**75.6**	97.8	**78.7**	48.5	**38.8**	75.0	**62.9**	84.2	**72.3**
	LLM2CLIP	78.5	0.0	96.0	0.1	98.4	0.2	51.0	0.0	75.9	0.0	84.8	0.1
	Ensemble	79.0	23.8	95.5	45.6	98.3	54.5	50.5	11.8	75.2	25.6	84.3	33.9
	MTA	75.9	15.3	94.2	29.9	97.2	37.7	46.5	6.7	72.0	14.7	81.5	19.1
	ROTA	**90.5**	**56.1**	**98.4**	**69.6**	**99.4**	**74.3**	**58.1**	**38.6**	**80.4**	**59.1**	**88.5**	**66.9**
CW	jina	76.2	11.8	94.6	23.1	**98.0**	32.5	**49.7**	7.6	**75.4**	17.7	**84.8**	23.7
	Ensemble	74.1	49.1	94.3	75.7	97.1	82.8	45.0	27.4	72.8	52.2	82.5	64.0
	MTA	65.6	32.1	87.0	52.3	91.7	59.4	40.2	18.7	66.4	36.3	76.9	45.4
	ROTA	**78.5**	**66.9**	**95.1**	**82.2**	97.8	**85.6**	48.5	**44.6**	75.0	**68.4**	84.2	**77.1**
	LLM2CLIP	78.5	4.1	96.0	11.1	98.4	15.8	51.0	1.5	75.9	3.8	84.8	5.7
	Ensemble	79.0	54.0	95.5	79.8	98.3	87.7	50.5	30.5	75.2	52.7	84.3	63.6
	MTA	75.9	46.2	94.2	70.3	97.2	78.1	46.5	22.0	72.0	41.6	81.5	50.8
	ROTA	**90.5**	**67.8**	**98.4**	**85.2**	**99.4**	**90.3**	**58.1**	**39.3**	**80.4**	**62.3**	**88.5**	**72.3**

3.4 Theoretical Analysis

OT margin stability. We analyze the final OT scoring step by comparing the clean cost matrix $C_i^\star$ with the cost matrix C_i induced by the attacked query after ROTA. Let:

$$\gamma_{\mathrm{OT}}(C) = \min_{j \neq y} [d_{\mathrm{OT}}(\mathbb{P}(x), \mathbb{Q}_j(t); C_j) - d_{\mathrm{OT}}(\mathbb{P}(x), \mathbb{Q}_y(t); C_y)]$$

denote the OT retrieval margin, where g_y is the ground-truth candidate. A larger $\gamma_{\mathrm{OT}}(C)$ means that the true candidate is separated from the nearest competing candidate by a larger distance gap.

Theorem 1. Fix the visual and textual view indices, unit-mass OT marginals, and entropic regularization coefficient across clean and perturbed inputs. Define the cost perturbation, with $\|\cdot\|_\infty$ denoting the maximum absolute entry, as:

$$\varepsilon_C := \max_i \|C_i - C_i^\star\|_\infty .$$

Then the OT margin satisfies:

$$\gamma_{\mathrm{OT}}(C) \geq \gamma_{\mathrm{OT}}(C^\star) - 2\varepsilon_C.$$

Thus, the ground-truth candidate remains top-ranked if $\gamma_{\mathrm{OT}}(C^\star) > 2\varepsilon_C$.

This bound shows that a larger clean retrieval margin allows greater changes in transport costs while preserving the top-ranked match. The proof is provided in Appendix A.2.

4 Experiments

4.1 Experimental Setup

Datasets and tasks. We evaluate robust multimodal retrieval on four benchmarks spanning standard image–text and composed retrieval: Flickr30K (Young et al., 2014), COCO (Lin et al., 2014),

Table 2: CIRR composed retrieval under image-side attacks. T embeds the reference image with its modification instruction; I embeds the target image.

Attack	Method	CIRR T2I						CIRR I2T					
		R@5		R@10		R@50		R@5		R@10		R@50	
		Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
PGD	jina	22.0	3.0	34.5	4.7	63.2	11.3	**19.6**	1.5	**35.2**	3.0	**64.5**	9.9
	Ensemble	21.1	11.8	33.8	20.9	62.8	46.1	19.1	12.0	34.2	21.4	63.9	47.6
	MTA	20.9	5.7	33.0	9.0	60.5	21.3	17.9	3.9	32.1	6.6	61.3	17.7
	ROTA	**22.1**	**17.2**	**34.8**	**29.1**	**64.1**	**57.4**	19.0	**18.5**	34.7	**34.1**	63.8	**63.8**
	LLM2CLIP	32.5	8.3	49.2	12.4	82.3	24.3	30.6	3.6	50.5	8.7	82.4	19.6
	Ensemble	33.6	21.6	50.6	34.7	**83.3**	64.6	32.3	19.1	52.4	32.1	82.7	59.2
	MTA	32.7	20.2	49.5	32.2	82.3	60.4	32.0	16.2	51.3	28.6	82.8	55.5
	ROTA	**35.4**	**26.0**	**52.0**	**39.6**	83.2	**69.8**	**33.7**	**28.0**	**53.4**	**45.3**	**84.5**	**75.0**
MI-FGSM	jina	22.0	1.6	34.5	2.7	63.2	6.3	**19.6**	0.9	**35.2**	1.9	**64.5**	5.9
	Ensemble	21.1	8.7	33.8	15.2	62.8	34.9	19.1	7.2	34.2	13.4	63.9	33.7
	MTA	20.9	5.4	33.0	8.2	60.5	17.9	17.9	2.7	32.1	4.8	61.3	13.5
	ROTA	**22.1**	**13.2**	**34.8**	**22.4**	**64.1**	**46.7**	19.0	**17.1**	34.7	**29.8**	63.8	**57.7**
	LLM2CLIP	32.5	6.0	49.2	9.2	82.3	19.4	30.6	2.0	50.5	4.8	82.4	12.6
	Ensemble	33.6	19.1	50.6	29.5	**83.3**	57.9	32.3	14.4	52.4	24.6	82.7	48.8
	MTA	32.7	17.9	49.5	28.5	82.3	51.6	32.0	11.7	51.3	21.0	82.8	42.0
	ROTA	**35.4**	**22.3**	**52.0**	**36.1**	83.2	**66.5**	**33.7**	**25.0**	**53.4**	**42.2**	**84.5**	**70.6**
CW	jina	22.0	4.1	34.5	7.5	63.2	17.9	**19.6**	3.5	**35.2**	6.9	**64.5**	18.1
	Ensemble	21.1	12.8	33.8	22.5	62.8	46.5	19.1	11.4	34.2	22.1	63.9	48.8
	MTA	20.9	9.2	33.0	14.7	60.5	30.4	17.9	6.4	32.1	12.3	61.3	30.4
	ROTA	**22.1**	**17.0**	**34.8**	**28.2**	**64.1**	**53.7**	19.0	**17.2**	34.7	**32.2**	63.8	**59.2**
	LLM2CLIP	32.5	5.7	49.2	10.3	82.3	23.5	30.6	3.4	50.5	8.7	82.4	21.2
	Ensemble	33.6	17.8	50.6	30.6	**83.3**	62.6	32.3	17.6	52.4	29.9	82.7	59.2
	MTA	32.7	19.8	49.5	32.1	82.3	61.9	32.0	16.9	51.3	30.2	82.8	58.3
	ROTA	**35.4**	**21.1**	**52.0**	**34.8**	83.2	**66.6**	**33.7**	**23.9**	**53.4**	**40.8**	**84.5**	**70.9**

CIRR (Liu et al., 2021), and FashionIQ (Wu et al., 2021), following the evaluation setup of E5-V (Jiang et al., 2024). We report both image-to-text (I2T) and text-to-image (T2I) retrieval on Flickr30K, COCO, and CIRR. For FashionIQ, we follow the standard composed retrieval protocol and report T2I results for the dress, shirt, and toptee categories. We use R@$1, 5, 10$ on Flickr30K and COCO, R@$5, 10, 50$ on CIRR, and category-wise R@$10, 50, 100$ on FashionIQ, where R@k denotes recall at rank k, namely the percentage of queries whose correct item appears in the top k results.

Models and comparisons. We use jina-embeddings-v4 (Günther et al., 2025) and LLM2CLIP (Huang et al., 2024) as the main backbones, with additional evaluations on mmE5 (Chen et al., 2025) and E5-V (Jiang et al., 2024). Each backbone is evaluated under PGD (Madry et al., 2018), MI-FGSM (Dong et al., 2018), and CW (Carlini & Wagner, 2017) attacks on image queries. We compare ROTA with Vanilla, Ensemble, and MTA (Zanella & Ben Ayed, 2024) where available. Clean retrieval provides a reference for assessing degradation under attack and subsequent recovery.

Implementation details. For attacked queries, ROTA builds multiple transformed visual embeddings and offline descriptive text views, then applies confidence-guided view selection and entropy-weighted transport scoring. Baseline configurations and implementation details are provided in Appendix A.3.

