MARP: TASK-GENERALIZED META-ATTACK FOR PROMOTION IN CLIP-BASED TEXT-TO-IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

CLIP-based text-to-image retrieval is vulnerable to ranking manipulation: an attacker can promote a target image under selected text queries. We study this threat from a task-level perspective: an attacker may need to solve many promotion tasks with different target images and query sets. Existing iterative attacks typically solve these tasks independently, optimizing a new perturbation for each test task without explicitly exploiting shared structure across tasks. This raises a natural question: can an attack strategy learn across such tasks and generalize to unseen ones? We formulate this problem as a *task-generalized meta-attack* and propose **MARP** (**M**eta **A**ttack for **R**etrieval **P**romotion), an amortized generator that produces an ℓ_∞-bounded task-specific perturbation in a single forward pass. MARP jointly conditions the generator on the target image and query set using frozen CLIP representations, and trains it with a bank-aware ranking surrogate that directly compares the adversarial target with competing images in the retrieval bank. On held-out attack tasks constructed from MSCOCO and Flickr30K, MARP achieves the best performance among the evaluated amortized or generator-based attacks while substantially reducing online attack-generation cost relative to iterative PGD. Task-level surrogate improvement is consistently associated with retrieval improvement, while further analyses characterize generalization across query-set regimes, task difficulty, and perturbation budgets.

1 INTRODUCTION

Text-to-image retrieval has become a common interface for visual search. Given a text query, the system returns a ranked list of database images. CLIP-style vision-language models make this process simple and scalable: a text encoder and an image encoder map queries and images into a shared embedding space, and retrieval is performed by ranking image-text similarity scores (Radford et al., 2021).

This ranking interface also changes the security objective. In classification, an attacker usually aims to change a predicted label. In retrieval, an attacker can instead manipulate exposure by changing the position of a target image in the ranked list. A target image does not need to become the top result for every query; increasing its exposure under a set of related queries can already introduce practical risks. This makes ranking promotion a natural security threat for modern multimodal retrieval systems.

Existing retrieval attacks have demonstrated that vision-language retrieval systems can be manipulated through adversarial perturbations (Zhou et al., 2020; 2021; Li et al., 2021b;a; Zhang et al., 2022; Hu et al., 2023). However, most existing approaches follow an instance-wise optimization paradigm: they solve a new optimization problem for each target image, query, or attack instance. While effective for individual cases, such a paradigm does not explicitly exploit shared structures among related promotion tasks and becomes inefficient when an attacker needs to perform many promotion operations under different target images and query sets.

Motivated by this limitation, we take a task-level view of retrieval promotion. A realistic promotion attack is naturally specified by an *attack task*, denoted as $\tau = (x, T)$, where x is a target image and $T = \{q_1, \ldots, q_m\}$ is the query set under which the target image should be promoted. As illustrated

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.

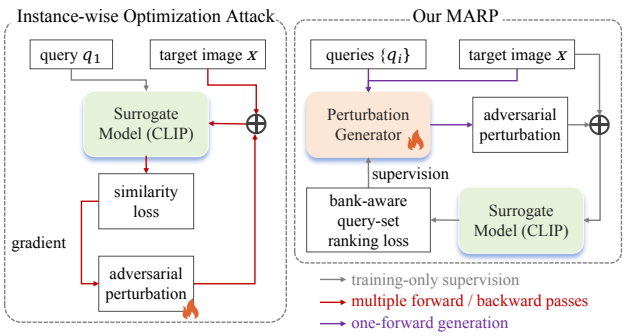

Figure 1: Motivation and comparison of attack paradigms. Instance-wise optimization attacks optimize a perturbation separately for each test task, whereas MARP learns a task-conditioned perturbation generator over a distribution of promotion tasks. The ranking loss supervises the generator during training, while inference only requires one forward pass.

in Fig. 1, this view shifts the attack paradigm from solving isolated perturbations to learning an attack strategy over a distribution of promotion tasks. Instead of optimizing a new perturbation for each test task, we aim to learn a task-conditioned generator that directly maps (x, T) to a task-specific perturbation. This formulation captures an important property of retrieval promotion: the optimal perturbation depends jointly on the target image and the query set. The same target image may require different perturbations for different query sets, while the same query set may induce different perturbations for different target images because their positions relative to the retrieval bank differ. Therefore, the key question is not only how to attack an individual retrieval instance, but how to learn an attack strategy that generalizes across unseen attack tasks.

We formulate this problem as a *task-generalized meta-attack*, where the attack strategy is learned over a distribution of attack tasks. Unlike universal perturbations that aim for task-agnostic transfer, our objective is to learn task-conditioned adaptation across unseen promotion tasks. During training, the attacker observes a collection of attack tasks and learns an amortized generator. During inference, the generator directly maps a new task to a task-specific perturbation without additional iterative optimization.

We propose **MARP** (**M**eta **A**ttack for **R**etrieval **P**romotion), a task-conditioned generator for CLIP-based retrieval promotion. MARP encodes the target image and query set using frozen CLIP representations and learns a task-conditioned mapping to an ℓ_∞-bounded perturbation. To preserve the ranking nature of retrieval promotion, MARP is trained with a bank-aware ranking surrogate that directly compares the adversarial target image against competing images in the retrieval bank under the queries of each task.

Beyond measuring aggregate attack success, we further analyze whether the optimization signal used to train the amortized generator is meaningfully related to its downstream retrieval behavior. Across both datasets, task-level improvement in the optimized surrogate is consistently associated with improvement in retrieval outcomes, even after controlling for the clean difficulty of the attack task. This analysis provides additional evidence that the proposed objective captures meaningful task-level promotion progress.

Our contributions are:

- We formulate CLIP-based text-to-image retrieval promotion as a *task-generalized meta-attack* and propose MARP, an amortized generator conditioned on the target image and query-set semantics.
- We introduce a bank-aware query-set ranking surrogate that directly optimizes the target against the retrieval bank and connect it to empirical outranking.
- Experiments on MSCOCO and Flickr30K show that MARP is strongest among evaluated amortized attacks while requiring substantially lower online cost than iterative optimization. Further analyses relate surrogate improvement to retrieval outcomes and assess robustness across query-set regimes, task difficulty, and perturbation budgets.

2 Related Work

Vision-language retrieval. CLIP learns image and text encoders from large-scale natural-language supervision and supports text-to-image retrieval through a shared embedding space (Radford et al., 2021). This dual-encoder structure is efficient, but it also means that a small image perturbation can change the relative similarity between a target image, many queries, and the retrieval bank.

Adversarial ranking and retrieval attacks. Ranking attacks aim to change the order of retrieved candidates rather than a single prediction. Adversarial ranking attacks that raise or lower selected candidates in deep ranking systems have been formulated (Zhou et al., 2020) and later extended to relative order manipulation (Zhou et al., 2021). QAIR studies query-efficient black-box attacks for image retrieval (Li et al., 2021b). These works motivate our ranking-based objective, but they do not study query-set-conditioned promotion in CLIP-based text-to-image retrieval.

Attacks on cross-modal and vision–language models. Cross-modal retrieval and vision–language models are vulnerable to adversarial perturbations. AACH attacks cross-modal Hamming retrieval (Li et al., 2021a), whereas Co-Attack jointly perturbs matched image–text pairs (Zhang et al., 2022). SGA improves transferability by disrupting matched-pair alignment with set-level guidance and can perturb both modalities (Lu et al., 2023). AdvCLIP instead learns downstream-agnostic universal patches for non-targeted degradation (Zhou et al., 2023). Neither directly addresses image-only promotion of a chosen target under a fixed query set and retrieval bank; adapting them would change their objectives and threat models. Recent work further studies black-box and transferable attacks against VLP/VLM models (Liu et al., 2025a; Chen et al., 2026; Hu et al., 2026; Liu et al., 2025b).

Promotion and poisoning attacks in multimodal retrieval. THA studies a Trojan-horse attack against text-to-image retrieval by embedding word-specific adversarial information into visible QR-code-like patches (Hu et al., 2023). Collisio studies data poisoning against text-to-image retrieval and targets semantically varied query forms through expectation over queries (Lazzaro et al., 2026). ViPro studies adversarial promotion in text-to-video retrieval and frames promotion as a realistic exposure-driven threat (Tian et al., 2026). In contrast, our attack is test-time rather than poisoning-based, uses imperceptible ℓ_∞-bounded perturbations rather than visible patterns, treats the query set as part of the task, and learns an amortized generator for unseen tasks.

Meta-learning and generator-based attacks. Meta-learning trains models over task distributions (Finn et al., 2017). Adversarial learning has also used the term "meta attack", for example in meta-gradient graph poisoning (Zügner & Günnemann, 2019). AnyAttack generates targeted perturbations from a single reference-image embedding (Zhang et al., 2025), whereas MARP conditions on target–query-set tasks and optimizes a bank-aware ranking objective. Our use is different: we do not optimize an inner loop or solve a bilevel poisoning problem. We learn an amortized attack generator over retrieval-promotion tasks. This also differs from universal perturbations (Moosavi-Dezfooli et al., 2017; Zheng et al., 2024) and image-conditioned generators such as AdvGAN (Xiao et al., 2018), because MARP conditions on both the image and the query-set-defined attack goal.

3 Task-Generalized Retrieval Promotion

3.1 Retrieval Promotion as an Attack Task

Let $\mathcal{D} = \{x_j\}_{j=1}^N$ be an image database and let q be a text query. A CLIP-style retrieval model has an image encoder $E_v(\cdot)$ and a text encoder $E_t(\cdot)$. With ℓ_2-normalized features, we write

$$v_j = E_v(x_j), \qquad z = E_t(q), \qquad s(q, x_j) = z^\top v_j, \tag{1}$$

and rank images by decreasing similarity. The rank of a target image x_a under query q is

$$\mathrm{Rank}(x_a; q) = 1 + \sum_{x_j \in \mathcal{D},\, j \neq a} \mathbb{I}[s(q, x_j) > s(q, x_a)], \tag{2}$$

where a smaller rank corresponds to higher retrieval exposure.

We define an attack task as

$$\tau = (x, T), \qquad T = \{q_1, \ldots, q_m\}, \tag{3}$$

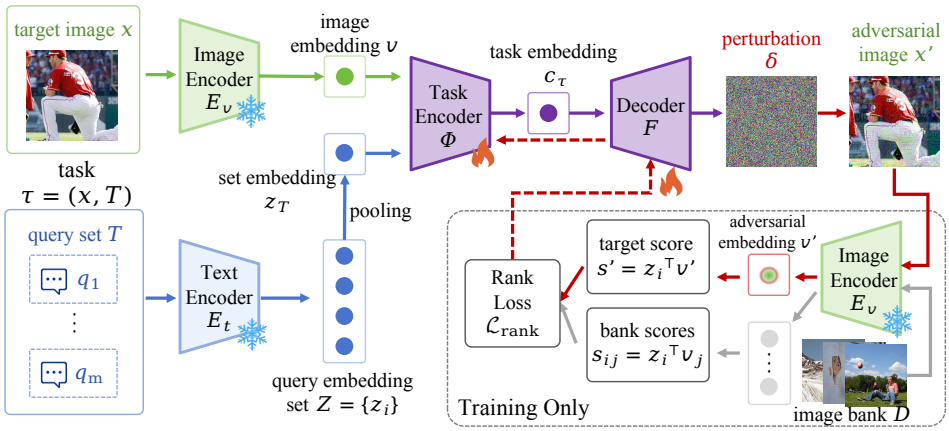

Figure 2: Overview of MARP. Frozen CLIP encoders produce representations of the target image and query set, which condition an ℓ_∞-bounded perturbation generator. Training uses a bank-aware ranking loss over the image bank, while inference requires one generator forward pass.

where x is the target image and T specifies the queries under which it should be promoted. For a bounded perturbation δ, let $x_\delta = \Pi_{[0,1]}(x + \delta)$ with $\|\delta\|_\infty \le \epsilon$. The task-level promotion objective is

$$\min_{\|\delta\|_\infty \le \epsilon} \frac{1}{|T|} \sum_{q_i \in T} \text{Rank}(x_\delta; q_i). \tag{4}$$

Thus, successful promotion moves the target upward across the ranked lists induced by the query set.

3.2 Threat Model

We consider a test-time attack with access to a fixed CLIP retrieval model, the target image, the query set, and the corresponding image-bank features. The attacker cannot modify the retrieval model, text queries, or other database images; only the target image may be perturbed within the ℓ_∞ budget defined above.

