PHASE: PACKET-LEVEL HIERARCHICAL ADVERSARIAL SEMANTIC EXPLOITATION FOR TARGETED STEERING OF RETRIEVAL-AUGMENTED DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented diffusion models (RDMs) enhance generation by conditioning the denoising process on a small set of retrieved examples. However, the dependence on external retrieved context introduces additional security risks at the retrieval–generation interface. Existing studies on RDMs security mainly focus on manipulating the knowledge base, retriever, or persistent trigger mechanism, while the security implications of the retrieved conditioning packet itself remain insufficiently studied. In this work, we investigate packet-level targeted steering in frozen RDMs and propose PHASE (Packet-level Hierarchical Adversarial Semantic Exploitation), a training-free and gradient-free adversarial framework that operates solely on the conditioning packet. Our central observation is that retrieval-space relevance is not a reliable indicator of generation-time influence: highly target-aligned retrieved examples do not necessarily form the packet that most effectively steers the downstream generation process. Based on this observation, PHASE combines heterogeneous candidate construction, source-aware semantic scoring, structured packet composition, and generation-aware probe selection to identify effective adversarial conditioning under a frozen generator. Extensive experiments across diverse generation settings, semantic evaluators, packet configurations, and retrieval conditions consistently demonstrate the effectiveness and robustness of the proposed framework. Further analyses reveal non-trivial interactions among packet members and a trade-off between retrieval compatibility and downstream steering effectiveness. These results show that security evaluation of retrieval-augmented generative models should jointly account for retrieval composition and generation-time behavior.

1 INTRODUCTION

Diffusion models have become a fundamental paradigm for image synthesis and multimodal generation (Ho et al., 2020; Song et al., 2020a; Rombach et al., 2022; Dhariwal & Nichol, 2021; Ramesh et al., 2022; Saharia et al., 2022). Retrieval-augmented diffusion models (RDMs) extend this paradigm with non-parametric visual memory, where a retriever selects external examples whose embeddings serve as conditioning signals during denoising (Blattmann et al., 2022; Sheynin et al., 2022). This mechanism improves model flexibility and reduces reliance on parametric knowledge. However, it also introduces an additional inference-time dependency: generation is determined not only by the model and user input, but also by the retrieved conditioning packet supplied to the generator.

Recent studies have begun to examine the security implications of this retrieval pathway. BadRDM manipulates retrieval components to associate triggers with attacker-specified retrieved content, while JOB poisons the knowledge base and jointly optimizes retrieval and generation in black-box RDMs (Fang et al., 2026; Lyu et al., 2026). These studies reveal vulnerabilities introduced by retrieval augmentation, yet they mainly focus on modifying how malicious context is retrieved through poisoned databases, retrievers, or persistent triggers. The conditioning packet after retrieval, namely the actual set of retrieved representations consumed by a frozen generator, remains comparatively underexplored. We therefore study targeted semantic steering in a complementary setting where both retriever and generator remain frozen and adversarial control is restricted to the retrieved packet. This setting is relevant to practical RDM pipelines in which retrieved context may pass through external

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.

retrieval services, middleware, caches, or orchestration layers before reaching the generator, creating an inference-time interface that can be influenced independently of model parameters.

However, this setting gives rise to a retrieval–generation challenge. Retrieval relevance alone does not fully determine generation because retrieved examples act jointly as conditioning signals during denoising (Blattmann et al., 2022; Sheynin et al., 2022). Prior work has also observed that retrieving target-related content does not necessarily guarantee target-aligned generation when retrieved evidence conflicts with other conditioning signals (Fang et al., 2026; Lyu et al., 2026). Consequently, simply replacing the packet with the most target-similar neighbors may be insufficient. Individual members may retain source-related evidence or provide redundant conditioning, while packet compositions can induce different downstream responses. Effective packet-level steering therefore requires considering query compatibility, target evidence, source suppression, packet composition, and the response of the frozen generator.

To study this retrieval–generation discrepancy, we propose PHASE (Packet-level Hierarchical Adversarial Semantic Exploitation), a training-free and gradient-free adversarial framework for packet-level targeted steering of frozen RDMs. PHASE treats retrieval-space relevance as a proposal signal rather than the final criterion for adversarial effectiveness. Instead of modifying the retriever or diffusion model, it operates directly on the conditioning packet and evaluates candidate packets according to their downstream semantic effects. Figure 1 illustrates this packet-level intervention at the retrieval–generation interface.