4.2 Main Results

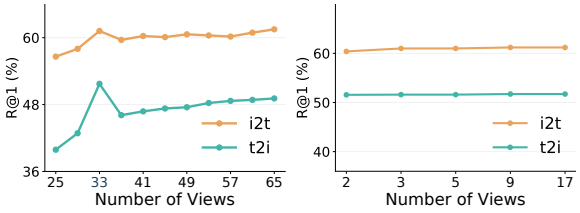
Image-to-text retrieval on Flickr30K and COCO. Table 1 reports image-to-text retrieval on Flickr30K and COCO. Under MI-FGSM, ROTA raises Flickr30K R@1 from 2.5 to 61.2 with jina and from 0.0 to 56.1 with LLM2CLIP. For jina, the gap to clean Vanilla narrows from 73.7 to 15.0 points. On COCO, ROTA reaches 72.3 R@10 with jina and 66.9 with LLM2CLIP, exceeding Ensemble (33.9) and MTA (19.1) on the latter. Text-to-image results are reported in Table 6 in Appendix B.1. To broaden the defense comparison, we report adversarial fine-tuning and additional test-time adaptation baselines in Table 11 in Appendix B.2.

Table 3: FashionIQ composed text-to-image retrieval under image-query attacks.

FashionIQ																	
		dress						shirt						toptee			
Attack	Method	R@10		R@50		R@100		R@10		R@50		R@100		R@10		R@50		R@100	
		Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
PGD	jina	**4.6**	0.2	12.7	0.2	19.3	0.7	13.8	1.7	28.0	5.0	35.8	8.1	9.5	1.0	21.9	2.6	30.4	4.4
	Ensemble	**4.6**	0.6	**13.7**	3.9	**20.0**	6.7	**15.6**	8.1	**29.7**	17.9	**37.5**	24.1	**10.3**	4.4	**23.1**	11.2	**31.2**	17.1
	MTA	4.3	0.4	12.8	1.9	18.5	3.8	12.7	5.2	26.7	11.6	35.5	17.2	8.7	3.2	21.0	8.2	29.3	11.2
	ROTA	4.5	**1.3**	13.1	**5.6**	19.8	**9.1**	13.6	**10.1**	28.4	**22.4**	36.7	**29.9**	9.7	**5.4**	21.9	**14.5**	29.9	**19.9**
	LLM2CLIP	12.2	0.1	29.1	0.5	40.7	0.8	29.1	1.2	47.4	2.8	57.6	4.6	24.4	0.7	42.9	1.8	53.4	2.8
	Ensemble	11.5	2.6	27.6	6.3	36.9	10.7	28.5	10.5	46.4	20.9	56.0	27.1	23.1	6.3	39.6	15.2	50.6	21.4
	MTA	11.2	2.1	27.2	8.0	38.2	12.7	26.9	11.5	44.5	21.0	54.8	27.5	22.1	6.5	39.8	15.0	49.6	21.3
	ROTA	**14.7**	**3.8**	**34.1**	**12.1**	**45.5**	**17.4**	**30.2**	**16.3**	**49.6**	**30.7**	**60.3**	**38.6**	**27.2**	**12.1**	**46.4**	**22.9**	**57.2**	**30.8**
MI-FGSM	jina	**4.6**	0.1	12.7	0.1	19.3	0.2	13.8	0.8	28.0	2.7	35.8	4.7	9.5	0.5	21.9	1.2	30.4	2.5
	Ensemble	**4.6**	0.3	**13.7**	2.3	**20.0**	4.1	**15.6**	6.4	**29.7**	13.3	**37.5**	19.5	**10.3**	3.1	**23.1**	8.6	**31.2**	13.2
	MTA	4.3	0.2	12.8	0.8	18.5	2.1	12.7	3.6	26.7	8.4	35.5	12.9	8.7	2.1	21.0	5.9	29.3	8.5
	ROTA	4.5	**1.0**	13.1	**3.9**	19.8	**6.4**	13.6	**8.0**	28.4	**18.4**	36.7	**25.5**	9.7	**4.2**	21.9	**11.4**	29.9	**16.2**
	LLM2CLIP	12.2	0.0	29.1	0.0	40.7	0.0	29.1	0.1	47.4	0.3	57.6	0.4	24.4	0.1	42.9	0.1	53.4	0.2
	Ensemble	11.5	0.8	27.6	2.3	36.9	4.2	28.5	5.2	46.4	11.2	56.0	16.1	23.1	2.7	39.6	6.6	50.6	10.5
	MTA	11.2	1.8	27.2	5.8	38.2	9.2	26.9	9.7	44.5	17.0	54.8	23.3	22.1	5.4	39.8	11.8	49.6	16.3
	ROTA	**14.7**	**3.1**	**34.1**	**8.6**	**45.5**	**13.8**	**30.2**	**14.2**	**49.6**	**27.0**	**60.3**	**33.8**	**27.2**	**8.8**	**46.4**	**18.5**	**57.2**	**24.6**
CW	jina	**4.6**	0.4	12.7	1.3	19.3	2.4	13.8	3.8	28.0	9.0	35.8	13.0	9.5	1.5	21.9	4.2	30.4	6.9
	Ensemble	**4.6**	**1.4**	**13.7**	**5.7**	**20.0**	9.7	**15.6**	9.2	**29.7**	20.1	**37.5**	27.2	**10.3**	5.5	**23.1**	13.7	**31.2**	**20.4**
	MTA	4.3	1.2	12.8	4.6	18.5	8.3	12.7	8.6	26.7	18.4	35.5	24.6	8.7	4.7	21.0	11.9	29.3	18.0
	ROTA	4.5	**1.4**	13.1	**5.7**	19.8	**9.8**	13.6	**9.7**	28.4	**21.5**	36.7	**28.8**	9.7	**5.6**	21.9	**14.2**	29.9	20.3
	LLM2CLIP	12.2	0.2	29.1	0.6	40.7	0.8	29.1	2.2	47.4	5.1	57.6	7.1	24.4	1.1	42.9	2.4	53.4	3.6
	Ensemble	11.5	2.0	27.6	7.3	36.9	11.4	28.5	11.3	46.4	22.7	56.0	29.7	23.1	6.8	39.6	15.9	50.6	21.2
	MTA	11.2	2.5	27.2	7.1	38.2	11.8	26.9	12.0	44.5	21.5	54.8	27.4	22.1	6.7	39.8	14.6	49.6	20.0
	ROTA	**14.7**	**3.7**	**34.1**	**10.9**	**45.5**	**16.6**	**30.2**	**16.4**	**49.6**	**30.4**	**60.3**	**38.7**	**27.2**	**10.0**	**46.4**	**22.5**	**57.2**	**29.6**

Table 4: Component ablation on Flickr30K with LLM2CLIP under two attacks.

Sec. 3.2		Sec. 3.3		R@1 (%)		
Image views	Query recon.	Text views	OT	Clean	PGD	MI-FGSM
×	×	×	×	78.5	0.0	0.0
✓	×	×	×	78.8	13.5	7.1
✓	×	✓	×	79.0	34.8	23.8
✓	✓	✓	×	88.9	58.2	48.4
✓	×	✓	✓	90.0	59.9	50.3
✓	✓	✓	✓	90.5	64.2	56.1

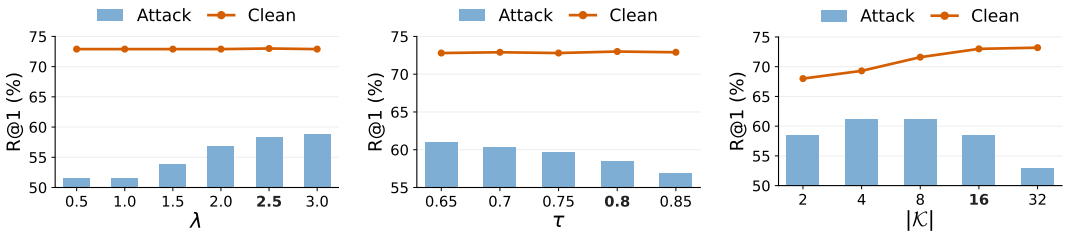


(a) Image view (b) Text view

Figure 2: Effect of image and text view counts on Flickr30K retrieval with jina under MI-FGSM.

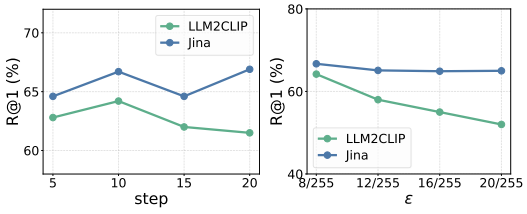
Text-to-image and image-to-text retrieval on CIRR. Table 2 reports CIRR composed retrieval in text-to-image and image-to-text settings. Under MI-FGSM, direct retrieval with jina obtains 6.3 T2I R@50 and 5.9 I2T R@50, while ROTA raises them to 46.7 and 57.7. The same trend holds for the stronger LLM2CLIP backbone: under PGD, ROTA improves T2I/I2T R@50 from 24.3/19.6 to 69.8/75.0. Across both directions, ROTA outperforms Ensemble and MTA on composed queries with attacked visual inputs.

Composed retrieval on FashionIQ. Table 3 evaluates FashionIQ composed text-to-image retrieval across dress, shirt, and toptee. Direct retrieval with attacked image evidence is highly unstable, especially on jina, where dress recall often falls below 1% and several shirt/toptee robust scores fall below 5%. ROTA achieves the best or tied-best robust recall in most reported columns; Ensemble is slightly higher for jina under CW on toptee R@100 (20.4 versus 20.3). On LLM2CLIP under PGD, it improves R@50 from 0.5 to 12.1 on dress, from 2.8 to 30.7 on shirt, and from 1.8 to 22.9 on toptee. On jina under MI-FGSM, R@100 increases from 4.7 to 25.5 on shirt and from 2.5 to 16.2



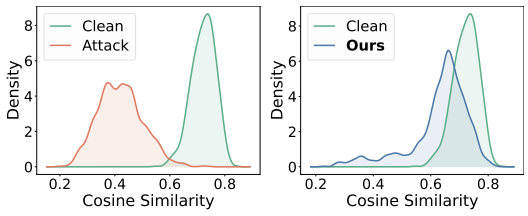
(a) Reconstruction strength λ (b) Consistency threshold τ (c) Selected views $|\mathcal{K}|$

Figure 3: Effect of reconstruction strength, consistency threshold, and selected view count on ROTA.