3.3 Amortized Task-Level Attack Formulation

Given training tasks $\mathcal{S}_{\text{train}} = \{\tau_a\}_{a=1}^M$ drawn from $p_{\text{train}}(\tau)$, we learn a task-conditioned generator $G_\theta : (x, T) \mapsto \delta_\tau$. Rather than solving Eq. (4) independently for every test task, we amortize attack optimization over the training-task distribution:

$$\min_\theta \mathbb{E}_{\tau \sim p_{\text{train}}} \left[\mathcal{L}_\tau \big(G_\theta(\tau) \big) \right], \tag{5}$$

where $\mathcal{L}_\tau$ is a differentiable surrogate of the discrete task-level rank objective.

For an unseen task (x, T), the trained generator directly produces a bounded perturbation without task-specific iterative optimization or parameter updates. We refer to this learning-over-tasks formulation as a *task-generalized meta-attack*. Section 4 instantiates this mapping with MARP.

4 MARP: Meta-Attack for Retrieval Promotion

Figure 2 summarizes the overall architecture of MARP.

4.1 Task Representation

For a task $\tau = (x, T)$, the frozen CLIP encoders produce the target feature $v = E_v(x)$ and query features $z_i = E_t(q_i)$. Database features $\mathcal{V} = \{E_v(x_j) : x_j \in \mathcal{D}\}$ are pre-computed and fixed. Since

T is an unordered set, we aggregate its query features by mean pooling and combine them with the target feature:

$$z_T = \frac{1}{|T|} \sum_{q_i \in T} z_i, \qquad c_\tau = \Phi_\theta([v; z_T]), \tag{6}$$

where c_τ is the task representation supplied to the perturbation generator. We use θ collectively for the trainable task-encoder and decoder parameters; the CLIP encoders remain frozen.

4.2 Task-Conditioned Perturbation Generation

A residual upsampling decoder F_θ maps c_τ to an image-sized perturbation field. We enforce the perturbation budget and valid image range by

$$\delta_\theta(x, T) = \mathrm{clip}\big(F_\theta(c_\tau), -\epsilon, \epsilon\big), \qquad x' = \Pi_{[0,1]}\big(x + \delta_\theta(x, T)\big). \tag{7}$$

At test time, the trainable attack module produces the perturbation in one forward pass, without task-specific gradient optimization. Decoder details and training pseudocode are provided in Appendix A.2.

4.3 Bank-Aware Query-Set Ranking Objective

The discrete rank in Eq. (2) is not differentiable. We therefore use a pairwise learning-to-rank surrogate (Burges et al., 2005). Let $\mathcal{D}^-(x) = \mathcal{D} \setminus \{x\}$ denote the image bank excluding the target. For query q_i, define the adversarial target score $s'_i = z_i^\top v'$ with $v' = E_v(x')$, and the bank score $s_{ij} = z_i^\top v_j$. We optimize

$$\begin{aligned} \mathcal{L}_i(x', q_i) &= \frac{1}{|\mathcal{D}^-(x)|} \sum_{x_j \in \mathcal{D}^-(x)} \mathrm{softplus}(s_{ij} - s'_i), \\ \mathcal{L}_{\mathrm{rank}}(\tau; \theta) &= \frac{1}{|T|} \sum_{q_i \in T} \mathcal{L}_i(x', q_i). \end{aligned} \tag{8}$$

The loss penalizes database images that score above or near the adversarial target and averages this penalty across the task queries. Ablations of the training objective and negative-bank construction are reported in Appendix C.2 and Appendix C.3, respectively. An idealized sampling interpretation of query-set averaging is provided in Appendix B.2.

Instantiating Eq. (5) with the bank-aware ranking surrogate, we optimize

$$\mathcal{L}_{\mathrm{train}}(\theta) = \mathbb{E}_{\tau \sim p_{\mathrm{train}}} \left[\mathcal{L}_{\mathrm{rank}}(\tau; \theta) \right]. \tag{9}$$

The CLIP encoders and image-bank features remain fixed.

4.4 Connection to Retrieval Rank

Although Eq. (8) is smooth, evaluation uses the discrete rank in Eq. (2). Let $R_i(x') = \mathrm{Rank}(x'; q_i)$.

Proposition 1 (Normalization-aware rank upper bound)**.** *Let*

$$N = |\mathcal{D}^-(x)|, \qquad c_- = \mathrm{softplus}(-2), \qquad c_0 = \ln 2,$$

and define

$$h(t) = \min\left\{1, \frac{t - c_-}{c_0 - c_-}\right\}.$$

Under ℓ_2-normalized cosine scores, for any query q_i and adversarial image x',

$$R_i(x') - 1 \le N h(\mathcal{L}_i(x', q_i)). \tag{10}$$

Consequently,

$$\frac{1}{|T|} \sum_{q_i \in T} R_i(x') \le 1 + N h(\mathcal{L}_{\mathrm{rank}}(\tau; \theta)). \tag{11}$$

Thus, the smooth objective $\mathcal{L}_{\mathrm{rank}}$ controls a nondecreasing upper bound on the average target rank; reducing the objective cannot increase this bound. The proof is given in Appendix B.1.

5 EXPERIMENTS

We evaluate MARP along four questions: **Q1**, its promotion effectiveness and online efficiency; **Q2**, the contribution of task conditioning; **Q3**, robustness across query-set regimes; and **Q4**, whether improvement in the training surrogate is associated with retrieval improvement on held-out tasks.

5.1 EXPERIMENTAL SETUP

Datasets and attack tasks. We construct retrieval-promotion task datasets from MSCOCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015) following the Karpathy splits (Karpathy & Fei-Fei, 2017). For each task $\tau = (x, T)$, we select a seed image and use its associated captions to define a semantic neighborhood in the frozen CLIP text space; nearby captions form the query set T, while different images are sampled as promotion targets x. Table 1 summarizes the default TopK-32 task populations, and full task construction and split details are provided in Appendix A.1.

Table 1: Retrieval splits and default TopK-32 attack-task statistics.

Dataset	Seed Images (Tr./Val./Te.)	Targets/Seed (Tr./Val./Te.)	Attack Tasks (Tr./Val./Te.)	$\lvert T\rvert$
MSCOCO	113,287 / 5,000 / 5,000	1 / 2 / 2	113,287 / 10,000 / 10,000	32
Flickr30K	29,000 / 1,014 / 1,000	5 / 5 / 2	145,000 / 5,070 / 2,000	32

Backbone and implementation. We use frozen OpenAI CLIP ViT-B/32 (Radford et al., 2021) with ℓ_2-normalized image and text embeddings and cosine-similarity retrieval. Unless otherwise specified, experiments use TopK-32 tasks, where the 32 nearest captions in the CLIP text space constitute the query set, and an ℓ_∞ budget of $\epsilon = 16/255$. MARP is trained for 100 epochs with AdamW at learning rate 10^{-4}. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU. Additional implementation and training details are provided in Appendix A.2.

Baselines. We compare MARP with per-task white-box attacks (FGSM (Goodfellow et al., 2014) and PGD (Madry et al., 2017)), a universal perturbation (UAP), conditioning variants (I-Only and Q-Only), and AnyAttack (Zhang et al., 2025) adapted to our retrieval-promotion setting (AA-bi and AA-cos). FGSM and PGD optimize the same bank-aware query-set ranking objective separately for each test task under an ℓ_∞ budget of $\epsilon = 16/255$. FGSM performs a single update with step size $\alpha = \epsilon$. PGD-n, where $n \in \{3, 5, 10\}$, performs n updates with step size $\alpha = \epsilon/n$. For AA-bi and AA-cos, we use the dataset-specific AnyAttack checkpoints released by the original authors, and evaluate them under the same perturbation budget. I-Only conditions only on the target image, whereas Q-Only conditions only on the aggregated query representation; both variants otherwise use the same bank-aware objective and training protocol as MARP.

Metrics and statistics. For an evaluation set of M tasks and evaluated targets $\tilde{x}_m$, the task-macro-averaged hit rate is

$$\mathrm{HR@}k = \frac{100}{M} \sum_{m=1}^{M} \frac{1}{|T_m|} \sum_{q_{m,i} \in T_m} \mathbb{I}[R_{m,i}(\tilde{x}_m) \le k], \qquad k \in \{1, 5, 10\}, \tag{12}$$

where $\tilde{x}_m$ is the clean or adversarial target and $R_{m,i}(\tilde{x}_m)$ is its rank for query $q_{m,i}$. We also report mean target rank (MR; lower is better) and online attack-generation time; the timing protocol is detailed in Appendix A.3. Inferential analyses use 10,000 seed-cluster bootstrap resamples with Holm correction within each analysis family, treating the seed image as the resampling cluster. Surrogate–outcome analysis additionally reports raw and partial Spearman correlations controlling for clean MR.

5.2 MAIN RESULTS

Table 2 evaluates promotion on held-out TopK-32 attack tasks. PGD-10 achieves the strongest overall promotion, whereas MARP obtains the best results among the evaluated amortized or generator-based attacks on all reported metrics for both datasets. This shows that task-level amortization

Table 2: Main promotion results on Flickr30K and MSCOCO TopK-32 tasks. HR@K is in percentage and MR is mean target rank. **Bold** denotes the overall best result and underline the best amortized or generator-based attack.

Dataset	Method	HR@1 ↑	HR@5 ↑	HR@10 ↑	MR ↓
Flickr30K	Clean	0.07	0.40	0.93	494.03
	AA-bi	0.08	1.17	2.81	398.60
	AA-cos	0.14	2.33	5.25	255.56
	FGSM	0.45	3.19	6.29	207.07
	PGD-3	12.69	36.39	48.96	40.20
	PGD-5	30.36	61.67	72.77	16.42
	PGD-10	**62.36**	**86.63**	**91.97**	**5.11**
	UAP	16.12	37.40	48.17	58.01
	I-Only	12.84	32.95	44.63	49.89
	Q-Only	42.88	65.73	73.36	28.93
	MARP	44.66	67.91	76.20	20.45
MSCOCO	Clean	0.01	0.09	0.20	2533.02
	AA-bi	0.03	0.39	0.95	1026.98
	AA-cos	0.03	0.31	0.68	1258.77
	FGSM	0.10	0.54	1.06	958.32
	PGD-3	6.86	16.94	24.11	124.92
	PGD-5	25.45	45.52	55.46	42.52
	PGD-10	**68.07**	**84.41**	**89.23**	**9.36**
	UAP	11.46	21.77	28.21	186.58
	I-Only	5.97	12.53	17.36	178.79
	Q-Only	51.32	64.63	69.84	75.70
	MARP	56.78	68.24	72.97	59.87

retains substantial promotion strength without solving a new optimization problem for every test task.

The iterative baselines reveal a clear optimization-depth effect. From FGSM to PGD-10, HR@1 increases from 0.45% to 62.36% and MR decreases from 207.07 to 5.11 on Flickr30K; the corresponding changes on MSCOCO are from 0.10% to 68.07% and from 958.32 to 9.36, respectively. Because these attacks share the same perturbation budget and ranking objective, the monotonic improvements from PGD-3 to PGD-5 and PGD-10 demonstrate the benefit of task-specific iterative optimization and show that a single gradient step is insufficient for this ranking problem.

MARP occupies a distinct point on this optimization continuum. It exceeds PGD-5 in HR@1 by 14.30 and 31.33 percentage points on Flickr30K and MSCOCO, respectively, although PGD-5 obtains lower MR by 4.03 and 17.35 ranks. This difference indicates that top-rank hit rates and average ranking depth capture complementary attack behavior. Relative to the task-agnostic UAP, MARP improves HR@1 by 28.54 and 45.32 percentage points and reduces MR by 64.7% and 67.9% on the two datasets.

MARP also provides substantially lower online attack-generation cost than PGD-10, requiring approximately 7.2 ms per task compared with 66–72 ms for PGD-10 across the two datasets. We use these measurements to characterize the practical latency gap rather than as a hardware-level microbenchmark; full timing details are provided in Appendix A.3.

5.3 Ablation Studies

5.3.1 Task Conditioning

We use the task-conditioning variants in Table 2 to examine what information should be exposed to the perturbation generator. This ablation only removes information from the generator input, while all methods are still trained and evaluated on the same attack tasks (x, T) with the same query-set ranking objective.

Table 3: Seed-level association between surrogate improvement and retrieval improvement. "Partial" denotes partial Spearman correlation controlling for clean mean rank. All eight tests in this analysis family remain significant after Holm correction ($p_{\mathrm{Holm}} = 0.0016$). Positive values indicate that larger surrogate improvement is associated with larger retrieval improvement.

Dataset	Outcome	Raw ρ	Partial ρ
MSCOCO	HR@1 gain	0.772	**0.821**
	MR reduction	0.319	**0.413**
Flickr30K	HR@1 gain	0.696	**0.782**
	MR reduction	0.420	**0.516**

Specifically, UAP learns a task-agnostic universal perturbation shared by all tasks. I-Only conditions the generator only on the target image, while the query set is still used in the ranking loss. Q-Only conditions the generator only on the query set, while the target image is still perturbed and evaluated under the same ranking objective. MARP conditions the generator on both the target image and the query set.