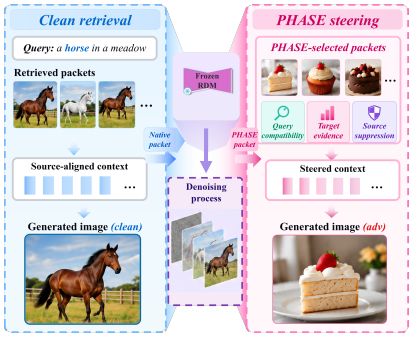


Figure 1: **PHASE at the RDM interface.** Native retrieval preserves the source concept, whereas a PHASE-selected packet redirects the same frozen RDM toward the attacker-specified target.

PHASE organizes packet construction as a progressive retrieval-to-generation process. Heterogeneous candidate construction broadens the search space with natural target examples, source/query anchors, source–target blends, localized target cues, and appearance-matched candidates. Source-aware semantic scoring then evaluates query compatibility, target alignment, directional shift, source suppression, and diversity. Structured packet composition further assembles candidates into multi-member proposals with complementary semantic roles, treating the packet as a joint conditioning unit rather than a set of independently ranked neighbors. Finally, generation-aware probe selection evaluates these proposals through the frozen RDM and retains the packet according to its realized downstream semantic effect. This hierarchy connects candidate-level retrieval evidence with packet-level generative influence without retriever fine-tuning, model updates, or diffusion gradients.

Our experiments evaluate both attack effectiveness and the mechanisms underlying PHASE. Across diverse source–target instances and generation configurations, PHASE substantially improves targeted generation over native retrieval and consistently outperforms its static counterpart. Controlled ablations isolate the contribution of generation-aware validation and show that retrieval-space scoring alone is insufficient for selecting the most effective packet. Independent vision–language evaluators, cross-seed and cross-sampler studies, and packet-size and subset analyses further examine evaluator dependence, transferability, and packet composition. Native-retrieval diagnostics additionally reveal that improving packet retrievability does not necessarily preserve generation-time steering strength. Together, these results support the central premise that packet-level vulnerability is governed jointly by retrieval composition and downstream generative response. We summarize our main contributions as follows:

- We propose PHASE, a training-free and gradient-free adversarial framework for targeted semantic steering of retrieval-augmented diffusion models. To the best of our knowledge, PHASE is the first packet-level framework to study targeted steering of frozen RDMs by manipulating the retrieved conditioning packet, without modifying the retriever, diffusion model, or model parameters.

- We develop a hierarchical packet construction strategy that integrates heterogeneous candidate construction, source-aware semantic scoring, structured multi-member packet composition, and generation-aware probe selection. In contrast to prior studies that primarily manipulate how malicious context is retrieved, PHASE operates directly on the conditioning packet consumed

by the generator and explicitly addresses the mismatch between retrieval-space relevance and downstream generative influence.

- We conduct extensive evaluations across generation settings, semantic evaluators, packet configurations, probe budgets, seeds, samplers, and native-retrieval conditions. The results demonstrate effective and transferable packet-level steering, validate the contribution of generation-aware selection through controlled ablations, characterize the role of packet composition, and identify the trade-off between packet retrievability and generation-time steering effectiveness.

2 Related Work

2.1 Retrieval-Augmented Diffusion Models

Diffusion models are widely used for image synthesis, with latent and large-scale text-conditioned variants further improving efficiency and generation quality (Ho et al., 2020; Song et al., 2020b;a; Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021; Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Podell et al., 2024). Retrieval-augmented diffusion models introduce external non-parametric memory by retrieving semantically relevant examples whose representations condition the denoising process (Blattmann et al., 2022; Sheynin et al., 2022). This coupling creates an additional inference-time dependency between retrieved context and generation. Our work studies the security implications of this conditioning interface.

2.2 Adversarial Security of Diffusion Models

Prior studies redirect diffusion models through poisoned training objectives, backdoor triggers, modified text encoders, or manipulated model components (Chou et al., 2023a; Chen et al., 2023; Chou et al., 2023b; Struppek et al., 2023; Zhai et al., 2023a). These methods primarily intervene in the model, training process, or prompt-side representation. In contrast, our setting leaves the generator unchanged and restricts adversarial manipulation to the conditioning packet supplied at inference time.

2.3 Security of Retrieval-Augmented Generative Systems

Retrieval augmentation introduces security risks because external memory can influence downstream behavior without directly modifying the core generator (Xue et al., 2024; Zou et al., 2025; Liu et al., 2023). RDM-specific studies remain limited. BadRDM learns malicious retrieval shortcuts and toxicity surrogates (Fang et al., 2026), while JOB poisons the knowledge base and jointly optimizes retrieval and target-aligned generation in black-box RDMs (Lyu et al., 2026). These approaches mainly manipulate how adversarial context is retrieved. PHASE instead studies the post-retrieval conditioning stage, where both retriever and generator remain frozen, enabling direct analysis of the mismatch between retrieval-space relevance and generation-time influence.