(a) Attack steps (b) Perturbation budget

Figure 4: Flickr30K image-to-text Recall@1 across attack steps and perturbation budgets.



(a) Attacked (b) ROTA

Figure 5: Cosine-similarity distributions before and after ROTA reconstruction.

on toptee. To evaluate ROTA on multi-frame queries, we report a preliminary video-to-text retrieval experiment in Figure 7 in Appendix B.3.

4.3 Ablation Study

Module ablation. Table 4 reports a component ablation under PGD and MI-FGSM. Image augmentation alone achieves R@1 scores of 13.5 and 7.1, respectively. Adding text views raises R@1 to 34.8 and 23.8, while further incorporating query reconstruction improves these scores to 58.2 and 48.4. Adding OT matching yields 64.2 and 56.1, corresponding to gains of 6.0 and 7.7 percentage points over reconstruction with image and text views. Without query reconstruction, image and text views with OT achieve 59.9 and 50.3, showing that reconstruction contributes an additional 4.3 and 5.8 points to the full framework. With OT, the R@1 range across the two attacks narrows from 9.8 to 8.1 points, indicating more consistent performance across the evaluated attacks.

Total number of views. We vary the total number of views, including the original view. Figure 2(a) shows initial gains from image views, followed by a plateau around 33 total views. Figure 2(b) shows smaller gains from additional text views: I2T R@1 rises from about 60.4% to 61.2%, and T2I R@1 from about 51.5% to 51.7%. Additional views therefore have diminishing returns in this setting.

Effect of hyperparameters. Figure 3 examines λ, τ, and $|\mathcal{K}|$ on Flickr30K with jina under MI-FGSM. Larger λ improves attacked R@1 from about 51.5% to nearly 59% over the tested range. Increasing τ from 0.65 to 0.85 reduces attacked recall. Here, τ switches between directional reconstruction and selected-view averaging; it does not filter views. The gains from larger λ and smaller τ are consistent with a greater contribution from directional reconstruction in this setting. Attacked recall peaks at approximately 4–8 selected views and decreases when more are retained, whereas clean recall generally increases with the selected-view count. This exposes a trade-off between clean and attacked recall. Together, Figures 2 and 3 suggest that a broad augmentation pool benefits from selectively retaining its reliable views.

Effect of attack strength. Figure 4 evaluates Flickr30K image-to-text R@1 under different attack steps and perturbation budgets. At 5, 10, 15, and 20 attack steps, LLM2CLIP obtains 62.8, 64.2, 62.0, and 61.5, respectively, while jina ranges from 64.6 to 66.9. Recall therefore varies only modestly over the tested step counts. In the separate budget sweep, LLM2CLIP R@1 decreases from 64.2 to 52.0, while jina remains around 65%–67%. The observed sensitivity is greater to perturbation magnitude than to step count in these sweeps; this does not establish attack convergence. To evaluate ROTA under adaptive attacks, we report an EOT/BPDA experiment targeting both original and augmented visual representations in Table 12 in Appendix C.1.

Similarity distribution. Figure 5 visualizes target-pair cosine similarities before and after ROTA reconstruction. Attacks reduce the mean similarity from 0.72 to 0.42, while reconstruction restores it to 0.63. Reconstruction thus recovers 70% of the observed decrease, leaving a 0.09 gap to the clean mean. Target-pair similarities do not account for competing candidates, so this mean shift alone does not establish improved rankings. To assess the separation between matched and non-matching candidates, we report a matching-margin analysis in Table 13 in Appendix C.2.

4.4 Running Time and Retrieval Visualization

Running time. Table 5 reports runtime on an NVIDIA H20 GPU, including image encoding. Vanilla takes 0.1579 s per query with 2.5 R@1, while ROTA achieves 61.2 R@1 in 0.8526 s. At nearly identical latency to Ensemble (0.8520 s), ROTA gains 31.2 R@1 points and 21.2 R@5 points. Its added cost over Vanilla is approximately 0.69 seconds per query; MTA takes 0.8619 s and reaches 10.0 R@1. We report a scalability experiment on approximately 2.5M COCO captions in Table 14 in Appendix C.3.

Table 5: Per-query runtime and I2T recall on Flickr30K with jina under MI-FGSM.

Method	Latency (s)	Retrieval	
		R@1	R@5
Vanilla	0.1579	2.5	6.0
Ensemble	0.8520	30.0	54.4
MTA	0.8619	10.0	17.0
ROTA	0.8526	**61.2**	**75.6**

Retrieval visualization. Figure 6(a) shows ROTA restoring the correct ice cream truck caption to rank 1 under attack. Without defense, the same perturbed image retrieves an unrelated caption about a person riding on a beach, despite retaining the truck's recognizable appearance. Figure 6(b) shows cross-view OT plans: Q denotes the reconstructed query, V1–V3 are transformed image views, and T0–T3 are the original candidate text and its descriptive views. Darker cells represent larger transport mass, with row and column sums of the full plans constrained by the view weights. These marginal constraints prevent a single highly similar image–text pair from absorbing all transport mass. Scores depend on both matching costs and plan entropy, so a concentrated transport plan alone does not determine the ranking.

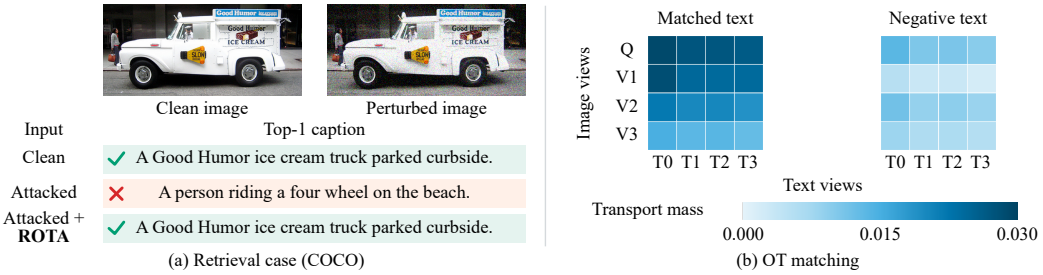


Figure 6: Illustrative visualization of retrieval rankings (a) and cross-view OT matching (b).

Limitations and Conclusion

Limitations. ROTA is a training-free defense that prioritizes plug-and-play robustness, and our current evaluation focuses on perturbation-bounded image-query attacks with representative retrieval backbones. Additional views yield diminishing gains, motivating further study of more diverse and informative views. Our video-to-text evaluation is preliminary and limited to one dataset and backbone; broader evaluation across video retrieval benchmarks remains future work.

Conclusion. We presented ROTA, a training-free framework that reconstructs visual evidence and aligns image and text views through entropy-weighted optimal transport. Selecting reliable transformed views reduces dependence on brittle single-view matching under attack. Experiments show consistent gains across standard and composed retrieval benchmarks, attacks, and backbones. The OT margin analysis establishes conditional scoring stability under bounded cost perturbations with fixed transport marginals. These findings suggest that inference-time cross-modal alignment can improve retrieval robustness without updating backbone parameters.

AI USE STATEMENT

Generative AI tools were used to assist with writing and revising the manuscript. The authors carefully reviewed and verified all AI-assisted content and take full responsibility for the accuracy and integrity of the final manuscript.

REPRODUCIBILITY STATEMENT

We describe the datasets, evaluation protocols, and baselines in Section 4. The algorithm and implementation details, including view construction and attack settings, are provided in Appendices A.1 and A.3.

REFERENCES

Akaispacelabs. Akai MovieActions 1K v1. Hugging Face Datasets, 2026.

Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*, 2022.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. mmE5: Improving multimodal multilingual embeddings via high-quality synthetic data. *arXiv preprint arXiv:2502.08468*, 2025.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. RandAugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with Cutout. arXiv preprint arXiv:1708.04552, 2017.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018.

Hao Fang, Jiawei Kong, Wenbo Yu, Bin Chen, Jiawei Li, Hao Wu, Shu-Tao Xia, and Ke Xu. One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4090–4100, 2025.

Jiyuan Fu, Zhaoyu Chen, Kaixun Jiang, Haijing Guo, Jiafeng Wang, Shuyong Gao, and Wenqiang Zhang. Improving adversarial transferability of vision-language pre-training models through collaborative multimodal interaction. *arXiv preprint arXiv:2403.10883*, 2024.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1762–1777. Association for Computational Linguistics, 2023.

Sajjad Ghiasvand, Haniyeh Ehsani Oskouie, Mahnoosh Alizadeh, and Ramtin Pedarsani. Few-shot adversarial low-rank fine-tuning of vision-language models. *arXiv preprint arXiv:2505.15130*, 2025.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, et al. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pp. 531–550, 2025.

Zongbo Han, Jialong Yang, Guangyu Wang, Junfan Li, Qianli Xu, Mike Zheng Shou, and Changqing Zhang. DOTA: Distributional test-time adaptation of vision-language models. In *Advances in Neural Information Processing Systems*, 2025.

Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Xiyang Dai, Dongdong Chen, Chong Luo, and Lili Qiu. LLM2CLIP: Powerful language model unlocks richer visual representation. arXiv preprint arXiv:2411.04997v3, 2024.

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.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-V: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14162–14171, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Zheng Li, Lijia Si, Caili Guo, Yang Yang, and Qiushi Cao. Data augmentation for text-based person retrieval using large language models. *arXiv preprint arXiv:2405.11971*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2125–2134, 2021.

Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 102–111, 2023.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022.

Zhengxin Pan, Haishuai Wang, Fangyu Wu, Bailing Zhang, Jiajun Bu, and Hongyang Chen. Learning image-text matching with optimal partial transport. *arXiv preprint arXiv:2603.14349*, 2026.

Juan C Pérez, Motasem Alfarra, Guillaume Jeanneret, Laura Rueda, Ali Thabet, Bernard Ghanem, and Pablo Arbeláez. Enhancing adversarial robustness via test-time transformation ensembling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 81–91, 2021.

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.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv preprint arXiv:2402.12336*, 2024.

Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-TPT: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29958–29967, 2025.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14274–14289, 2022.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Pu Tian, Yalong Wu, Sudheera Balagam, Kaushik Kona, Bhuvanya Sirigineedi, Tanmai Bhashyam, Manideep Nalumachu, Varshini Kotha, and Justyn Pibil. On robustness of vision-language models under adversarial perturbations. In *2025 IEEE International Conference on Future Machine Learning and Data Science (FMLDS)*, pp. 56–61. IEEE, 2025.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin, Heidelberg, 2009.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: a systematic review of methods and future directions. *Proceedings of the IEEE*, 112(11):1716–1754, 2024.

Xin Wang, Kai Chen, Jiaming Zhang, Jingjing Chen, and Xingjun Ma. TAPT: Test-time adversarial prompt tuning for robust inference in vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19910–19920, 2025.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11307–11317, 2021.

Peng Xie, Yequan Bie, Jianda Mao, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. Chain of attack: On the robustness of vision-language models against transfer-based adversarial attacks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14679–14689, 2025.

Jingning Xu, Haochen Luo, and Chen Liu. PDA: Text-augmented defense framework for robust vision-language models against adversarial image attacks. *arXiv preprint arXiv:2604.01010*, 2026.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23783–23793, 2024.

Yi Zeng, Han Qiu, Gerard Memmi, and Meikang Qiu. A data augmentation-based defense method against adversarial attacks in neural networks. In *International Conference on Algorithms and Architectures for Parallel Processing*, pp. 274–289. Springer, 2020.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Guanqi Zhan, Yuanpei Liu, Kai Han, Weidi Xie, and Andrew Zisserman. ELIP: Enhanced visual-language foundation models for image retrieval. In *2025 International Conference on Content-Based Multimedia Indexing (CBMI)*, pp. 1–8. IEEE, 2025.

Kun Zhang, Jingyu Li, Zhe Li, Jingjing Zhang, Fan Li, Yandong Liu, Rui Yan, Zihang Jiang, Nan Chen, Lei Zhang, et al. Composed multi-modal retrieval: A survey of approaches and applications. *arXiv preprint arXiv:2503.01334*, 2025.

Tingwei Zhang, Rishi Jha, Eugene Bagdasaryan, and Vitaly Shmatikov. Adversarial illusions in Multi-Modal embeddings. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 3009–3025. USENIX Association, 2024a.

Tingwei Zhang, Fnu Suya, Rishi Jha, Collin Zhang, and Vitaly Shmatikov. Adversarial hubness in multi-modal retrieval. *arXiv preprint arXiv:2412.14113*, 2024b.

Mo Zhou, Le Wang, Zhenxing Niu, Qilin Zhang, Nanning Zheng, and Gang Hua. Adversarial attack and defense in deep ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5306–5324, 2024a.

Wanqi Zhou, Shuanghao Bai, Danilo P Mandic, Qibin Zhao, and Badong Chen. Revisiting the adversarial robustness of vision language models: a multimodal perspective. *arXiv preprint arXiv:2404.19287*, 2024b.

Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. AdvCLIP: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 6311–6320, 2023.

Yuhan Zhu, Yuyang Ji, Zhiyu Zhao, Gangshan Wu, and Limin Wang. AWT: Transferring vision-language models via augmentation, weighting, and transportation. *Advances in Neural Information Processing Systems*, 37:25561–25591, 2024.

CONTENTS

1 Introduction **1**

2 Related Work **2**

3 Methodology **3**

3.1 Preliminaries 3

3.2 Robust Query Representation 3

3.3 Cross-Modal Transport Matching 4

3.4 Theoretical Analysis 5

4 Experiments **5**

4.1 Experimental Setup 5

4.2 Main Results 6

4.3 Ablation Study 8

4.4 Running Time and Retrieval Visualization 9

A Method Details **15**

A.1 ROTA Algorithm 15

A.2 Proofs for Theoretical Analysis 15

A.3 Implementation Details 16

B Extended Retrieval Evaluation **16**

B.1 Additional Retrieval Results 16

B.2 Additional Adaptation Baselines 17

B.3 Preliminary Video-to-Text Extension 17

C Supplementary Analyses **22**

C.1 Adaptive Attack Evaluation 22

C.2 Cross-Modal Matching Margins 22

C.3 Large-Scale Retrieval Efficiency 23

C.4 Retrieval Examples 24

A METHOD DETAILS

A.1 ROTA ALGORITHM

Algorithm 1 ROTA inference

Require: Adversarial query x^{adv}, gallery $\mathcal{G}$, transforms $\mathcal{A}$, text views $\mathcal{T}(g_i)$, τ, λ.
Ensure: Ranked list $\mathcal{R}$.
1: Encode original and transformed views $\{\mathbf{x}^{(n)}\}_{n=0}^{N}$.
2: Select reliable view indices $\mathcal{K} \subseteq \{1, \dots, N\}$ using visual uncertainty $\{H_I^{(n)}\}$.
3: Compute displacement $\bar{\Delta}$ and coherence η from $\mathcal{K}$.
4: Set $\hat{\mathbf{x}}^{(0)} \leftarrow \mathbf{x}^{(0)} + \lambda\bar{\Delta}$ if $\eta > \tau$; otherwise average views in $\mathcal{K}$.
5: Set $\hat{\mathbf{x}}^{(n)} \leftarrow \mathbf{x}^{(n)}$ for $n \in \mathcal{K}$ and form $\hat{\mathcal{X}}$.
6: Form $\mathbb{P}$ over $n \in \{0\} \cup \mathcal{K}$ with normalized weights a_n.
7: **for** each $g_i \in \mathcal{G}$ **do**
8: Form $\mathbb{Q}_i$ over text views $m = 0, \dots, M$ with normalized weights $b_{i,m}$.
9: Compute $C_i(n, m) = 1 - \cos(\hat{\mathbf{x}}^{(n)}, \mathbf{t}_i^{(m)})$.
10: Score $\gamma(g_i \mid x^{\mathrm{adv}}) = -d_{\mathrm{OT}}(\mathbb{P}, \mathbb{Q}_i; C_i)$ with Sinkhorn.
11: **end for**
12: **return** candidates sorted by $\gamma(g_i \mid x^{\mathrm{adv}})$.

A.2 PROOFS FOR THEORETICAL ANALYSIS

Proof of Theorem 1. We use the notation of Theorem 1. With fixed view indices, unit-mass marginals, and regularization coefficient ρ, the clean and perturbed OT problems share the same feasible set and entropy term. Every feasible $\mathbf{T} \geq 0$ satisfies $\sum_{n,m} \mathbf{T}(n, m) = 1$, giving:

$$|\langle \mathbf{T}, C_i - C_i^\star \rangle| \leq \|C_i - C_i^\star\|_\infty \sum_{n,m} \mathbf{T}(n, m) \leq \varepsilon_C.$$

The objectives thus differ pointwise by at most ε_C. Taking minima gives:

$$|d_{\mathrm{OT}}(\mathbb{P}(x), \mathbb{Q}_i(t); C_i) - d_{\mathrm{OT}}(\mathbb{P}(x), \mathbb{Q}_i(t); C_i^\star)| \leq \varepsilon_C.$$

For each $j \neq y$, the competing distance can decrease by at most ε_C, while the ground-truth distance can increase by at most ε_C. Taking the minimum of their differences yields:

$$\gamma_{\mathrm{OT}}(C) \geq \gamma_{\mathrm{OT}}(C^\star) - 2\varepsilon_C.$$

Hence $\gamma_{\mathrm{OT}}(C^\star) > 2\varepsilon_C$ preserves the ground-truth candidate's top rank.