As shown in Table 2 and Table 5, query conditioning provides a substantial fraction of the promotion strength: Q-Only is considerably stronger than the task-agnostic and I-Only alternatives. Joint image–query conditioning further improves the aggregate attack performance on both datasets, showing that the target image provides complementary task-specific information beyond the query set alone. The magnitude of this additional benefit, however, varies with dataset, metric, and task difficulty.

We examine this dependence in more detail by stratifying the frozen test tasks according to clean mean rank in Appendix F.1. On MSCOCO, MARP obtains significant HR@1 and mean-rank advantages over Q-Only in all three difficulty tertiles, and the mean-rank advantage reaches 28.37 ranks on Hard tasks. On Flickr30K, MARP obtains significant mean-rank advantages over Q-Only across Easy, Medium, and Hard tasks, with the advantage increasing to 17.50 ranks on the Hard stratum. These results indicate that joint conditioning provides additional benefit, particularly for deeper ranking improvement, while the magnitude of that benefit depends on dataset, metric, and task difficulty. Appendix F.2 further evaluates the conditioning variants across perturbation budgets. The results show that the benefit of target-image conditioning varies with the perturbation budget, suggesting that joint conditioning becomes more valuable when sufficient perturbation capacity is available for task-specific adaptation.

5.3.2 Objective and Bank Design

We further investigate whether the retrieval-aware supervision is necessary for learning effective promotion strategies. Removing the bank-aware ranking objective or replacing the full retrieval bank with reduced negative subsets consistently weakens promotion performance. These results indicate that optimizing relative positions against the retrieval bank provides a more faithful training signal than optimizing against partial negatives. Detailed comparisons are provided in Appendix C.2 and Appendix C.3.

5.3.3 Surrogate–Outcome Association

The training objective is optimized in surrogate space, whereas attack success is ultimately measured by retrieval rank. We therefore examine whether task-level improvement in the optimized surrogate is associated with improvement in held-out retrieval outcomes. For each seed, we average its two target tasks and measure surrogate improvement together with HR@1 gain and mean-rank reduction. We report both raw Spearman correlation and partial Spearman correlation after controlling for canonical clean mean rank.

Surrogate improvement is positively associated with both retrieval outcomes on both datasets. The strongest relationships appear for HR@1 gain, with partial correlations of 0.821 on MSCOCO and 0.782 on Flickr30K. Mean-rank reduction shows more moderate but still consistent associations. Importantly, controlling for clean mean rank does not remove these relationships, indicating that they cannot be reduced simply to easier tasks exhibiting larger gains.

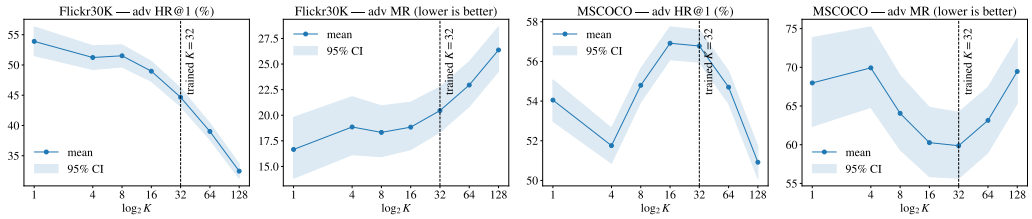

Figure 3: Query-set-size generalization of a single MARP checkpoint trained at $K = 32$. Points show the mean adversarial HR@1 and MR, and shaded regions denote 95% seed-cluster bootstrap confidence intervals. The horizontal axis is shown on a base-2 logarithmic scale.

These results provide empirical support that optimization progress in the task surrogate tracks downstream promotion progress on unseen tasks. Because the analysis compares naturally varying improvements across frozen test tasks, we interpret the results as statistical associations rather than causal effects. Full confidence intervals and corrected significance tests are reported in Appendix D.

5.4 Generalization Across Query-Set Sizes

To examine whether MARP relies on the query-set size observed during training, we freeze the checkpoint trained on TopK-32 tasks and evaluate it at $K \in \{1, 4, 8, 16, 32, 64, 128\}$. This evaluation changes the task specification at test time while keeping the attack model fixed, providing a direct test of task-level generalization beyond the training query-set regime. We use source-anchored nested query sets: smaller sets are canonical prefixes of the frozen TopK-32 set, whereas larger sets retain all frozen TopK-32 members and append the nearest outside queries. All variants preserve the task, seed image, target image, retrieval bank, checkpoint, and perturbation budget.

On MSCOCO, performance forms a plateau around $K = 16$–32: HR@1 is nearly unchanged between these two settings, while $K = 32$ achieves the lowest mean rank. Performance is weaker at both $K = 8$ and $K = 64$ in terms of mean rank. On Flickr30K, promotion effectiveness decreases as the nested query set expands from $K = 8$ to $K = 128$, with HR@1 decreasing and mean rank increasing. These results show that MARP can operate at query-set sizes not observed during training, but the most favorable evaluated query-set regime is dataset dependent.

The statistical comparisons are post-hoc and are reported in Appendix E.1. Because changing K also changes the identity and semantic composition of the added queries, this experiment does not isolate a causal effect of cardinality alone. We therefore interpret it as a test of generalization across source-anchored query-set regimes.

Appendix E.2 evaluates radius-defined adaptive query sets; their distinct dataset-dependent patterns show sensitivity to semantic-region construction.

Qualitative analysis. Appendix G reports mechanically selected high-, typical-, and low-gain cases under a rule fixed before image inspection, with queries kept in frozen order.

6 Conclusion

We introduced MARP, a task-generalized meta-attack that generates task-conditioned promotion perturbations without per-task iterative optimization. Across MSCOCO and Flickr30K, MARP is the strongest evaluated amortized attack. Its bank-aware surrogate upper-bounds empirical outranking, and task-level surrogate improvement is associated with retrieval improvement. Conditioning and query-set analyses further reveal dataset- and construction-dependent generalization. Our evaluation is limited to one frozen CLIP backbone and two benchmarks. Training also assumes white-box access to the victim encoders and image-bank features, while the cost of the full-bank ranking objective scales with the retrieval-bank size. Future work should investigate scalable approximations to full-bank supervision, transfer to black-box retrieval systems and retrieval pipelines using stronger vision-language encoders or LLM/VLM-based rerankers, and defenses against amortized attacks through robust training, perturbation detection, and retrieval-time consistency checks.

Reproducibility Statement

We provide detailed information to support reproducibility throughout the paper and appendix. Section 5.1 specifies the datasets, retrieval backbone, attack-task populations, perturbation budget, baselines, evaluation metrics, and statistical protocol. Appendix A.1 describes the complete attack-task construction procedure, and Appendix A.2 provides implementation and training details. Theoretical proofs and additional analyses are included in the corresponding appendix sections. The attack-generation timing protocol is documented in Appendix A.3. We provide anonymized code and configuration files as supplementary material.

Ethics Statement

This work studies adversarial promotion vulnerabilities in text-to-image retrieval systems and is therefore inherently dual-use. The proposed techniques could potentially be misused to increase the exposure of malicious or undesired content in retrieval systems. Our experiments are conducted only on public image-caption benchmarks and offline retrieval models; we do not attack deployed services or interact with real users. We present these results to better characterize retrieval-security risks and to support the development and evaluation of future defenses.

AI Use Statement

Generative AI tools assisted with language editing, code drafting, and consistency checks. All AI-assisted outputs were reviewed and verified by the authors. The authors remain responsible for all claims, citations, experimental results, and released artifacts.

References

Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96, 2005.

Wutao Chen, Huaqin Zou, Chen Wan, and Lifeng Huang. A two-stage globally-diverse adversarial attack for vision-language pre-training models. *arXiv preprint arXiv:2601.12304*, 2026.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Chengyin Hu, Jiaju Han, Xuemeng Sun, Qike Zhang, Luwei Yang, Lehan Sun, Jiahuan Long, Yiwei Wei, and Jiujiang Guo. Xspa: Crafting imperceptible x-shaped sparse adversarial perturbations for transferable attacks on vlms. arXiv preprint arXiv:2603.28568, 2026.

Fan Hu, Aozhu Chen, and Xirong Li. Towards making a trojan-horse attack on text-to-image retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676, April 2017. ISSN 0162-8828.

Dario Lazzaro, Raffaele Mura, Antonio Emanuele Cinà, Giuseppe Laurita, Gianni Vercelli, Luca Oneto, Battista Biggio, and Fabio Roli. Poison once, fool many: Practical poisoning attacks against text-to-image retrieval systems. *Knowledge-Based Systems*, 334:115090, 2026.

Chao Li, Shangqian Gao, Cheng Deng, Wei Liu, and Heng Huang. Adversarial attack on deep cross-modal hamming retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2218–2227, 2021a.

Xiaodan Li, Jinfeng Li, Yuefeng Chen, Shaokai Ye, Yuan He, Shuhui Wang, Hang Su, and Hui Xue. QAIR: Practical query-efficient black-box attacks for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3330–3339, 2021b.

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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *ECCV 2014*, pp. 740–755, 2014.

Han Liu, Jiaqi Li, Zhi Xu, Xiaotong Zhang, Xiaoming Xu, Fenglong Ma, Yuanman Li, and Hong Yu. Hqa-vlattack: Towards high quality adversarial attack on vision-language pre-trained models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 170967–170986, 2025a.

Kuanrong Liu, Siyuan Liang, Cheng Qian, Ming Zhang, and Xiaochun Cao. Bridging the task gap: Multi-task adversarial transferability in CLIP and its derivatives. *arXiv preprint arXiv:2509.23917*, 2025b.

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. *arXiv preprint arXiv:1706.06083*, 2017.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1765–1773, 2017.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2641–2649, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Qiwei Tian, Chenhao Lin, Zhengyu Zhao, Shuai Liu, Qian Li, and Chao Shen. Adversarial video promotion against text-to-video retrieval. *IEEE Transactions on Information Forensics and Security*, 2026.

Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 3905–3911, 2018.

Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5005–5013, 2022.

Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19900–19909. IEEE, 2025.

Haonan Zheng, Wen Jiang, Xinyang Deng, and Wenrui Li. Sample-agnostic adversarial perturbation for vision-language pre-training models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9749–9758, 2024.

Mo Zhou, Zhenxing Niu, Le Wang, Qilin Zhang, and Gang Hua. Adversarial ranking attack and defense. In *European Conference on Computer Vision*, pp. 781–799. Springer, 2020.

Mo Zhou, Le Wang, Zhenxing Niu, Qilin Zhang, Yinghui Xu, Nanning Zheng, and Gang Hua. Practical relative order attack in deep ranking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16413–16422, 2021.

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.

Daniel Zügner and Stephan Günnemann. Adversarial attacks on graph neural networks via meta learning. In *7th International Conference on Learning Representations (ICLR)*, 2019.

A Experimental and Reproducibility Details

This section provides the task-construction, implementation, training, and timing details underlying the experimental protocol in Section 5.1.

A.1 Attack Task Construction

Attack tasks are constructed independently within each dataset split using frozen CLIP features. For each seed image, its associated captions define a local semantic query region; one or more different images from the same split are then sampled as promotion targets. Each resulting task follows the generic notation $\tau = (x, T)$ in Eq. (3), where x is the target image and $T = T_s$ is the query set induced by seed image x_s.

Retrieval index. For a dataset split, let $D = \{x_j\}_{j=1}^{N}$ denote the image set and $Q = \{q_i\}_{i=1}^{M}$ the caption pool. Each image is associated with its ground-truth captions. We encode all images and captions with the frozen CLIP encoders,

$$V = \{v_j = E_v(x_j)\}_{j=1}^{N}, \qquad Z = \{z_i = E_t(q_i)\}_{i=1}^{M},$$

and ℓ_2-normalize all features.

Seed-induced query region. We construct the query set from a seed image rather than from the promotion target itself. For seed image x_s, let A_s denote the indices of its associated captions. For every caption q_i in the split, we compute its minimum cosine distance to the seed-caption anchors:

$$d_s(q_i) = \min_{a \in A_s} \left(1 - z_i^\top z_a\right). \tag{13}$$

Smaller $d_s(q_i)$ indicates that q_i lies closer to the local semantic region represented by the seed captions.

TopK query-set construction. The default attack tasks use fixed-cardinality TopK query sets. For a given K, we select

$$T_s^{\mathrm{topk}} = \mathrm{TopK}_{q_i \in Q}\big(-d_s(q_i)\big). \tag{14}$$

All selected queries receive equal weight. Unless otherwise specified, the default experiments use $K = 32$, yielding a query set of identical cardinality for every task.