3 Method

Figure 2 illustrates the overall pipeline of PHASE. Given a query image, its source class, and an attacker-specified target class, PHASE constructs a heterogeneous candidate pool, evaluates candidate semantics in the native retrieval space, assembles structured multi-member packet proposals, and finally validates these proposals through the frozen generator. The resulting packet is selected according to its downstream semantic effect while both the retriever and diffusion model remain unchanged. We first formulate the packet-level steering problem and threat model, and then describe the four components of PHASE.

3.1 Problem Setup and Threat Model

An RDM consists of a retriever R, a knowledge base $\mathcal{D} = \{(x_i, e_i)\}_{i=1}^{N}$, and a frozen diffusion generator G_θ, where x_i denotes a database image and $e_i \in \mathbb{R}^d$ its native retrieval embedding. Given a query image x_q, let $q \in \mathbb{R}^d$ denote its native query embedding. The retriever returns a conditioning

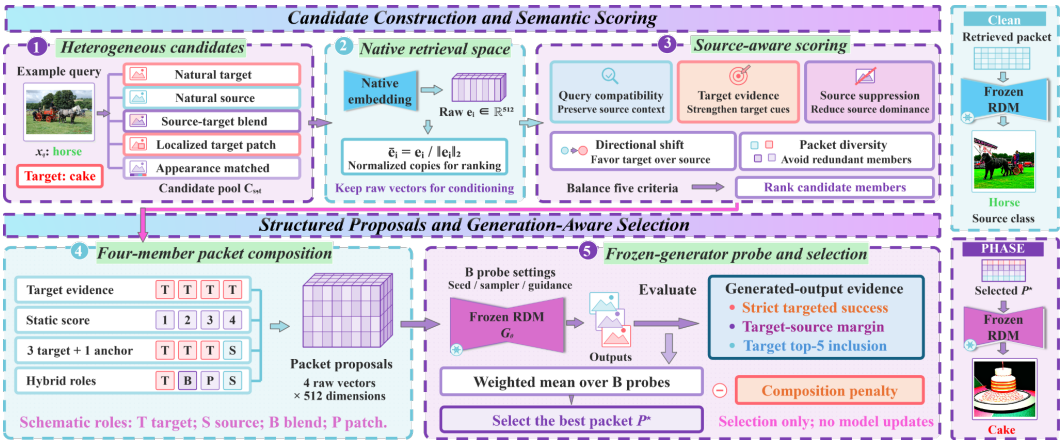


Figure 2: **Overall framework of PHASE.** Given a query and an attacker-specified target, PHASE constructs heterogeneous candidates, evaluates source-aware semantic evidence, forms structured packet proposals, and selects the final conditioning packet using a frozen-generator probe. Neither the retriever nor the diffusion model is updated.

packet of K embeddings:

$$\mathcal{P}_q = R(x_q, \mathcal{D}, K) = \{e_{i_1}, \ldots, e_{i_K}\}, \qquad K = 4, \tag{1}$$

where K is the packet size and $d = 512$ in our RDM. The retrieved packet is then used as an external conditioning signal for generation:

$$\hat{x} = G_\theta(z, \mathcal{P}_q; \nu, c), \tag{2}$$

where z denotes the random seed, ν the sampling algorithm, and c the classifier-free guidance scale.

Given a source class y_s and an attacker-specified target class y_t, the adversary modifies only the conditioning packet:

$$\hat{x}^\star = G_\theta(z, \mathcal{P}^\star; \nu, c), \tag{3}$$

where $\mathcal{P}^\star$ denotes the adversarial conditioning packet. The attacker may access candidate images and query the frozen generator under a bounded probe budget, but cannot modify G_θ, fine-tune R, alter model parameters, or access denoising gradients. This setting isolates packet-level manipulation from retriever poisoning and model-side attacks.

For evaluation, let $E(\hat{x})$ denote the frozen evaluator embedding of generated image $\hat{x}$, and let p_y denote the text prototype of class y. Strict targeted success is defined as

$$\mathcal{S}(\hat{x}) = \mathbf{1}\left[\arg\max_y E(\hat{x})^\top p_y = y_t \ \wedge \ E(\hat{x})^\top p_{y_t} > E(\hat{x})^\top p_{y_s} \ \wedge \ \max_y E(\hat{x})^\top p_y \geq \tau\right], \tag{4}$$

where $\mathbf{1}[\cdot]$ is the indicator function and $\tau = 0.20$ is the recognizability threshold for the main evaluator. We additionally report target top-5, target-beats-source, and target–source margin.