Connection to image-space perturbations. For image-to-text retrieval under the fixed configuration of Theorem 1, also fix the unit-norm candidate text views, augmentation realizations, and reconstruction branch. Let $\widehat{\mathbf{z}}_n(\mathbf{x})$ denote the n-th normalized visual representation used by OT. Assume these complete mappings, including augmentation, encoding, reconstruction, and normalization, are all L-Lipschitz in ℓ_2 on the admissible ℓ_∞ ball of radius ϵ around the clean image $\mathbf{x} \in \mathbb{R}^D$. For $\|\boldsymbol{\delta}\|_\infty \leq \epsilon$, unit-norm text views and Cauchy–Schwarz give:

$$\begin{aligned} |C_i(n, m) - C_i^\star(n, m)| &\leq \|\widehat{\mathbf{z}}_n(\mathbf{x} + \boldsymbol{\delta}) - \widehat{\mathbf{z}}_n(\mathbf{x})\|_2 \\ &\leq L\|\boldsymbol{\delta}\|_2 \leq L\sqrt{D}\,\epsilon. \end{aligned}$$

Thus $\varepsilon_C \leq L\sqrt{D}\,\epsilon$, and Theorem 1 yields:

$$\gamma_{\mathrm{OT}}(C) \geq \gamma_{\mathrm{OT}}(C^\star) - 2L\sqrt{D}\,\epsilon.$$

The top rank is preserved if $\gamma_{\mathrm{OT}}(C^\star) > 2L\sqrt{D}\,\epsilon$. This connection is conditional: transforms such as quantization may be discontinuous even with fixed randomness. We do not estimate a certified bound on L or certify the full adaptive ROTA pipeline.

A.3 Implementation Details

View configuration. For Flickr30K with LLM2CLIP, the reference configuration uses 32 image augmentations and 8 text augmentations, in addition to the original views.

Entropy computation and weighting. In image-to-text retrieval, visual-view entropy is computed over the full gallery using the base text embeddings. Text-view entropy is computed separately for each candidate over the refined image-view set, including the reconstructed query. Text weighting favors concentrated responses as a heuristic for more specific visual correspondences.

Retrieval parameters. We set the entropic regularization coefficient to $\rho = 1$ and run Sinkhorn for at most 1,000 iterations. Convergence is checked every 20 iterations with a stopping tolerance of 10^{-4}.

Baseline configurations. Vanilla scores the attacked query directly with the backbone's similarity function. Ensemble averages similarities over augmented image and text views. MTA (Zanella & Ben Ayed, 2024) provides an iterative test-time adaptation baseline where available. Gallery splits and candidate IDs are fixed across clean and attacked evaluations for each backbone.

Descriptive text views. E5-V generates descriptive text views with task-specific inputs. For Flickr30K and COCO, each candidate caption is inserted into the text-only continuation template `<sent>\n Summary above sentence in one word:`. CIRR and FashionIQ use image-conditioned prompts containing the query modification text; FashionIQ also specifies the garment category. E5-V may return a short continuation despite the one-word request. We concatenate each continuation with its prompt and re-encode it with the evaluated backbone's text encoder. The original text is encoded independently and retained as the base view.

Text-view caching. For Flickr30K and COCO, fixed candidate-text descriptions can be encoded once and reused across image queries. CIRR and FashionIQ prompts depend on the composed query, making their precomputation query-specific. Text-view weights and OT scores are computed for each query from the refined image views.

Image transforms. We sample image views with a RandAugment-style (Cubuk et al., 2020) transformation pool: shearing, rotation, translation, auto-contrast, equalization, inversion, solarization, posterization, color, contrast, brightness, sharpness, and cutout (DeVries & Taylor, 2017). Randomly sampled operations are applied sequentially with a shared magnitude. Directional geometric operations are reversed with probability 0.5.

Attack settings. All backbones and datasets share the default attack settings: a perturbation budget of $\epsilon = 8/255$, a step size of $8/255$, and 10 attack iterations, with MI-FGSM momentum decay 1.0 and CW parameters $C = 100$, $\kappa = 0$.

B Extended Retrieval Evaluation

B.1 Additional Retrieval Results

We evaluate additional retrieval directions and backbones. Recall values are percentages; Clean and Rob. denote clean and attacked settings.

Text-to-image retrieval with jina and LLM2CLIP. Table 6 reports T2I retrieval on Flickr30K and COCO with jina and LLM2CLIP. Under PGD, ROTA raises Flickr30K R@10 from 14.1 for direct retrieval to 82.2 with jina; under MI-FGSM, it raises COCO R@10 from 0.0 to 45.2 with LLM2CLIP.

Image-to-text retrieval with mmE5 and E5-V. Table 7 evaluates I2T retrieval with mmE5 and E5-V. On COCO under PGD, ROTA raises mmE5 R@10 from 61.2 for direct retrieval to 80.6. On Flickr30K under MI-FGSM, E5-V R@10 increases from 0.8 to 53.5.

Text-to-image retrieval with mmE5 and E5-V. Table 8 evaluates T2I retrieval with mmE5 and E5-V. Under PGD, ROTA raises mmE5 COCO R@10 from 54.6 for direct retrieval to 76.9; under MI-FGSM, it raises E5-V Flickr30K R@10 from 1.6 to 49.2.

Composed retrieval on CIRR. Table 9 evaluates mmE5 and E5-V on CIRR in both retrieval directions. Under MI-FGSM, ROTA raises T2I/I2T R@50 from 73.3/70.3 for direct retrieval to

Table 6: Text-to-image retrieval on Flickr30K and COCO with textual queries and adversarially perturbed image candidates using jina and LLM2CLIP.

Attack	Method	Flickr30K						COCO					
		R@1		R@5		R@10		R@1		R@5		R@10	
		Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
PGD	jina	72.8	3.4	91.4	9.8	95.0	14.1	46.1	3.0	72.0	8.1	81.3	12.2
	Ensemble	71.0	39.9	91.8	66.1	95.7	75.8	44.1	22.5	70.1	44.1	79.5	55.7
	MTA	61.9	15.5	84.0	26.1	90.0	31.5	40.6	10.7	65.3	20.9	74.9	26.9
	ROTA	**75.7**	**54.2**	**93.0**	**76.0**	**96.3**	**82.2**	**47.7**	**35.1**	**73.6**	**61.0**	**82.7**	**71.2**
	LLM2CLIP	59.1	0.0	82.6	0.1	88.9	0.2	34.8	0.0	60.2	0.0	70.4	0.1
	Ensemble	61.5	29.1	84.4	54.2	90.1	64.7	36.0	13.8	61.7	30.7	72.0	40.3
	MTA	56.6	26.7	80.9	45.0	88.0	53.5	32.8	13.9	58.1	28.5	68.7	36.2
	ROTA	**75.3**	**39.3**	**93.1**	**58.8**	**96.1**	**65.7**	**46.4**	**23.0**	**71.4**	**43.4**	**79.9**	**53.4**
MI-FGSM	jina	72.8	1.6	91.4	4.3	95.0	6.8	46.1	1.1	72.0	3.6	81.3	5.8
	Ensemble	71.0	29.9	91.8	54.3	95.7	64.2	44.1	16.6	70.1	35.6	79.5	46.0
	MTA	61.9	11.2	84.0	19.4	90.0	23.8	40.6	7.3	65.3	15.0	74.9	19.4
	ROTA	**75.7**	**51.7**	**93.0**	**65.9**	**96.3**	**72.9**	**47.7**	**28.9**	**73.6**	**52.2**	**82.7**	**62.7**
	LLM2CLIP	59.1	0.0	82.6	0.0	88.9	0.0	34.8	0.0	60.2	0.0	70.4	0.0
	Ensemble	61.5	21.8	84.4	41.9	90.1	51.8	36.0	9.5	61.7	22.4	72.0	30.6
	MTA	56.6	12.6	80.9	22.3	88.0	27.1	32.8	6.1	58.1	12.7	68.7	16.2
	ROTA	**75.3**	**31.7**	**93.1**	**46.7**	**96.1**	**52.9**	**46.4**	**18.1**	**71.4**	**35.9**	**79.9**	**45.2**
CW	jina	72.8	8.6	91.4	21.5	95.0	29.1	46.1	5.1	72.0	13.7	81.3	20.3
	Ensemble	71.0	45.7	91.8	73.4	95.7	82.0	44.1	27.1	70.1	51.4	79.5	62.6
	MTA	61.9	32.4	84.0	50.4	90.0	57.9	40.6	20.3	65.3	37.1	74.9	45.3
	ROTA	**75.7**	**56.3**	**93.0**	**79.0**	**96.3**	**85.4**	**47.7**	**37.0**	**73.6**	**62.8**	**82.7**	**73.3**
	LLM2CLIP	59.1	2.7	82.6	7.1	88.9	10.8	34.8	0.9	60.2	2.7	70.4	4.3
	Ensemble	61.5	47.9	84.4	73.5	90.1	81.9	36.0	23.7	61.7	45.2	72.0	56.2
	MTA	56.6	32.6	80.9	53.7	88.0	61.8	32.8	15.9	58.1	31.8	68.7	40.1
	ROTA	**75.3**	**51.9**	**93.1**	**75.6**	**96.1**	**82.4**	**46.4**	**27.9**	**71.4**	**50.9**	**79.9**	**61.7**

91.0/92.2 with mmE5, and from 14.0/11.1 to 50.8/40.3 with E5-V. Ensemble remains stronger on E5-V under CW at all three I2T ranks.

Composed retrieval on FashionIQ. Table 10 reports FashionIQ retrieval across three garment categories. Under MI-FGSM, ROTA raises mmE5 R@50 from 16.6/39.5/31.1 for direct retrieval to 36.1/55.2/51.5 for dress/shirt/toptee. Gains also hold for E5-V, although Ensemble achieves higher shirt R@100 under MI-FGSM and CW.

B.2 Additional Adaptation Baselines

Comparison with adaptation baselines. Table 11 compares adversarial fine-tuning and test-time adaptation on Flickr30K I2T retrieval with LLM2CLIP. TeCoA is the strongest additional baseline under all three attacks, reaching 35.5, 33.2, and 40.0 R@1 under PGD, MI-FGSM, and CW, respectively. ROTA achieves 64.2, 56.1, and 67.8 without parameter updates, exceeding the reported alternatives in each setting.