Source-anchored nested query sets. For the query-set-regime analysis in Section 5.4, we construct nested variants around the frozen TopK-32 task. We first compute a canonical ordering using cosine distances in the frozen construction feature space, evaluated in CPU float64 with ascending query index used to break ties. For $K < 32$, we retain the first K members of the canonically ordered frozen TopK-32 set. For $K > 32$, we retain all frozen TopK-32 members and append the closest queries outside that set.

The construction feature bank is used only to determine query membership and ordering; the frozen evaluation feature bank supplies the embeddings used by the retrieval evaluator. At $K = 32$, the reference task membership is preserved exactly. Recomputed distances are used only to establish the canonical ordering and construct the smaller or extended query sets; they never modify the frozen TopK-32 membership.

Radius-defined query sets. For robustness analysis, we additionally construct adaptive query sets using a cosine-distance radius ρ:

$$T_s^{\mathrm{radius}}(\rho) = \{q_i \in Q : d_s(q_i) \le \rho\}. \tag{15}$$

Unlike TopK construction, radius-defined sets have variable cardinality. Changing ρ generally changes both the number and semantic composition of the selected queries. The corresponding experiment is therefore interpreted as robustness to an alternative query-set construction rather than as an isolated causal intervention on query-set size.

Algorithm 1 Attack Task Construction

Require: Image-caption split (D, Q), frozen E_v, E_t, construction mode, K or ρ, targets per seed n_t

Encode all images and captions to obtain V and Z
for each seed image $x_s \in D$ **do**
  Let A_s be the caption indices associated with x_s
  for each caption $q_i \in Q$ **do**
    Compute $d_s(q_i) = \min_{a \in A_s}(1 - z_i^\top z_a)$
  end for
  if mode is TopK **then**
    Select the K captions with the smallest $d_s(q_i)$ as T_s
  else
    Select $T_s = \{q_i : d_s(q_i) \le \rho\}$
  end if
  Sample n_t targets from $D \setminus \{x_s\}$
  for each sampled target x **do**
    Set $T \leftarrow T_s$, create task $\tau = (x, T)$, and save its metadata
  end for
end for

Target sampling. After constructing the seed-induced query set T_s, we sample one or more target images from the same split while excluding the seed image:

$$x \sim D \setminus \{x_s\}, \qquad T = T_s, \qquad \tau = (x, T). \tag{16}$$

The promotion target is therefore not required to be semantically aligned with the seed-induced query region, matching the setting in which an attacker seeks to promote a selected image under a group of semantically related queries.

Task metadata. For each task, we store the seed and target image identifiers, query identifiers and texts, query embeddings, and construction metadata. We also record analysis quantities such as clean mean rank, retrieval hit indicators, query-set size, intra-query cosine similarity, and seed–target feature similarity. These quantities are used for evaluation and analysis; the attack generator itself receives the target image and query-set representation defined in Section 4.1.

Split protocol. Attack tasks are constructed independently within the training, validation, and test retrieval splits. For MSCOCO, the 113,287 training seeds use one target per seed, yielding 113,287 training tasks; the 5,000 validation and 5,000 test seeds use two targets per seed, yielding 10,000 tasks in each evaluation split. For Flickr30K, the 29,000 training and 1,014 validation seeds use five targets per seed, yielding 145,000 and 5,070 tasks, respectively, while the 1,000 test seeds use two targets per seed, yielding 2,000 test tasks.

The generator is trained only on training tasks, checkpoint selection uses validation tasks, and all reported test results use the held-out test-task population. No test task is used for training or checkpoint selection.

Algorithm 1 summarizes the construction procedure.

A.2 Implementation and Training Details

Trainable attack module. The CLIP image and text encoders remain frozen throughout training. The trainable attack module is implemented as a `ReBootGenerator`, consisting of a task-condition encoder and a residual upsampling decoder. Image-bank features are pre-computed once and remain fixed during optimization.

Task-condition encoder. For an attack task $\tau = (x, T)$, the frozen CLIP encoders produce a 512-dimensional target-image feature and 512-dimensional query features. Query features are mean-pooled over the valid queries using the query mask; no additional query weights are used. Using

Algorithm 2 Training MARP

Require: Training tasks $\mathcal{S}_{\text{train}}$, frozen E_v, E_t, image bank $\mathcal{D}$, perturbation budget ϵ

```
Pre-compute image-bank features V
for each mini-batch of attack tasks do
    for each task τ = (x, T) do
        Encode the target image and valid task queries with frozen CLIP
        Mean-pool the valid query features and compute c_τ
        Generate adversarial image x' using Eq. (7)
        Re-encode x' with frozen E_v to obtain v' = E_v(x')
        Compute L_rank using Eq. (8)
    end for
    Update the condition encoder and perturbation decoder using AdamW
end for
Select the checkpoint with the lowest validation loss
```

the notation of Eq. (6), the pooled query feature z_T is concatenated with the target-image feature $v = E_v(x)$ to form a 1024-dimensional input:

$$h_\tau = [v; z_T] \in \mathbb{R}^{1024}. \tag{17}$$

The condition encoder is a two-layer MLP,

$$1024 \to 1024 \to 512, \tag{18}$$

with LayerNorm and GELU in the intermediate layer. Its output is the 512-dimensional task representation c_τ supplied to the perturbation decoder. This implementation instantiates the task representation in Eq. (6).

Perturbation decoder. The decoder first projects the 512-dimensional task representation into a $256 \times 14 \times 14$ feature map. It then applies five residual blocks interleaved with four nearest-neighbor $\times 2$ upsampling blocks:

$$\begin{aligned} 512 \xrightarrow{\text{FC}} 256 \times 14 \times 14 &\xrightarrow{\text{Res}(256)} \xrightarrow{\text{Up}(256\to128)} \xrightarrow{\text{Res}(128)} \\ &\xrightarrow{\text{Up}(128\to64)} \xrightarrow{\text{Res}(64)} \xrightarrow{\text{Up}(64\to32)} \xrightarrow{\text{Res}(32)} \\ &\xrightarrow{\text{Up}(32\to16)} \xrightarrow{\text{Res}(16)} \xrightarrow{\text{Conv}_{3\times3}} 3 \times 224 \times 224. \end{aligned} \tag{19}$$

Thus, the decoder contains five residual blocks and four upsampling blocks. The convolutional decoder blocks use BatchNorm2d and LeakyReLU with negative slope 0.2; each residual block applies an activation after adding the residual connection. The final 3×3 convolution has no output activation. Its output is subsequently clipped to $[-\epsilon, \epsilon]$, and the perturbed image is projected to the valid image range according to Eq. (7).

Optimization configuration. Unless otherwise specified, both MSCOCO and Flickr30K models are trained for 100 epochs with mini-batch size 32. We use AdamW with learning rate 10^{-4}, weight decay 0, and the default momentum parameters $(\beta_1, \beta_2) = (0.9, 0.999)$. The learning rate remains constant throughout training; no learning-rate scheduler is used. All reported checkpoints are trained with random seed 42. Model selection is performed using the validation set, and the checkpoint with the lowest validation loss is retained for evaluation.

Training procedure. Each mini-batch contains attack tasks $\tau = (x, T)$. The frozen CLIP encoders first compute target-image and query features, from which the condition encoder produces c_τ. The decoder then generates the bounded adversarial image x'. We re-encode x' with the frozen CLIP image encoder and optimize the bank-aware query-set ranking objective in Eq. (8). Only the condition encoder and perturbation decoder are updated.

Algorithm 2 summarizes the training procedure.

Table 4: Online attack-generation latency on Flickr30K and MSCOCO TopK-32 test tasks. All methods are measured under the same unified protocol on a single NVIDIA GeForce RTX 4090 GPU with native batch size 1. Lower is better.

Method	Flickr30K (ms/task) ↓	MSCOCO (ms/task) ↓
UAP	2.03	2.09
I-Only	4.53	4.26
Q-Only	5.61	5.58
MARP	7.25	7.20
PGD-10	65.95	72.23

AnyAttack adaptation. The released AnyAttack decoder generates a perturbation from an image-based semantic reference, whereas our task $\tau = (x, T)$ contains a promotion target x and a query set T induced by seed image x_s. We therefore encode x_s as AnyAttack's semantic reference, feed its CLIP feature to the pretrained decoder, clip the resulting perturbation to $[-\epsilon, \epsilon]$, and add it to the promotion target x before projecting the image to $[0, 1]$. Thus, AnyAttack receives the seed image as a proxy for the query-set semantics but does not receive the query texts directly. We use the released dataset-specific `bi` and `cos` decoder checkpoints without fine-tuning and denote them AA-bi and AA-cos, respectively. Both variants are evaluated on the same frozen TopK-32 tasks, CLIP ViT-B/32 retrieval backbone, image bank, target-replacement protocol, perturbation budget, and HR/MR metrics as MARP.

A.3 Attack Generation Time Measurement

We measure online attack-generation latency under a unified protocol on a single NVIDIA GeForce RTX 4090 GPU. All methods are evaluated with native batch size 1. For each method and dataset, we first execute 64 warm-up tasks and then measure 512 test tasks. The measurement is repeated three times using the same ordered task subset, and we report the mean latency per task over the timed executions. GPU execution time is measured using CUDA events with explicit synchronization.

The timed region is defined consistently across attack paradigms. For amortized or generator-based attacks, it includes all online condition encoding required by the method, generator execution, perturbation projection, adversarial-image construction, and final adversarial CLIP image encoding. For MARP, this includes both the target-image CLIP encoding and the query-set CLIP text encoding before the generator forward pass. For iterative attacks such as PGD, the timed region includes online condition encoding, all forward/backward computations and perturbation-update steps required by the attack objective, perturbation projection, and final adversarial CLIP image encoding.

We exclude one-time model and checkpoint initialization, task and image loading, image resizing and cropping, text tokenization, pre-computation of the retrieval image bank, final retrieval ranking and metric computation, serialization, and disk I/O. Image-bank computations that are required inside an attack objective, such as the bank-aware ranking loss used by PGD, remain inside the timed region. Thus, the reported latency isolates the online computation required to generate an adversarial example rather than the end-to-end retrieval-system latency.

Table 4 reports the resulting latency. Across both datasets, the amortized methods require only a few milliseconds per task. MARP requires 7.25 ms on Flickr30K and 7.20 ms on MSCOCO, whereas PGD-10 requires 65.95 ms and 72.23 ms, respectively. This corresponds to approximately $9.1\times$ and $10.0\times$ faster online attack generation for MARP on the two datasets.

B Additional Theoretical Analysis

This section provides the proof of the rank-surrogate connection in Proposition 1 and an idealized sampling interpretation of the query-set average used by the ranking objective.

B.1 Proof of Proposition 1

Proof. For query q_i, define

$$u_{ij} = s(q_i, x_j) - s(q_i, x') = z_i^\top v_j - z_i^\top v'.$$

By ℓ_2 normalization, $u_{ij} \in [-2, 2]$. For every such u_{ij},

$$\mathbb{I}[u_{ij} > 0] \le h(\mathrm{softplus}(u_{ij}))\,.$$

Indeed, if $u_{ij} > 0$, then $\mathrm{softplus}(u_{ij}) > \mathrm{softplus}(0) = c_0$, and the right-hand side equals 1. If $u_{ij} \le 0$, the indicator equals 0, while $\mathrm{softplus}(u_{ij}) \ge \mathrm{softplus}(-2) = c_-$ ensures that the right-hand side is nonnegative.

The function h is nondecreasing and concave on $[c_-, \mathrm{softplus}(2)]$. Therefore, using the rank definition and Jensen's inequality,

$$\begin{aligned} R_i(x') - 1 &= \sum_{x_j \in \mathcal{D}^-(x)} \mathbb{I}[u_{ij} > 0] \\ &\le \sum_{x_j \in \mathcal{D}^-(x)} h(\mathrm{softplus}(u_{ij})) \\ &\le N h\left(\frac{1}{N} \sum_{x_j \in \mathcal{D}^-(x)} \mathrm{softplus}(u_{ij}) \right) \\ &= N h(\mathcal{L}_i(x', q_i))\,. \end{aligned}$$

This proves Eq. (10). Applying Jensen's inequality once more across the task queries gives

$$\begin{aligned} \frac{1}{|T|} \sum_{q_i \in T} \big(R_i(x') - 1\big) &\le \frac{N}{|T|} \sum_{q_i \in T} h(\mathcal{L}_i(x', q_i)) \\ &\le N h\left(\frac{1}{|T|} \sum_{q_i \in T} \mathcal{L}_i(x', q_i) \right) \\ &= N h(\mathcal{L}_{\mathrm{rank}}(\tau; \theta))\,, \end{aligned}$$

which proves Eq. (11). □

The normalization-aware form removes the artificial $\mathrm{softplus}(-2)/\ln 2 \approx 0.183$ floor of the simpler uncentered softplus bound. Nevertheless, it remains a worst-case certificate rather than a calibrated numerical estimate of rank. In particular, a direct sufficient condition for the bound to certify $R_i(x') \le k$ is

$$\mathcal{L}_i(x', q_i) - c_- \le \frac{k-1}{N}(c_0 - c_-),$$

which becomes increasingly stringent as the retrieval bank grows. The bound is therefore primarily used to justify the ranking surrogate's optimization direction, while its empirical relationship with retrieval outcomes is evaluated separately.