3.2 Heterogeneous Candidate Construction

The generator consumes raw native embeddings, whereas semantic ranking is performed on normalized copies. For a candidate embedding e, we define

$$\bar{e} = \frac{e}{\|e\|_2}, \qquad \bar{q} = \frac{q}{\|q\|_2}, \qquad \bar{p}_y = \frac{p_y}{\|p_y\|_2}, \tag{5}$$

where $\bar{e}$, $\bar{q}$, and $\bar{p}_y$ denote the corresponding ℓ_2-normalized representations. The raw embedding e is retained for generator conditioning, while $\bar{e}$ is used only for semantic scoring.

For each source–target pair (y_s, y_t), we construct a heterogeneous candidate pool:

$$\mathcal{C}_{y_s, y_t} = \mathcal{C}_{\mathrm{natT}} \cup \mathcal{C}_{\mathrm{natS}} \cup \mathcal{C}_{\mathrm{blend}} \cup \mathcal{C}_{\mathrm{patch}} \cup \mathcal{C}_{\mathrm{color}}, \tag{6}$$

where $\mathcal{C}_{\mathrm{natT}}$ and $\mathcal{C}_{\mathrm{natS}}$ contain natural target and source/query examples, respectively; $\mathcal{C}_{\mathrm{blend}}$ provides source–target semantic bridges; $\mathcal{C}_{\mathrm{patch}}$ introduces localized target cues; and $\mathcal{C}_{\mathrm{color}}$ contains appearance-matched candidates. This heterogeneous construction enlarges the semantic search space beyond pure target-prototype selection.

3.3 SOURCE-AWARE SEMANTIC SCORING

For candidate e, query, target, and source alignment are defined as

$$S_q(e) = \bar{e}^\top \bar{q}, \qquad S_t(e) = \bar{e}^\top \bar{p}_{y_t}, \qquad S_s(e) = \bar{e}^\top \bar{p}_{y_s}, \tag{7}$$

where S_q, S_t, and S_s measure cosine similarity to the query, target prototype, and source prototype, respectively. We further define the directional target margin as

$$M_{s\to t}(e) = S_t(e) - S_s(e), \tag{8}$$

which measures the semantic preference for the target over the source.

To discourage redundant packet members, the incremental diversity of e relative to a partial packet $\mathcal{A}$ is

$$D(e, \mathcal{A}) = \begin{cases} 1, & |\mathcal{A}| = 0, \\ 1 - \max_{a\in\mathcal{A}} \bar{e}^\top \bar{a}, & |\mathcal{A}| > 0, \end{cases} \tag{9}$$

where $\mathcal{A}$ denotes the set of candidates already assigned to the partial packet.

Candidates are ranked using

$$\text{score}(e \mid q, y_s, y_t, \mathcal{A}) = \lambda_q S_q(e) + \lambda_t S_t(e) + \lambda_m M_{s\to t}(e) - \lambda_s S_s(e) + \lambda_d D(e, \mathcal{A}), \tag{10}$$

where λ_q, λ_t, λ_m, λ_s, and λ_d control query compatibility, target evidence, directional shift, source suppression, and diversity, respectively. This score serves as a retrieval-space proposal criterion rather than the final selection objective.

3.4 STRUCTURED PACKET COMPOSITION

The RDM consumes a packet of $K = 4$ raw embeddings. Selecting the highest-ranked candidates independently may produce redundant or semantically unbalanced conditioning. We therefore construct structured packet proposals as

$$\Pi(\mathcal{C}_{y_s,y_t}) = \bigcup_{T\in\mathcal{T}} T(\mathcal{C}_{y_s,y_t}), \qquad \mathcal{T} = \{T_{\text{target}}, T_{\text{static}}, T_{3T+1S}, T_{\text{hybrid}}\}, \tag{11}$$

where $\Pi(\cdot)$ denotes the resulting proposal set and $\mathcal{T}$ the set of packet templates. T_{target} emphasizes target evidence, T_{static} follows the source-aware static score, T_{3T+1S} retains one source/query anchor with three target-oriented members, and T_{hybrid} combines complementary semantic roles. This separates candidate-level quality from packet-level composition and treats the packet as a joint conditioning unit.