B.3 Preliminary Video-to-Text Extension

Multi-frame query retrieval. Figure 7 reports video-to-text recall on the 1,000-clip Akai MovieActions dataset (Akaispacelabs, 2026). We use frozen CLIP ViT-B/32 with eight uniformly sampled frames per video and original plus offline-generated candidate text views. Under PGD, ROTA raises R@1 from 1.2 to 23.0 and R@10 from 6.0 to 51.5, while clean R@1 changes from 20.4 to 21.4. ROTA gains extend across recall cutoffs in this preliminary multi-frame setting.

Table 7: Image-to-text retrieval on Flickr30K and COCO with mmE5 and E5-V.

Attack	Method	Flickr30K						COCO					
		R@1		R@5		R@10		R@1		R@5		R@10	
		Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
PGD	mmE5	80.1	44.0	**96.8**	69.3	**99.5**	78.1	**51.3**	27.3	**77.8**	50.2	**86.6**	61.2
	Ensemble	76.2	67.7	96.0	90.7	99.0	96.1	48.8	41.2	75.8	68.6	84.6	79.2
	MTA	71.2	63.0	92.1	85.7	97.2	91.8	42.6	36.0	69.7	63.1	79.4	74.0
	ROTA	**80.7**	**74.2**	96.7	**93.5**	98.9	**97.0**	48.2	**42.6**	75.6	**70.4**	84.4	**80.6**
	E5-V	**89.1**	0.9	**98.5**	2.7	**99.4**	5.1	**62.1**	4.6	**83.4**	10.4	**89.7**	14.8
	Ensemble	85.7	13.7	97.9	28.4	**99.4**	38.9	59.1	14.0	81.9	29.2	89.1	38.2
	MTA	74.2	8.7	93.2	20.7	96.7	27.7	49.0	10.1	74.1	21.4	82.5	29.3
	ROTA	87.2	**34.9**	97.4	**53.8**	98.8	**61.0**	60.4	**27.5**	81.9	**46.7**	88.6	**55.9**
MI-FGSM	mmE5	80.1	22.5	**96.8**	40.2	**99.5**	49.5	**51.3**	26.0	**77.8**	48.7	**86.6**	60.0
	Ensemble	76.2	57.1	96.0	83.7	99.0	89.3	48.8	39.8	75.8	66.6	84.6	77.3
	MTA	71.2	46.9	92.1	71.4	97.2	79.8	42.6	35.3	69.7	61.4	79.4	72.6
	ROTA	**80.7**	**66.7**	96.7	**88.3**	98.9	**93.1**	48.2	**41.6**	75.6	**68.6**	84.4	**79.2**
	E5-V	**89.1**	0.2	**98.5**	0.4	**99.4**	0.8	**62.1**	1.0	**83.4**	2.6	**89.7**	4.3
	Ensemble	85.7	6.5	97.9	19.6	**99.4**	25.8	59.1	6.1	81.9	14.9	89.1	21.7
	MTA	74.2	4.7	93.2	10.0	96.7	12.9	49.0	3.7	74.1	8.9	82.5	13.1
	ROTA	87.2	**29.9**	97.4	**45.7**	98.8	**53.5**	60.4	**17.8**	81.9	**34.2**	88.6	**44.0**
CW	mmE5	80.1	25.2	**96.8**	45.2	**99.5**	55.3	**51.3**	28.6	**77.8**	52.3	**86.6**	63.6
	Ensemble	76.2	59.1	96.0	84.2	99.0	90.4	48.8	40.3	75.8	68.0	84.6	78.7
	MTA	71.2	49.8	92.1	75.7	97.2	84.0	42.6	36.6	69.7	63.1	79.4	73.9
	ROTA	**80.7**	**65.8**	96.7	**88.9**	98.9	**93.8**	48.2	**41.9**	75.6	**70.0**	84.4	**79.8**
	E5-V	**89.1**	47.8	**98.5**	62.6	**99.4**	68.6	**62.1**	25.0	**83.4**	37.3	**89.7**	41.8
	Ensemble	85.7	67.1	97.9	88.1	**99.4**	93.5	59.1	38.7	81.9	62.5	89.1	73.4
	MTA	74.2	53.6	93.2	74.9	96.7	81.7	49.0	23.5	74.1	38.7	82.5	45.5
	ROTA	87.2	**73.8**	97.4	**90.9**	98.8	**94.1**	60.4	**52.1**	81.9	**73.6**	88.6	**81.4**

Table 8: Text-to-image retrieval on Flickr30K and COCO with mmE5 and E5-V.

Attack	Method	Flickr30K						COCO					
		R@1		R@5		R@10		R@1		R@5		R@10	
		Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
PGD	mmE5	**72.7**	36.9	**91.5**	61.7	**95.5**	71.3	**46.4**	22.2	71.7	43.8	80.7	54.6
	Ensemble	71.9	62.3	91.2	85.3	95.2	91.0	45.9	37.3	71.4	63.2	81.0	73.7
	MTA	66.0	57.8	87.9	80.9	93.1	87.5	41.0	34.9	66.1	59.0	76.1	69.2
	ROTA	72.6	**67.2**	91.4	**88.1**	95.4	**92.8**	45.9	**40.7**	**71.9**	**66.7**	**81.2**	**76.9**
	E5-V	**81.2**	1.8	**95.6**	4.8	**97.7**	7.2	52.1	3.8	76.4	10.3	84.7	14.8
	Ensemble	78.7	14.2	95.1	32.3	97.6	42.5	50.0	11.3	75.1	26.1	83.8	35.4
	MTA	67.6	11.3	89.4	23.9	94.1	31.2	42.6	9.8	67.1	22.3	76.6	30.3
	ROTA	81.1	**24.0**	**95.6**	**45.8**	**97.7**	**54.7**	**52.4**	**16.0**	**76.7**	**33.9**	**84.9**	**43.7**
MI-FGSM	mmE5	**72.7**	18.3	**91.5**	36.6	**95.5**	45.3	**46.4**	21.2	71.7	41.8	80.7	52.5
	Ensemble	71.9	51.9	91.2	76.4	95.2	83.6	45.9	35.5	71.4	61.1	81.0	72.0
	MTA	66.0	42.7	87.9	65.8	93.1	73.8	41.0	33.4	66.1	57.5	76.1	68.1
	ROTA	72.6	**59.0**	91.4	**81.8**	95.4	**88.1**	45.9	**38.9**	**71.9**	**65.1**	**81.2**	**75.6**
	E5-V	**81.2**	0.4	**95.6**	1.1	**97.7**	1.6	52.1	1.0	76.4	3.1	84.7	5.3
	Ensemble	78.7	9.6	95.1	23.8	97.6	32.9	50.0	6.2	75.1	15.9	83.8	22.9
	MTA	67.6	6.6	89.4	13.6	94.1	17.8	42.6	4.5	67.1	11.4	76.6	16.4
	ROTA	81.1	**19.8**	**95.6**	**40.3**	**97.7**	**49.2**	**52.4**	**10.2**	**76.7**	**23.6**	**84.9**	**32.0**
CW	mmE5	**72.7**	21.6	**91.5**	41.4	**95.5**	50.3	**46.4**	23.6	71.7	45.2	80.7	56.1
	Ensemble	71.9	53.4	91.2	77.6	95.2	85.0	45.9	36.6	71.4	62.6	81.0	73.1
	MTA	66.0	47.2	87.9	70.6	93.1	79.4	41.0	35.0	66.1	59.8	76.1	70.3
	ROTA	72.6	**60.7**	91.4	**83.0**	95.4	**88.8**	45.9	**39.6**	**71.9**	**66.2**	**81.2**	**76.4**
	E5-V	**81.2**	44.4	**95.6**	60.8	**97.7**	67.1	52.1	23.8	76.4	35.5	84.7	39.7
	Ensemble	78.7	61.1	95.1	83.2	97.6	89.0	50.0	30.1	75.1	50.6	83.8	60.1
	MTA	67.6	50.1	89.4	72.5	94.1	79.6	42.6	22.7	67.1	37.1	76.6	43.0
	ROTA	81.1	**67.0**	**95.6**	**88.1**	**97.7**	**92.7**	**52.4**	**38.8**	**76.7**	**63.1**	**84.9**	**72.9**

Table 9: CIRR composed retrieval in both directions with mmE5 and E5-V.