B.2 Query-Set Risk Under an Idealized Sampling Model

We provide an idealized sampling interpretation of the query-set average in Eq. (8). The analysis considers a fixed adversarial image and assumes that the evaluation queries are sampled independently from an underlying local query distribution. These assumptions are deliberately stronger than our experimental setting: MARP generates x' as a function of the query set, and our TopK and radius-defined query sets are deterministic. Accordingly, the result below should be viewed only as an interpretation of query averaging, not as a generalization guarantee for the learned attack.

Let $\mathcal{Q}_\tau$ denote an idealized local query distribution associated with task τ. For a fixed adversarial image x', write the per-query loss in Eq. (8) in query-index-free form as

$$\mathcal{L}(x', q) = \frac{1}{|\mathcal{D}^-(x)|} \sum_{x_j \in \mathcal{D}^-(x)} \mathrm{softplus}\big(E_t(q)^\top v_j - E_t(q)^\top E_v(x')\big)\,. \tag{20}$$

Thus, $\mathcal{L}(x', q_i) = \mathcal{L}_i(x', q_i)$ in Eq. (8).

The corresponding population risk is

$$\mathcal{R}_{\mathcal{Q}_\tau}(x') = \mathbb{E}_{q\sim\mathcal{Q}_\tau}\left[\mathcal{L}(x', q)\right]. \tag{21}$$

For an i.i.d. query sample $T = \{q_1, \ldots, q_m\}$ from $\mathcal{Q}_\tau$, define

$$\widehat{\mathcal{R}}_T(x') = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(x', q_i). \tag{22}$$

Proposition 2 (Query-set concentration for a fixed adversarial image)**.** *Assume that all CLIP features are ℓ_2-normalized and that no additional logit scale is applied. Then*

$$\mathrm{softplus}(-2) \le \mathcal{L}(x', q) \le \mathrm{softplus}(2),$$

and the width of this normalization-implied interval is

$$\mathrm{softplus}(2) - \mathrm{softplus}(-2) = 2.$$

Let x' be fixed independently of $q_1, \ldots, q_m \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_\tau$. Then, with probability at least $1-\eta$,

$$\left|\widehat{\mathcal{R}}_T(x') - \mathcal{R}_{\mathcal{Q}_\tau}(x')\right| \le \sqrt{\frac{2\ln(2/\eta)}{m}}. \tag{23}$$

Proof. For each $x_j \in \mathcal{D}^-(x)$, define

$$u_j(q) = E_t(q)^\top v_j - E_t(q)^\top E_v(x').$$

Because all embeddings are ℓ_2-normalized, both inner products lie in $[-1, 1]$, and hence $u_j(q) \in [-2, 2]$. Monotonicity of the softplus function therefore gives

$$\mathrm{softplus}(-2) \le \mathrm{softplus}(u_j(q)) \le \mathrm{softplus}(2).$$

Averaging over the image bank preserves these bounds, so the same interval contains $\mathcal{L}(x', q)$. Moreover,

$$\mathrm{softplus}(2) - \mathrm{softplus}(-2) = \ln\frac{1+\exp(2)}{1+\exp(-2)} = 2.$$

For fixed x', the random variables $\mathcal{L}(x', q_1), \ldots, \mathcal{L}(x', q_m)$ are therefore i.i.d. and bounded in an interval of width 2. Hoeffding's inequality gives

$$\Pr\left(\left|\widehat{\mathcal{R}}_T(x') - \mathcal{R}_{\mathcal{Q}_\tau}(x')\right| \ge t\right) \le 2\exp\left(-\frac{mt^2}{2}\right).$$

Setting the right-hand side to η proves Eq. (23). □

Proposition 2 formalizes only the sampling intuition that averaging bounded per-query losses reduces estimation error at the usual $O(m^{-1/2})$ rate when the prediction is fixed independently of the queries. It does not apply directly to MARP's learned output $x' = G_\theta(x, T)$, which depends on the query set, nor to the deterministic TopK and radius-defined constructions used in our experiments. Generalization under those task constructions is therefore assessed empirically on held-out tasks and query-set regimes rather than claimed from this proposition.

C Component and Objective Ablations

This section isolates three design choices in MARP: the information exposed to the generator, the training objective, and the construction of the negative image bank used by the ranking surrogate.

C.1 TASK CONDITIONING

We first examine which components of an attack task $\tau = (x, T)$ should condition the perturbation generator. The compared variants differ only in the information supplied to the generator: UAP uses neither target-image nor query-set conditioning, I-Only uses the target image, Q-Only uses the query set, and MARP uses both. The query set remains part of the training objective for all task-specific variants, so this ablation isolates generator conditioning rather than changing the attack task itself.

Table 5: Task-conditioning ablation at $\epsilon = 16/255$. The Image and Query columns indicate which task components condition the perturbation generator.

Dataset	Method	Image	Query	HR@1 ↑	HR@10 ↑	MR ↓
Flickr30K	UAP	No	No	16.12	48.17	58.01
	I-Only	Yes	No	12.84	44.63	49.89
	Q-Only	No	Yes	42.88	73.36	28.93
	MARP	Yes	Yes	**44.66**	**76.20**	**20.45**
MSCOCO	UAP	No	No	11.46	28.21	186.58
	I-Only	Yes	No	5.97	17.36	178.79
	Q-Only	No	Yes	51.32	69.84	75.70
	MARP	Yes	Yes	**56.78**	**72.97**	**59.87**

Query conditioning accounts for a substantial fraction of the amortized promotion strength on both datasets: Q-Only is markedly stronger than the task-agnostic UAP and I-Only variants. Adding target-image conditioning further improves all reported aggregate metrics relative to Q-Only. The incremental gain is nevertheless smaller than the gain from introducing query conditioning and, as analyzed in Appendix F.1, varies with dataset, metric, and task difficulty. These results support the target image and query set as complementary task information without implying that their contributions are uniform across task regimes.

C.2 TRAINING OBJECTIVE

We next hold the generator architecture and task input (x, T) fixed and vary only the training objective. Let $z_i = E_t(q_i)$, $v' = E_v(x')$, and define the mean query embedding and its normalized direction as

$$\mu_T = \frac{1}{|T|} \sum_{q_i \in T} z_i, \qquad \bar{z}_T = \frac{\mu_T}{\|\mu_T\|_2}. \tag{24}$$

The mean-query similarity control maximizes similarity to the normalized query centroid:

$$\mathcal{L}_{\text{mean-sim}} = -\bar{z}_T^\top v'. \tag{25}$$

The query-set similarity control instead maximizes the average per-query similarity:

$$\mathcal{L}_{\text{qs-sim}} = -\frac{1}{|T|} \sum_{q_i \in T} z_i^\top v'. \tag{26}$$

All objective-ablation experiments use uniformly weighted TopK-32 queries. For $\mu_T \neq 0$, $\mathcal{L}_{\text{qs-sim}} = \|\mu_T\|_2 \mathcal{L}_{\text{mean-sim}}$: the two controls share the same attraction direction for a fixed task but differ in task-dependent scaling during training. Neither objective uses the retrieval bank. The default bank-aware objective instead compares the promoted target with competing images through Eq. (8).

Table 6: Training-objective ablation at $\epsilon = 16/255$. All variants use the same generator architecture and task input (x, T).

Dataset	Objective	Bank	HR@1 ↑	HR@5 ↑	HR@10 ↑	MR ↓
Flickr30K	Mean-query similarity	No	44.61	67.74	75.54	21.86
	Query-set similarity	No	42.41	65.94	74.07	24.79
	Bank-aware ranking	Yes	**44.66**	**67.91**	**76.20**	**20.45**
MSCOCO	Mean-query similarity	No	55.40	67.11	71.95	63.53
	Query-set similarity	No	50.10	63.47	69.04	68.19
	Bank-aware ranking	Yes	**56.78**	**68.24**	**72.97**	**59.87**

The bank-aware ranking objective gives the strongest result on every reported metric for both datasets. The improvement over mean-query similarity is modest on Flickr30K and more pronounced on MSCOCO, while query-set similarity is weaker in both cases. Thus, the ablation supports explicitly modeling retrieval competitors as the default training objective, while also showing that simpler similarity-based objectives can remain competitive in some settings.

C.3 Negative-Bank Construction

The default objective in Eq. (8) averages over all non-target images. To compare alternative negative-bank constructions, let $\mathcal{S}_i \subseteq \mathcal{D}^-(x)$ denote the negative set used for query q_i and define

$$\ell_i(\mathcal{S}_i) = \begin{cases} \dfrac{1}{|\mathcal{S}_i|} \displaystyle\sum_{x_j \in \mathcal{S}_i} \operatorname{softplus}\big(z_i^\top v_j - z_i^\top v'\big), & |\mathcal{S}_i| > 0, \\ 0, & |\mathcal{S}_i| = 0, \end{cases} \qquad \mathcal{L}_{\mathcal{S}} = \frac{1}{|T|} \sum_{q_i \in T} \ell_i(\mathcal{S}_i). \tag{27}$$

The full-bank configuration uses $\mathcal{S}_i = \mathcal{D}^-(x)$. The violating-only configuration uses the query-specific set of images that currently outrank the adversarial target:

$$\mathcal{V}_i(x') = \big\{x_j \in \mathcal{D}^-(x) : z_i^\top v_j > z_i^\top v'\big\}, \qquad \mathcal{S}_i = \mathcal{V}_i(x'). \tag{28}$$

This set is recomputed from the current adversarial feature at each forward pass; if it is empty, the corresponding per-query loss is zero as specified in Eq. (27).

For the random control, a size-B subset is sampled uniformly without replacement for each task-level loss evaluation and shared across its queries:

$$\mathcal{R}^B \sim \operatorname{Unif}\big\{\mathcal{S} \subseteq \mathcal{D}^-(x) : |\mathcal{S}| = B\big\}, \qquad \mathcal{S}_i = \mathcal{R}^B. \tag{29}$$

For top-B hard negatives, $\mathcal{H}_i^B$ contains the B highest-scoring non-target images for query q_i:

$$\begin{gathered} \mathcal{H}_i^B \subseteq \mathcal{D}^-(x), \qquad |\mathcal{H}_i^B| = B, \\ \min_{x_j \in \mathcal{H}_i^B} z_i^\top v_j \geq \max_{x_k \in \mathcal{D}^-(x) \setminus \mathcal{H}_i^B} z_i^\top v_k, \qquad \mathcal{S}_i = \mathcal{H}_i^B. \end{gathered} \tag{30}$$

Table 7: Effect of negative-bank construction at $\epsilon = 16/255$. Random negatives provide a size-matched control, while hard-negative variants retain the highest-scoring competitors. Full bank denotes the default objective used by MARP.

Dataset	Negative Set	B	HR@1 ↑	HR@5 ↑	HR@10 ↑	MR ↓
Flickr30K	Violating-only	–	36.42	62.60	71.90	24.35
	Random	32	37.56	62.43	71.48	25.60
	Hard	16	44.31	67.83	75.90	21.66
		32	44.37	67.78	75.85	22.22
		64	40.57	63.61	71.85	28.36
		128	39.45	61.95	70.15	30.38
		256	40.81	64.04	72.32	26.76
	Full bank	–	**44.66**	**67.91**	**76.20**	**20.45**
MSCOCO	Violating-only	–	46.62	62.38	68.69	60.96
	Random	32	53.81	66.01	71.01	64.66
	Hard	16	54.58	66.49	71.35	67.38
		32	50.51	64.16	69.74	60.74
		64	53.09	65.34	70.47	63.46
		128	50.25	63.19	68.68	71.78
		256	54.58	66.13	71.00	67.69
	Full bank	–	**56.78**	**68.24**	**72.97**	**59.87**

The full-bank objective gives the strongest aggregate performance on both datasets. On Flickr30K, the $B = 16$ and $B = 32$ hard-negative variants approach the full-bank result closely and substantially outperform the random control. On MSCOCO, restricted hard-negative variants are less consistent across metrics and none uniformly matches the full bank. These results indicate that high-scoring competitors contain useful supervision, but restricting the negative bank is not uniformly equivalent to the full-bank objective. We therefore retain the full image bank as the default training configuration.

D Full Surrogate–Outcome Analysis

Section 5.3.3 summarizes the task-level association between improvement in the optimized surrogate and improvement in retrieval outcomes. Here we report the complete correlation analysis, including confidence intervals and multiplicity-corrected significance tests.

Analysis protocol. The test population contains two target tasks for each seed image. Because these tasks share the same seed-induced query region, we first average the two targets within each seed and use the seed as the statistical unit. Retrieval improvement is oriented so that larger values always indicate stronger promotion: HR@1 is represented by its gain over the clean task, whereas mean rank is represented by its reduction from the clean task.

We report both raw Spearman correlation and partial Spearman correlation after controlling for the canonical clean mean rank. Confidence intervals are obtained from 10,000 seed-cluster bootstrap resamples with random seed 42, and Holm correction is applied jointly over the eight dataset–outcome–correlation tests.

Across both datasets, larger surrogate improvements are consistently associated with stronger retrieval improvements. After controlling for clean task difficulty, the partial Spearman correlations remain positive for both HR@1 gain and mean-rank reduction. The association is strongest for HR@1 gain ($\rho = 0.821$ on MSCOCO and $\rho = 0.782$ on Flickr30K), while mean-rank reduction shows more moderate but still positive associations ($\rho = 0.413$ and $\rho = 0.516$, respectively).

These results show that variation in surrogate improvement tracks variation in held-out retrieval outcomes beyond what is explained by clean mean rank alone. However, the analysis exploits naturally occurring variation across frozen test tasks and does not manipulate surrogate improvement

Table 8: Seed-level Spearman association between decrease in the dataset-specific training-objective component and retrieval improvement. Two target tasks are averaged per seed; partial correlations control for clean mean rank. CIs use 10,000 seed-cluster resamples (seed 42), and Holm correction covers all eight registered tests. Association is not causation.

Dataset	Outcome	ρ type	ρ	95% CI	raw p	Holm p
MSCOCO	HR@1 gain	Partial	0.821^{**}	[0.811, 0.831]	$< .001$	0.002
		Raw	0.772^{**}	[0.759, 0.783]	$< .001$	0.002
	MR reduction	Partial	0.413^{**}	[0.393, 0.433]	$< .001$	0.002
		Raw	0.319^{**}	[0.293, 0.344]	$< .001$	0.002
Flickr30K	HR@1 gain	Partial	0.782^{**}	[0.753, 0.808]	$< .001$	0.002
		Raw	0.696^{**}	[0.660, 0.730]	$< .001$	0.002
	MR reduction	Partial	0.516^{**}	[0.473, 0.558]	$< .001$	0.002
		Raw	0.420^{**}	[0.367, 0.472]	$< .001$	0.002

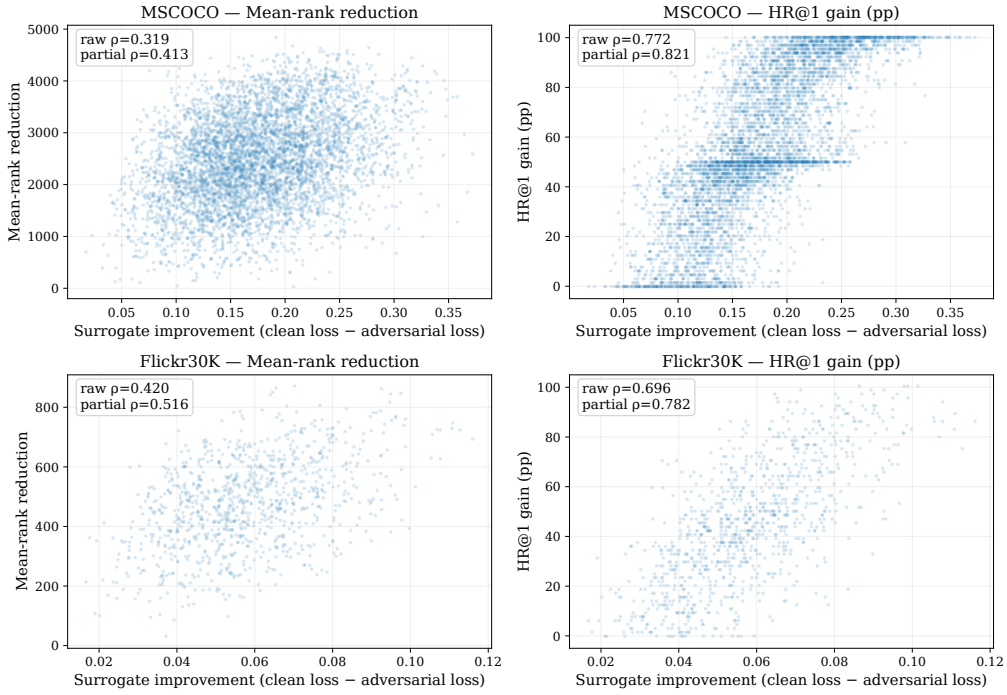

Figure 4: Seed-level association between improvement in the optimized task surrogate and retrieval improvement. Each point represents one seed after averaging its two target tasks. The relationship is observational and should not be interpreted causally.

independently. Accordingly, it supports an association between optimization progress and retrieval progress, not a causal effect of the surrogate on attack success.

E Generalization Across Query-Set Regimes

The default MARP checkpoint is trained on fixed-cardinality TopK-32 tasks. We evaluate how this frozen checkpoint behaves under two alternative query-set regimes without retraining: source-anchored nested TopK sets and radius-defined adaptive sets. Because both protocols change query membership together with cardinality, they characterize generalization across query-set regimes rather than an isolated causal effect of query-set size.

Table 9: Descriptive statistics across query-set sizes. HR values are percentages; MR is the mean rank (lower is better). Cohesion is the mean pairwise cosine similarity among the queries in a task (K=1 uses the explicit singleton convention, with cohesion = 1 by definition and standard deviation 0). Seed–query alignment is the mean cosine similarity between the seed-image embedding and the text embeddings of all queries in the evaluated query set T. Cohesion and alignment are descriptive; association with attack success is not causal.

Dataset	K	Cohesion	Seed–query align.	HR@1 (%)	HR@5 (%)	HR@10 (%)	MR
MSCOCO	1	1.0000	0.1462	54.05	68.44	72.97	67.98
	4	0.7906	0.1491	51.76	66.61	71.80	69.93
	8	0.8100	0.1524	54.80	68.09	73.06	64.06
	16	0.8227	0.1545	56.92	68.92	73.65	60.27
	32	0.8195	0.1560	56.78	68.24	72.97	59.87
	64	0.8065	0.1571	54.70	66.22	71.10	63.14
	128	0.7869	0.1582	50.92	62.78	67.86	69.46
Flickr30K	1	1.0000	0.1683	53.90	79.35	84.25	16.66
	4	0.7480	0.1692	51.24	76.58	82.69	18.85
	8	0.7469	0.1745	51.52	75.35	81.63	18.32
	16	0.7476	0.1787	48.98	72.26	79.52	18.84
	32	0.7426	0.1814	44.66	67.91	76.20	20.45
	64	0.7334	0.1829	39.01	63.12	72.13	22.95
	128	0.7212	0.1841	32.45	57.01	67.03	26.37

E.1 Source-Anchored Nested Query Sets

We use the source-anchored construction described in Appendix A.1. Starting from the frozen TopK-32 task, we evaluate

$$K \in \{1, 4, 8, 16, 32, 64, 128\}$$

while keeping the checkpoint, seed image, target image, image bank, and perturbation budget fixed. For $K < 32$, the query set is a canonical prefix of the frozen TopK-32 set; for $K > 32$, all TopK-32 members are retained and the nearest outside queries are appended.

For descriptive analysis, we additionally report query coherence,

$$\mathrm{Coh}(T) = \begin{cases} 1, & |T| = 1, \\ \dfrac{2}{m(m-1)} \displaystyle\sum_{i<j} E_t(q_i)^\top E_t(q_j), & m = |T| > 1, \end{cases} \tag{31}$$

together with the seed–query alignment between the seed image and all queries in the selected query set,

$$\mathrm{Align}(x_s, T) = \frac{1}{m} \sum_{q_i \in T} E_t(q_i)^\top E_v(x_s). \tag{32}$$

These quantities are descriptive diagnostics and are not used to infer causal relationships with attack success.

The descriptive results show that the most favorable evaluated query-set regime is dataset dependent. On MSCOCO, performance is strongest around the intermediate $K = 16$–32 regimes, whereas on Flickr30K promotion generally weakens as the nested set expands beyond the smaller-K conditions. Thus the checkpoint remains effective outside its training regime, but performance does not vary monotonically with K in a dataset-independent manner.

The paired analysis treats $K = 32$ as the reference training regime. The main post-hoc family compares $K \in \{8, 16, 64\}$ with $K = 32$ using seed-cluster bootstrap confidence intervals and Holm correction. The extended $K \in \{1, 4, 128\}$ comparisons are reported descriptively with confidence intervals only. Because changing K changes both query cardinality and query composition, these effects should be interpreted as differences between source-anchored query-set regimes rather than as isolated effects of cardinality.

Table 10: Main analysis: $K \in \{8, 16, 64\}$ vs the trained condition $K{=}32$ (*post-hoc* paired cluster bootstrap; cluster = seed image; 10,000 resamples; seed 42; single Holm family across 24 comparisons (3 K × 2 datasets × 4 metrics)). The reported effect $\Delta{=}$`k32_advantage_effect` is defined as $\mathrm{metric}(K{=}32) - \mathrm{metric}(K)$ for higher-is-better metrics (HR) and $\mathrm{metric}(K) - \mathrm{metric}(K{=}32)$ for lower-is-better metrics (MR); positive Δ always favors the trained condition $K{=}32$. Units: HR in percentage points (pp), MR in rank.

Dataset	K	Metric	Δ	95% CI	Win rate	raw p	Holm p
MSCOCO	8	adv HR@1 (pp)	1.98	[1.68, 2.29]	0.546	0.0002	0.0048
		adv HR@5 (pp)	0.15	[-0.13, 0.42]	0.486	0.3	0.7655
		adv HR@10 (pp)	-0.10	[-0.35, 0.16]	0.478	0.4676	0.7655
		adv MR (rank)	4.18	[2.85, 5.55]	0.459	0.0002	0.0048
	16	adv HR@1 (pp)	-0.14	[-0.31, 0.03]	0.508	0.12	0.48
		adv HR@5 (pp)	-0.68	[-0.83, -0.52]	0.478	0.0002	0.0048
		adv HR@10 (pp)	-0.69	[-0.83, -0.54]	0.474	0.0002	0.0048
		adv MR (rank)	0.40	[-0.29, 1.08]	0.464	0.2552	0.7655
	64	adv HR@1 (pp)	2.08	[1.93, 2.23]	0.598	0.0002	0.0048
		adv HR@5 (pp)	2.02	[1.88, 2.16]	0.607	0.0002	0.0048
		adv HR@10 (pp)	1.86	[1.73, 2.00]	0.605	0.0002	0.0048
		adv MR (rank)	3.27	[2.71, 3.84]	0.656	0.0002	0.0048
Flickr30K	8	adv HR@1 (pp)	-6.86	[-7.76, -5.99]	0.366	0.0002	0.0048
		adv HR@5 (pp)	-7.44	[-8.13, -6.75]	0.268	0.0002	0.0048
		adv HR@10 (pp)	-5.43	[-6.01, -4.86]	0.297	0.0002	0.0048
		adv MR (rank)	-2.13	[-3.07, -1.16]	0.194	0.0002	0.0048
	16	adv HR@1 (pp)	-4.31	[-4.81, -3.82]	0.347	0.0002	0.0048
		adv HR@5 (pp)	-4.34	[-4.77, -3.93]	0.303	0.0002	0.0048
		adv HR@10 (pp)	-3.32	[-3.68, -2.95]	0.340	0.0002	0.0048
		adv MR (rank)	-1.62	[-2.10, -1.13]	0.253	0.0002	0.0048
	64	adv HR@1 (pp)	5.65	[5.26, 6.03]	0.754	0.0002	0.0048
		adv HR@5 (pp)	4.79	[4.45, 5.13]	0.756	0.0002	0.0048
		adv HR@10 (pp)	4.07	[3.75, 4.38]	0.748	0.0002	0.0048
		adv MR (rank)	2.50	[2.09, 2.89]	0.799	0.0002	0.0048

E.2 Radius-Defined Adaptive Query Sets

Using the radius-defined construction in Eq. (15), we evaluate the same frozen TopK-32 checkpoint under adaptive query sets without retraining. Unlike the nested TopK protocol, the radius construction produces variable-cardinality query sets and changes membership according to a semantic-distance threshold.