Attack	Method	CIRR T2I						CIRR I2T					
		R@5		R@10		R@50		R@5		R@10		R@50	
		Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
PGD	mmE5	62.1	34.0	75.8	49.4	92.9	77.6	55.8	32.0	71.0	45.9	91.7	74.4
	Ensemble	61.5	52.3	75.4	68.1	**93.4**	90.2	54.9	46.8	69.9	63.1	91.1	87.6
	MTA	54.5	46.4	69.4	60.7	90.9	84.9	49.8	40.5	65.8	55.9	87.7	81.3
	ROTA	**62.2**	**55.9**	**75.9**	**71.6**	**93.4**	**91.8**	**59.1**	**57.2**	**74.6**	**72.1**	**93.8**	**92.6**
	E5-V	63.6	14.6	77.1	21.7	93.7	46.6	55.7	12.8	71.4	19.9	**92.8**	44.0
	Ensemble	**64.7**	35.5	**78.3**	50.2	**94.3**	79.2	55.2	33.6	**72.4**	47.6	92.4	76.6
	MTA	64.4	20.0	77.5	28.0	94.0	50.9	55.2	16.8	71.4	25.7	91.9	52.4
	ROTA	63.6	**45.0**	78.1	**59.9**	92.8	**85.8**	**55.9**	**36.7**	71.3	**50.2**	92.4	**80.1**
MI-FGSM	mmE5	62.1	31.8	75.8	45.5	92.9	73.3	55.8	28.9	71.0	42.7	91.7	70.3
	Ensemble	61.5	50.4	75.4	65.5	**93.4**	89.5	54.9	44.7	69.9	61.0	91.1	86.3
	MTA	54.5	44.8	69.4	58.8	90.9	84.3	49.8	39.0	65.8	54.1	87.7	80.5
	ROTA	**62.2**	**54.5**	**75.9**	**70.2**	**93.4**	**91.0**	**59.1**	**56.1**	**74.6**	**71.9**	**93.8**	**92.2**
	E5-V	63.6	2.6	77.1	4.8	93.7	14.0	55.7	1.4	71.4	3.0	**92.8**	11.1
	Ensemble	**64.7**	10.7	**78.3**	17.8	**94.3**	41.1	55.2	7.6	**72.4**	12.2	92.4	30.7
	MTA	64.4	6.2	77.5	9.2	94.0	20.9	55.2	2.6	71.4	5.3	91.9	16.1
	ROTA	63.6	**16.9**	78.1	**25.6**	92.8	**50.8**	**55.9**	**11.2**	71.3	**18.0**	92.4	**40.3**
CW	mmE5	62.1	33.5	75.8	47.4	92.9	76.1	55.8	30.8	71.0	44.0	91.7	73.5
	Ensemble	61.5	50.0	75.4	65.9	**93.4**	89.3	54.9	45.7	69.9	61.2	91.1	86.2
	MTA	54.5	45.6	69.4	60.9	90.9	85.4	49.8	40.8	65.8	55.2	87.7	81.2
	ROTA	**62.2**	**54.0**	**75.9**	**70.1**	**93.4**	**90.9**	**59.1**	**55.9**	**74.6**	**71.0**	**93.8**	**91.9**
	E5-V	63.6	46.5	77.1	55.8	93.7	71.8	55.7	38.1	71.4	50.7	**92.8**	73.3
	Ensemble	**64.7**	52.3	**78.3**	65.3	**94.3**	87.0	55.2	**46.0**	**72.4**	**61.9**	92.4	**86.5**
	MTA	64.4	43.9	77.5	54.6	94.0	73.6	55.2	36.1	71.4	49.8	91.9	75.2
	ROTA	63.6	**54.9**	78.1	**68.8**	92.8	**90.6**	**55.9**	43.8	71.3	61.0	92.4	85.6

Table 10: FashionIQ composed text-to-image retrieval with mmE5 and E5-V.

		FashionIQ															
		dress						shirt						toptee			
Attack	Method	R@10		R@50		R@100		R@10		R@50		R@100		R@10		R@50		R@100	
		Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.	Clean	Rob.
PGD	mmE5	28.8	6.8	52.2	17.4	63.0	24.6	**43.5**	24.7	62.5	39.7	70.7	49.0	41.0	17.4	62.8	32.0	73.6	41.6
	Ensemble	22.9	13.6	47.5	31.3	58.9	43.0	40.4	32.4	60.2	52.4	69.3	61.8	37.2	26.7	59.1	46.6	70.7	56.9
	MTA	21.7	11.4	40.0	26.3	49.9	37.6	36.1	29.6	55.6	48.4	65.5	58.4	33.1	23.9	53.1	41.7	63.9	51.8
	ROTA	**29.3**	**17.7**	**52.9**	**39.5**	**63.9**	**51.1**	42.5	**36.7**	**62.6**	**56.2**	**71.3**	**65.3**	**41.2**	**31.1**	**64.2**	**53.1**	**74.5**	**63.8**
	E5-V	**31.8**	4.0	**56.8**	13.3	**68.7**	21.6	**49.2**	17.6	**72.0**	35.9	**81.3**	47.5	**47.2**	13.6	**73.0**	32.0	**82.1**	42.3
	Ensemble	28.9	10.4	53.9	27.7	65.6	38.6	49.0	31.4	70.9	53.6	79.5	64.9	45.1	26.2	70.5	48.9	79.4	60.4
	MTA	20.2	7.4	38.3	17.9	48.0	25.8	33.4	20.2	53.1	36.5	63.1	47.8	31.0	16.1	51.7	31.7	62.7	42.1
	ROTA	28.3	**11.2**	51.7	**30.4**	63.5	**41.7**	46.3	**32.6**	67.4	**54.7**	76.2	**65.3**	42.9	**26.9**	68.0	**50.6**	76.8	**63.0**
MI-FGSM	mmE5	28.8	6.1	52.2	16.6	63.0	23.6	**43.5**	23.7	62.5	39.5	70.7	48.2	41.0	17.4	62.8	31.1	73.6	41.2
	Ensemble	22.9	12.8	47.5	29.4	58.9	41.6	40.4	31.9	60.2	51.2	69.3	60.4	37.2	24.9	59.1	44.4	70.7	55.7
	MTA	21.7	11.0	40.0	25.5	49.9	35.2	36.1	29.0	55.6	47.1	65.5	57.7	33.1	22.9	53.1	40.2	63.9	50.4
	ROTA	**29.3**	**16.6**	**52.9**	**36.1**	**63.9**	**47.5**	42.5	**36.1**	**62.6**	**55.2**	**71.3**	**63.6**	**41.2**	**29.8**	**64.2**	**51.5**	**74.5**	**61.5**
	E5-V	**31.8**	1.3	**56.8**	3.9	**68.7**	8.0	**49.2**	6.0	**72.0**	15.5	**81.3**	22.9	**47.2**	4.2	**73.0**	13.3	**82.1**	20.4
	Ensemble	28.9	3.8	53.9	12.4	65.6	19.9	49.0	16.4	70.9	31.8	79.5	**42.2**	45.1	11.6	70.5	27.2	79.4	37.5
	MTA	20.2	2.9	38.3	7.8	48.0	12.6	33.4	10.3	53.1	20.8	63.1	29.6	31.0	7.5	51.7	16.6	62.7	24.7
	ROTA	28.3	**5.2**	51.7	**14.0**	63.5	**22.0**	46.3	**17.8**	67.4	**32.7**	76.2	42.1	42.9	**11.9**	68.0	**28.9**	76.8	**39.6**
CW	mmE5	28.8	7.1	52.2	17.8	63.0	26.3	**43.5**	25.1	62.5	41.4	70.7	50.4	41.0	18.9	62.8	33.7	73.6	43.7
	Ensemble	22.9	13.2	47.5	31.3	58.9	43.2	40.4	32.1	60.2	51.9	69.3	61.8	37.2	26.1	59.1	45.9	70.7	57.0
	MTA	21.7	12.1	40.0	27.6	49.9	38.5	36.1	30.3	55.6	48.7	65.5	59.4	33.1	23.6	53.1	42.2	63.9	53.6
	ROTA	**29.3**	**17.4**	**52.9**	**37.9**	**63.9**	**48.9**	42.5	**36.3**	**62.6**	**55.5**	**71.3**	**65.2**	**41.2**	**29.4**	**64.2**	**52.2**	**74.5**	**62.8**
	E5-V	**31.8**	7.6	**56.8**	16.0	**68.7**	22.6	**49.2**	33.6	**72.0**	50.0	**81.3**	58.6	**47.2**	23.6	**73.0**	38.0	**82.1**	46.4
	Ensemble	28.9	12.8	53.9	30.3	65.6	41.2	49.0	**39.5**	70.9	61.2	79.5	**72.3**	45.1	29.9	70.5	52.1	79.4	64.0
	MTA	20.2	8.7	38.3	19.5	48.0	27.8	33.4	26.8	53.1	44.7	63.1	54.8	31.0	19.7	51.7	36.5	62.7	45.2
	ROTA	28.3	**14.1**	51.7	**32.3**	63.5	**43.1**	46.3	**39.5**	67.4	**61.5**	76.2	71.3	42.9	**31.3**	68.0	**54.3**	76.8	**66.3**

Table 11: Additional adaptation baselines on Flickr30K with LLM2CLIP: image-to-text R@1 under PGD, MI-FGSM, and CW.