The radius grids are selected without using attack outcomes. Because the same numerical radius induces substantially different cardinality distributions on MSCOCO and Flickr30K, we use dataset-specific grids calibrated to a sequence of median query-set sizes:

$$\text{MSCOCO:} \quad \rho \in \{0,\ 0.09,\ 0.115,\ 0.135,\ 0.16,\ 0.185\},$$
$$\text{Flickr30K:} \quad \rho \in \{0,\ 0.15,\ 0.185,\ 0.21,\ 0.23,\ 0.25\}.$$

We refer to these conditions as seed-only, median-$K8$, median-$K16$, median-$K32$, median-$K64$, and median-$K128$. The labels indicate the calibrated median cardinality and do not imply a fixed number of queries for every task.

The radius-defined evaluation exhibits strongly dataset-dependent behavior. On MSCOCO, every evaluated radius condition is weaker than the frozen TopK-32 reference on both reported effects. On Flickr30K, the smaller-radius conditions improve HR@1 relative to TopK-32, whereas the effect reverses as the radius expands. Thus, behavior under radius-defined sets differs qualitatively across the two datasets and from the source-anchored nested TopK regimes.

These results should not be interpreted as causal effects of query-set cardinality. Radius expansion jointly changes query identity, composition, and size, and the resulting cardinality distributions are strongly right-skewed. For example, the MSCOCO condition calibrated to median $K \approx 32$ has me-

Table 11: Extended-range analysis (appendix; outside the Holm family): $K \in \{1, 4, 128\}$ vs K=32. Point effect and 95% CI only; no p-values are reported by design.

Dataset	K	Metric	Δ	95% CI	Win rate
MSCOCO	1	adv HR@1 (pp)	2.73	[2.04, 3.44]	0.459
		adv HR@5 (pp)	-0.20	[-0.81, 0.44]	0.410
		adv HR@10 (pp)	-0.00	[-0.58, 0.58]	0.406
		adv MR (rank)	8.11	[4.92, 11.40]	0.309
	4	adv HR@1 (pp)	5.02	[4.58, 5.47]	0.558
		adv HR@5 (pp)	1.63	[1.23, 2.03]	0.485
		adv HR@10 (pp)	1.17	[0.80, 1.54]	0.472
		adv MR (rank)	10.06	[8.07, 12.14]	0.443
	128	adv HR@1 (pp)	5.86	[5.60, 6.11]	0.678
		adv HR@5 (pp)	5.47	[5.23, 5.71]	0.702
		adv HR@10 (pp)	5.10	[4.88, 5.33]	0.703
		adv MR (rank)	9.59	[8.65, 10.56]	0.791
Flickr30K	1	adv HR@1 (pp)	-9.24	[-11.03, -7.46]	0.416
		adv HR@5 (pp)	-11.44	[-12.76, -10.10]	0.250
		adv HR@10 (pp)	-8.05	[-9.22, -6.82]	0.256
		adv MR (rank)	-3.80	[-5.63, -1.82]	0.120
	4	adv HR@1 (pp)	-6.58	[-7.80, -5.39]	0.414
		adv HR@5 (pp)	-8.66	[-9.60, -7.73]	0.270
		adv HR@10 (pp)	-6.49	[-7.29, -5.70]	0.281
		adv MR (rank)	-1.60	[-2.93, -0.16]	0.185
	128	adv HR@1 (pp)	12.21	[11.60, 12.83]	0.805
		adv HR@5 (pp)	10.90	[10.36, 11.45]	0.849
		adv HR@10 (pp)	9.17	[8.69, 9.66]	0.862
		adv MR (rank)	5.92	[5.34, 6.49]	0.884

Table 12: Paired effects of radius-defined evaluation relative to the frozen TopK-32 condition. Positive values favor the radius condition. HR@1 effects are percentage points; MR effects are oriented so that positive values indicate lower mean rank under the radius condition.

Dataset	Radius condition	ρ	ΔHR@1	ΔMR
MSCOCO	Seed-only	0	-5.13	-9.41
	Median-$K8$	0.090	-1.88	-4.43
	Median-$K16$	0.115	-1.19	-2.56
	Median-$K32$	0.135	-1.39	-2.49
	Median-$K64$	0.160	-3.02	-4.22
	Median-$K128$	0.185	-6.64	-9.31
Flickr30K	Seed-only	0	+6.66	+1.23
	Median-$K8$	0.150	+6.02	+1.62
	Median-$K16$	0.185	+1.70	+0.32
	Median-$K32$	0.210	-2.63	-1.41
	Median-$K64$	0.230	-7.11	-4.05
	Median-$K128$	0.250	-11.68	-6.75

dian $|T| = 31$, mean $|T| = 65.39$, and maximum $|T| = 400$; the corresponding Flickr30K condition has median $|T| = 34.5$, mean $|T| = 60.92$, and maximum $|T| = 406$. The experiment therefore supports a narrower conclusion: the generalization behavior of a TopK-32-trained checkpoint is sensitive to how the semantic query region is constructed.

F Robustness Across Task Difficulty and Perturbation Budget

We further examine two factors that can alter the operating regime of the attack: the clean difficulty of the promotion task and the available ℓ_∞ perturbation budget. Difficulty analysis uses the frozen

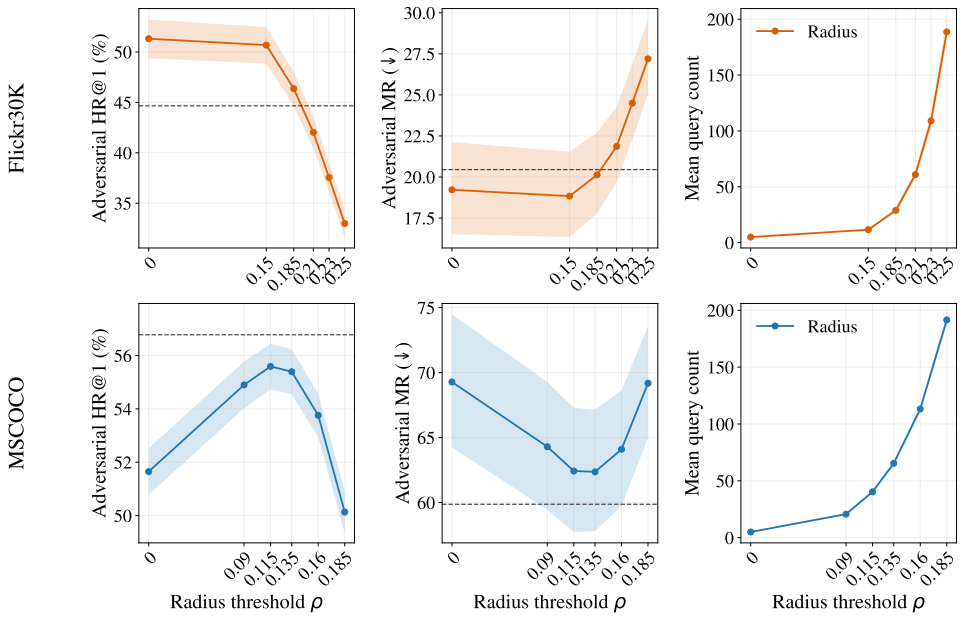

Figure 5: Promotion under radius-defined adaptive query sets using the frozen TopK-32 MARP checkpoint. Changing ρ alters both query-set cardinality and composition, so the experiment measures robustness to query-set construction rather than a causal effect of cardinality.

TopK-32 test population, while the budget analysis compares independently trained checkpoints under different perturbation constraints.

F.1 Task-Difficulty Stratification

We examine whether the incremental value of task conditioning varies with the clean difficulty of the promotion task. For each frozen TopK-32 test task, we define difficulty by the clean mean target rank

$$D_{\text{rank}}(\tau) = \frac{1}{|T|} \sum_{q_i \in T} R_i(x), \tag{33}$$

where larger values indicate that the target is more difficult to retrieve before attack. Within each dataset, tasks are partitioned once into Easy, Medium, and Hard tertiles according to D_{rank}. This assignment is frozen and reused for every compared method.

The full stratified results show that query-conditioned attacks remain much stronger than the task-agnostic UAP and I-Only variants across the three difficulty strata. The gap between MARP and Q-Only is substantially smaller, so we use paired within-stratum comparisons to isolate the incremental effect of additionally conditioning on the target image.

At the default budget $\epsilon = 16/255$, the paired results are dataset- and metric-dependent. On MSCOCO, MARP obtains significant HR@1 and mean-rank advantages over Q-Only in all three difficulty strata. Its HR@1 advantage remains similar across strata (5.20–5.72 percentage points), whereas its mean-rank advantage increases from 5.18 ranks on Easy tasks to 28.37 ranks on Hard tasks. On Flickr30K, the mean-rank advantage is significant in all three strata and is largest on Hard tasks; the HR@1 advantage is significant on Easy and Hard tasks but not on Medium tasks. These results suggest that target-image conditioning consistently improves ranking depth, while its incremental HR@1 benefit is smaller and less uniform on Flickr30K.

To examine whether the MARP–Q-Only difference itself changes systematically from Easy to Hard tasks, we additionally report Hard-minus-Easy interaction contrasts in Table 15.

As shown in Table 15, the positive MR interactions indicate that MARP's mean-rank advantage increases from Easy to Hard tasks relative to all three comparators on both datasets. In contrast, its

Table 13: Difficulty-stratified frozen TopK-32 performance. Difficulty tertiles are defined once from canonical clean mean rank. Brackets are seed-cluster bootstrap 95% CIs (10,000 resamples, seed 42); HR@1 is in percentage points.

Dataset	Stratum	Method	Adv. HR@1 ↑	Adv. MR ↓
MSCOCO	Easy	UAP	13.35 [12.55, 14.18]	106.22 [100.11, 112.70]
		I-Only	7.10 [6.49, 7.74]	122.73 [117.79, 127.98]
		Q-Only	55.05 [53.76, 56.37]	36.58 [33.48, 39.87]
		MARP	60.78 [59.49, 62.09]	31.40 [28.26, 34.69]
	Medium	UAP	11.02 [10.29, 11.79]	179.06 [167.94, 190.44]
		I-Only	5.76 [5.24, 6.32]	176.43 [168.21, 184.64]
		Q-Only	50.68 [49.36, 52.04]	72.26 [65.08, 79.72]
		MARP	56.14 [54.78, 57.50]	58.30 [51.96, 65.15]
	Hard	UAP	9.93 [9.21, 10.67]	275.29 [258.04, 293.32]
		I-Only	5.04 [4.53, 5.57]	237.24 [225.63, 248.99]
		Q-Only	48.21 [46.81, 49.58]	118.29 [106.86, 130.44]
		MARP	53.42 [52.00, 54.80]	89.92 [79.75, 100.85]
Flickr30K	Easy	UAP	19.17 [17.36, 20.99]	33.98 [29.97, 38.27]
		I-Only	15.63 [14.20, 17.08]	28.14 [25.51, 31.04]
		Q-Only	48.04 [45.66, 50.36]	14.59 [12.39, 17.13]
		MARP	49.99 [47.54, 52.39]	11.42 [9.67, 13.53]
	Medium	UAP	15.22 [13.56, 16.91]	52.81 [47.12, 59.01]
		I-Only	12.50 [11.11, 13.92]	44.01 [40.12, 48.23]
		Q-Only	42.59 [40.15, 45.05]	21.96 [18.93, 25.34]
		MARP	43.90 [41.37, 46.43]	17.22 [14.86, 19.82]
	Hard	UAP	13.92 [12.33, 15.55]	87.19 [77.15, 97.68]
		I-Only	10.37 [9.12, 11.69]	77.56 [69.87, 85.75]
		Q-Only	38.04 [35.51, 40.65]	50.24 [42.16, 58.66]
		MARP	40.09 [37.48, 42.72]	32.73 [27.31, 38.57]

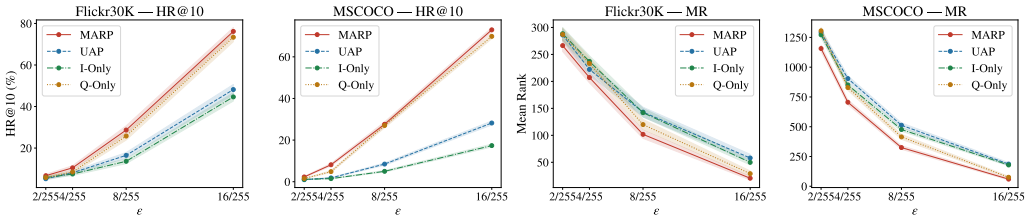

Figure 6: Perturbation-budget sensitivity on Flickr30K and MSCOCO. Each point evaluates an independently trained checkpoint under the indicated ℓ_∞ budget. Shaded regions denote 95% seed-cluster bootstrap confidence intervals from 10,000 resamples.

HR@1 advantage decreases from Easy to Hard relative to I-Only and UAP, whereas the corresponding HR@1 interactions relative to Q-Only are not statistically distinguishable from zero after Holm correction on either dataset. We treat these interaction results as exploratory evidence of metric- and comparator-dependent heterogeneity rather than as support for a universal monotonic relationship between clean task difficulty and the benefit of target-image conditioning.

F.2 Perturbation Budget Sensitivity

We next evaluate robustness across perturbation budgets $\epsilon \in \{2/255, 4/255, 8/255, 16/255\}$. Each budget uses an independently trained checkpoint under the corresponding ℓ_∞ constraint, while the TopK-32 attack-task construction and retrieval backbone remain fixed.

Across the fully reported configurations, larger perturbation budgets generally yield stronger promotion. For MARP specifically, HR@10 increases monotonically and mean rank decreases monotonically as ϵ grows on both datasets. On Flickr30K, MARP achieves the strongest reported performance

Table 14: Paired MARP advantages within clean-difficulty tertiles. Positive effects favor MARP; HR@1 effects are percentage points and MR effects are ranks. CIs resample seed clusters. Holm correction is over the prespecified 36-comparison family; $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$ after correction.

Dataset	Stratum	Comparator	Metric	Effect	95% CI	raw p	Holm p
MSCOCO	Easy	I-Only	HR@1	53.681**	[52.400, 54.959]	$< .001$	0.007
			MR	91.338**	[86.960, 95.860]	$< .001$	0.007
		Q-Only	HR@1	5.724**	[5.003, 6.438]	$< .001$	0.007
			MR	5.180**	[2.639, 7.630]	$< .001$	0.007
		UAP	HR@1	47.434**	[46.178, 48.699]	$< .001$	0.007
			MR	74.827**	[70.092, 79.687]	$< .001$	0.007
	Medium	I-Only	HR@1	50.372**	[49.068, 51.716]	$< .001$	0.007
			MR	118.129**	[111.719, 124.548]	$< .001$	0.007
		Q-Only	HR@1	5.453**	[4.714, 6.193]	$< .001$	0.007
			MR	13.958**	[9.159, 18.877]	$< .001$	0.007
		UAP	HR@1	45.114**	[43.852, 46.401]	$< .001$	0.007
			MR	120.759**	[112.645, 129.210]	$< .001$	0.007
	Hard	I-Only	HR@1	48.380**	[47.053, 49.712]	$< .001$	0.007
			MR	147.321**	[138.523, 156.000]	$< .001$	0.007
		Q-Only	HR@1	5.204**	[4.474, 5.934]	$< .001$	0.007
			MR	28.366**	[21.434, 35.699]	$< .001$	0.007
		UAP	HR@1	43.482**	[42.185, 44.757]	$< .001$	0.007
			MR	185.368**	[173.144, 198.443]	$< .001$	0.007
Flickr30K	Easy	I-Only	HR@1	34.361**	[32.372, 36.315]	$< .001$	0.007
			MR	16.724**	[14.652, 18.976]	$< .001$	0.007
		Q-Only	HR@1	1.954**	[0.653, 3.237]	0.003	0.007
			MR	3.170**	[1.921, 4.455]	$< .001$	0.007
		UAP	HR@1	30.819**	[28.890, 32.702]	$< .001$	0.007
			MR	22.564**	[19.395, 25.878]	$< .001$	0.007
	Medium	I-Only	HR@1	31.395**	[29.370, 33.533]	$< .001$	0.007
			MR	26.782**	[23.833, 29.897]	$< .001$	0.007
		Q-Only	HR@1	1.307	[-0.028, 2.643]	0.057	0.057
			MR	4.736**	[2.600, 6.969]	$< .001$	0.007
		UAP	HR@1	28.683**	[26.641, 30.734]	$< .001$	0.007
			MR	35.589**	[31.056, 40.511]	$< .001$	0.007
	Hard	I-Only	HR@1	29.716**	[27.653, 31.803]	$< .001$	0.007
			MR	44.821**	[40.310, 49.508]	$< .001$	0.007
		Q-Only	HR@1	2.046**	[0.778, 3.278]	0.002	0.007
			MR	17.505**	[12.392, 22.761]	$< .001$	0.007
		UAP	HR@1	26.173**	[24.276, 28.176]	$< .001$	0.007
			MR	54.459**	[47.202, 62.136]	$< .001$	0.007

within every evaluated budget block. On MSCOCO, MARP achieves the strongest reported performance among the evaluated amortized variants at every evaluated perturbation budget.

The comparison between MARP and Q-Only also indicates that the incremental value of target-image conditioning can depend on the available perturbation budget. This dependence is further analyzed through the difficulty stratification and the complete budget-specific evaluation presented in this section.

G Deterministically Selected Qualitative Cases

Qualitative examples can be sensitive to manual selection, so we fix the selection rule before inspecting the corresponding images. Within each dataset, we select the seed with the maximum seed-level relative MR reduction as the high-gain case, the remaining seed closest to the median as the typical-gain case, and the seed with the minimum reduction as the low-gain case. These labels indicate relative positions within each dataset's empirical gain distribution rather than absolute at-

Table 15: Hard-minus-Easy differences in MARP's paired advantage. Positive values indicate a larger MARP advantage on Hard tasks. Seed clusters are jointly resampled so seeds whose two targets occupy different strata remain dependent. Holm correction covers 12 tests. $^{**}p < .01$ after Holm correction.

Dataset	Comparator	Metric	Interaction	95% CI	raw p	Holm p
MSCOCO	I-Only	HR@1	-5.301**	[-7.147, -3.500]	$< .001$	0.002
		MR	55.982**	[46.311, 65.511]	$< .001$	0.002
	Q-Only	HR@1	-0.521	[-1.532, 0.489]	0.324	0.648
		MR	23.186**	[15.708, 30.752]	$< .001$	0.002
	UAP	HR@1	-3.952**	[-5.721, -2.141]	$< .001$	0.002
		MR	110.542**	[97.385, 124.026]	$< .001$	0.002
Flickr30K	I-Only	HR@1	-4.645**	[-7.396, -1.892]	$< .001$	0.003
		MR	28.096**	[23.061, 33.334]	$< .001$	0.002
	Q-Only	HR@1	0.092	[-1.706, 1.866]	0.901	0.901
		MR	14.334**	[9.151, 19.754]	$< .001$	0.002
	UAP	HR@1	-4.646**	[-7.229, -2.013]	$< .001$	0.002
		MR	31.895**	[23.833, 39.885]	$< .001$	0.002

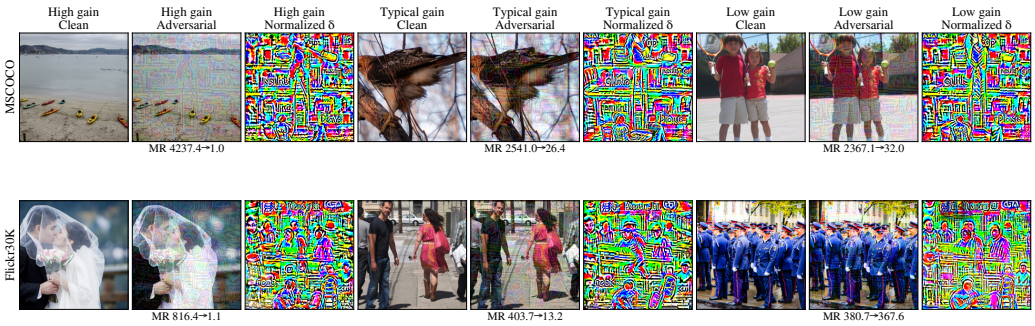

Figure 7: Deterministically selected high-gain, typical-gain, and low-gain cases on MSCOCO and Flickr30K. The labels indicate relative positions within each dataset's seed-level gain distribution. Case selection is fixed from retrieval outcomes before image inspection, and the displayed queries follow the frozen query ordering.

tack success or failure. For each selected seed, we mechanically choose the target whose outcome is closest to the seed-level mean, and display the first four queries in the frozen ordering. This protocol is designed to make the qualitative panel reflect the range of observed attack outcomes rather than only visually favorable examples.

The examples illustrate substantial heterogeneity across attack tasks: some targets are strongly promoted across the selected query region, whereas others show limited rank improvement. These cases are intended as descriptive illustrations of the quantitative results rather than as additional statistical evidence for attack effectiveness or failure modes.

Table 16: Promotion performance across perturbation budgets on TopK-32 tasks. HR@K is reported in percentage and MR denotes mean target rank. Bold indicates the best result within each dataset–budget block. All evaluated perturbations satisfy the corresponding ℓ_∞ budget.

Dataset	ϵ	Method	HR@1 ↑	HR@5 ↑	HR@10 ↑	MR ↓
Flickr30K	$2/255$	UAP	0.66	2.95	5.21	286.03
		I-Only	0.84	3.41	5.78	288.42
		Q-Only	0.87	3.39	5.81	287.03
		MARP	**1.03**	**4.05**	**6.70**	**266.39**
	$4/255$	UAP	1.14	4.61	8.10	222.16
		I-Only	1.13	4.36	7.50	237.00
		Q-Only	1.21	4.93	8.47	233.28
		MARP	**1.71**	**6.41**	**10.48**	**207.19**
	$8/255$	UAP	2.70	10.15	16.49	143.03
		I-Only	2.04	8.18	13.63	142.00
		Q-Only	6.34	17.85	25.77	119.95
		MARP	**7.64**	**20.40**	**28.68**	**101.82**
	$16/255$	UAP	16.12	37.40	48.17	58.01
		I-Only	12.84	32.95	44.63	49.89
		Q-Only	42.88	65.73	73.36	28.93
		MARP	**44.66**	**67.91**	**76.20**	**20.45**
MSCOCO	$2/255$	UAP	0.17	0.58	0.97	1290.94
		I-Only	0.23	0.71	1.15	1271.57
		Q-Only	0.33	0.97	1.55	1305.51
		MARP	**0.54**	**1.50**	**2.30**	**1156.65**
	$4/255$	UAP	0.39	1.14	1.84	904.35
		I-Only	0.28	0.87	1.47	850.96
		Q-Only	1.33	3.26	4.85	828.75
		MARP	**2.93**	**6.07**	**8.14**	**705.58**
	$8/255$	UAP	2.63	5.93	8.47	512.57
		I-Only	1.26	3.25	4.99	476.71
		Q-Only	13.96	22.39	27.01	414.61
		MARP	**15.06**	**23.16**	**27.59**	**325.14**
	$16/255$	UAP	11.46	21.77	28.21	186.58
		I-Only	5.97	12.53	17.36	178.79
		Q-Only	51.32	64.63	69.84	75.70
		MARP	**56.78**	**68.24**	**72.97**	**59.87**

Table 17: Mechanically selected qualitative cases. High-, typical-, and low-gain cases are determined from seed-level relative MR reduction before image inspection; one representative target is selected by distance to its seed-level outcome. The displayed queries are the first four in the frozen order, not hand-picked.

Dataset	Case	Task	Clean MR	Adv. MR	Rel. MR Red.	Δsur.
MSCOCO	High gain	task_008350	4237.44	1.00	99.98%	0.2907
Queries: the baseball player is hoping for a hit; a baseball player grabbing a bat in a game; a baseball player holding a bat near home plate; a batter readies to swing at first plate						
MSCOCO	Typical gain	task_005594	2540.97	26.44	98.96%	0.1139
Queries: a woman has headphones on and is holding two cell phones; a woman with beautiful breast sitting at a table; a man sitting at a table in a wheelchair while on a phone; a woman on a cell phone with a bag						
MSCOCO	Low gain	task_008206	2367.09	32.03	98.65%	0.1186
Queries: a man wearing a suit and a tie staring; a man wearing a grey suit and tie looking at something; a man with a tie and a suit; a man who is wearing a suit and tie						
Flickr30K	High gain	task_001978	816.38	1.09	99.87%	0.0967
Queries: a large group of people of various ages and genders sit outside together; some of the people in the crowd are having a drink; the peoples sitting in the chair; people sitting at tables at some sort of festival						
Flickr30K	Typical gain	task_001862	403.66	13.19	96.73%	0.0472
Queries: the man in the yellow pants is raising his arms; a child wearing a yellow shirt is jumping up and down; the woman in the brown shirt is sitting on a bright red bench; a skateboard is grinding on a curb with his skateboard						
Flickr30K	Low gain	task_001924	380.69	367.59	3.44%	0.0006
Queries: woman singing with a microphone next to a person playing guitar; a band performs in a dimly lit setting; a group of musicians are performing on a stage in front of a crowd; on stage photo of small band performing for theater audience						