Method	Adaptation type	PGD	MI-FGSM	CW
Direct retrieval	None	0.0	0.0	4.1
TeCoA (Mao et al., 2022)	Adversarial fine-tuning	35.5	33.2	40.0
FARE (Schlarmann et al., 2024)	Unsupervised adversarial fine-tuning	21.4	17.9	24.2
TPT (Shu et al., 2022)	Test-time prompt tuning	35.0	18.7	32.5
TDA (Karmanov et al., 2024)	Test-time adaptation	14.4	2.1	21.9
DOTA (Han et al., 2025)	Test-time adaptation	13.9	2.1	20.8
ROTA	No parameter updates	**64.2**	**56.1**	**67.8**

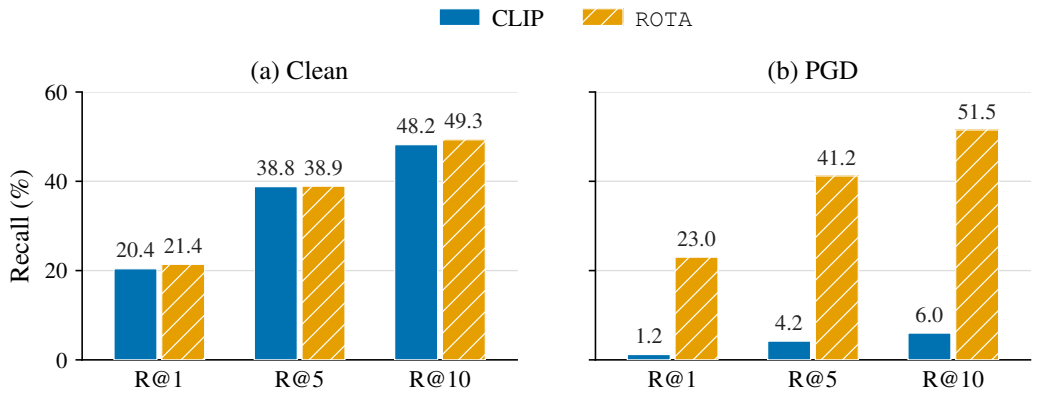


Figure 7: Video-to-text recall on the 1,000-clip Akai MovieActions dataset with frozen CLIP ViT-B/32, using eight frames per video.

C Supplementary Analyses

C.1 Adaptive Attack Evaluation

Retrieval under EOT/BPDA attacks. Table 12 evaluates Flickr30K retrieval with LLM2CLIP under attacks targeting both original-image and augmented-view representations. EOT averages gradients over eight stochastic views per step, and BPDA uses straight-through approximations for non-differentiable transforms, with 10 steps and $\epsilon = 8/255$. ROTA achieves I2T/T2I R@1 of $73.70/61.54$, compared with $5.50/2.78$ for Vanilla, retaining useful retrieval performance under this finite attack configuration.

Table 12: Flickr30K recall with LLM2CLIP under adaptive EOT/BPDA attacks: eight EOT samples per step, 10 steps, and $\epsilon = 8/255$.

Method	T2I			I2T		
	R@1	R@5	R@10	R@1	R@5	R@10
Vanilla	2.78	7.98	11.86	5.50	12.10	17.70
ROTA	**61.54**	**81.54**	**86.52**	**73.70**	**87.40**	**90.80**

C.2 Cross-Modal Matching Margins

Matched and competing scores. Table 13 measures the separation between the strongest matched caption and the hardest negative on Flickr30K with LLM2CLIP. PGD changes the mean cosine margin from 0.0300 to -0.0547, while ROTA obtains a positive mean OT margin of 0.0072. The positive OT margin reflects average separation on a different scale from cosine similarity.

Table 13: Matching margins and I2T recall on Flickr30K with LLM2CLIP under PGD. Scores and margins are query averages. The per-query margin is the highest of five matched-caption scores minus the highest non-matching score. Clean/Attack use cosine scores, while ROTA uses OT scores; their magnitudes are not directly comparable.

Setting	Best matched score	Hardest-negative score	Mean margin	I2T R@1	I2T R@5
Clean	0.3270	0.2971	0.0300	78.5	96.0
Attack	0.1937	0.2484	-0.0547	0.0	0.4
Attack + ROTA	0.0623	0.0550	0.0072	64.2	80.2

C.3 LARGE-SCALE RETRIEVAL EFFICIENCY

Recall and resource cost. Table 14 evaluates LLM2CLIP on approximately 2.5M COCO captions (Lin et al., 2014), with timings including image encoding. Experiments are conducted on an NVIDIA A100 GPU, and image encoding accounts for most of the end-to-end latency. For Top-50 reranking, we first score each candidate caption by averaging the pairwise cosine similarities between the query's image views and the candidate's text views. We retain the 50 highest-scoring captions and rerank them using OT, aligning the query's image-view distribution with each retained caption's text-view distribution, including its original and augmented views. This restricts Sinkhorn computation to 50 candidates, while the initial similarity scoring still covers the full gallery. Using 32 image views, Top-50 ROTA reaches 63.5 PGD R@1 versus 45.0 for Ensemble at similar latency (0.95 versus 0.87 s/query). Full-gallery scoring reaches 67.6 R@1 but raises peak memory from 55.03 to 59.04 GiB. Eight-view Top-50 reranking reduces latency to 0.57 s/query with 57.0 R@1, providing a lower-cost option with reduced recall.

Table 14: LLM2CLIP I2T retrieval over approximately 2.5M COCO captions. Latency includes image encoding. View counts denote augment image views; ROTA uses eight augmented text views.

Method	Clean R@1	PGD R@1	Time/query (s)	Peak GPU memory (GiB)
Vanilla	69.4	12.9	0.18	5.97
Ensemble, 32 views	69.6	45.0	0.87	55.03
ROTA, 8 views + Top-50	69.5	57.0	0.57	54.06
ROTA, 32 views + Top-50	69.8	63.5	0.95	55.03
ROTA, 32 views + full gallery	**70.0**	**67.6**	1.14	59.04

C.4 RETRIEVAL EXAMPLES

We present additional retrieval examples from COCO and Flickr30K in Figure 8, covering diverse objects, activities, and scenes. Each example compares the top-1 captions retrieved from clean images, adversarially perturbed images, and perturbed images processed by ROTA. Although the perturbed images retain recognizable visual content, direct retrieval returns captions describing mismatched objects or activities. In these examples, ROTA restores the clean-image top-1 captions, illustrating its ability to recover semantically relevant retrieval results under image-side attacks.

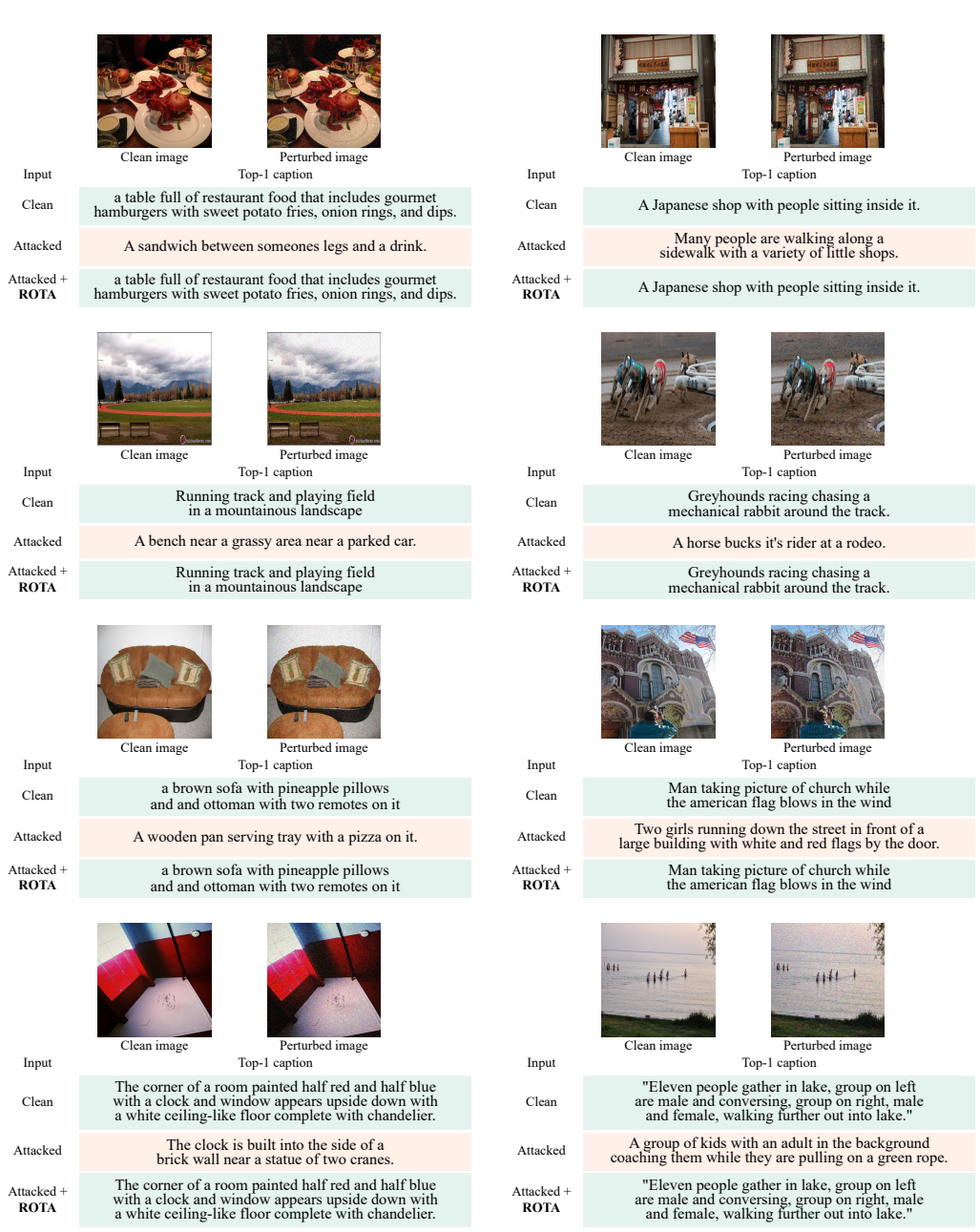


Figure 8: Retrieval examples on COCO (left) and Flickr30K (right). Each case shows clean, attacked, and ROTA top-1 captions.