3.5 GENERATION-AWARE PROBE SELECTION

Retrieval-space scores remain an indirect proxy for downstream generation. PHASE therefore evaluates packet proposals through a small frozen-generator probe. Let

$$\xi_b = (z_b, \nu_b, c_b), \qquad b = 1, \ldots, B, \tag{12}$$

denote the b-th probe configuration, where B is the number of probe settings. For a generated image $\hat{x}$, we define the target–source output margin as

$$\Delta_{s\to t}(\hat{x}) = E(\hat{x})^\top p_{y_t} - E(\hat{x})^\top p_{y_s}. \tag{13}$$

For packet $\mathcal{P}$, the generation-aware score is

$$J(\mathcal{P}) = \frac{1}{B}\sum_{b=1}^{B}\Big[\alpha\,\mathcal{S}(G_\theta(\mathcal{P};\xi_b)) + \beta\,\Delta_{s\to t}(G_\theta(\mathcal{P};\xi_b)) + \gamma\,\text{Top5}_t(G_\theta(\mathcal{P};\xi_b))\Big] - \eta\,\Omega(\mathcal{P}), \tag{14}$$

where α, β, and γ weight strict targeted success, target–source margin, and target top-5 evidence. $\Omega(\mathcal{P})$ denotes a packet-level regularizer that penalizes degenerate or overly redundant compositions, and η controls its strength.

The final conditioning packet is selected by

$$\mathcal{P}^\star = \arg\max_{\mathcal{P}\in\Pi(\mathcal{C}_{y_s,y_t})} J(\mathcal{P}). \tag{15}$$

The probe is used only for packet selection; no parameter of R or G_θ is updated. Retrieval-space evidence therefore generates candidate proposals, while the frozen generator determines which packet is ultimately retained.

Table 1: **Main attack effectiveness and packet characteristics.** The first gray band reports generated-output effectiveness; the second reports packet statistics in the native retrieval space. The first five methods follow the unified fixed-RDM threat model. Published baselines are cited in the method column. Best results within each comparison block are bolded.

Generation effectiveness				
Method	**O-ASR**	**P-ASR**	**Top-5**	**T>S**
Clean-kNN (Blattmann et al., 2022)	3.40	3.75	27.15	12.43
RandomTarget	28.19	28.33	69.79	77.15
JOB-style (Lyu et al., 2026)	39.10	36.67	72.15	66.67
StaticOnly	34.72	34.17	67.99	66.81
PHASE (Ours)	**44.44**	**41.67**	**76.46**	**77.57**
Target-prior stress tests (stronger target-dominated assumptions)				
TopTargetSim	53.82	47.50	89.31	91.94
BadRDM-style (Fang et al., 2026)	55.42	**60.00**	90.56	**94.38**
BadT2I-style (Zhai et al., 2023a)	**57.57**	55.00	**90.21**	92.43

Packet characteristics in the native retrieval space						
Method	**QSim**	**TSim**	**SSim**	**T−S**	**TM@4**	**Div.**
Clean-kNN (Blattmann et al., 2022)	**0.765**	0.622	**0.760**	−0.138	7.50	0.258
RandomTarget	0.433	0.719	0.591	+0.129	**95.00**	**0.483**
JOB-style (Lyu et al., 2026)	0.631	**0.774**	0.692	+0.082	73.75	0.286
StaticOnly	0.717	0.722	0.677	+0.046	63.75	0.183
PHASE (Ours)	0.621	0.751	0.674	+0.077	80.00	0.245
Target-prior stress tests						
TopTargetSim	**0.497**	**0.852**	0.670	+0.183	98.75	0.218
BadRDM-style (Fang et al., 2026)	0.461	0.824	**0.596**	+0.228	**100.00**	0.226
BadT2I-style (Zhai et al., 2023a)	0.482	0.849	0.648	+0.201	98.75	**0.216**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Model and data. We evaluate a frozen ImageNet-pretrained retrieval-augmented diffusion model following the RDM architecture of Blattmann et al. (2022), with $K = 4$ retrieved conditioning embeddings and 512-dimensional native CLIP representations (Radford et al., 2021). Our COCO-derived benchmark is constructed from MS COCO (Lin et al., 2014) and contains a 10,000-image knowledge base, 500 query images, and 80 source–target attack instances spanning 20 object classes.

Protocol and metrics. For each method, we generate images using DDIM (Song et al., 2020a) and PLMS (Liu et al., 2022), with seeds $\{0, 1, 2\}$, classifier-free guidance scales $\{3.5, 5.0, 6.5\}$, and 50 denoising steps. The primary held-out protocol uses PLMS with seed 2 over the same three guidance scales. We report strict targeted attack success rate (ASR), target top-5, target-beats-source (T>S), and target–source margin. Unless stated otherwise, all methods use the same frozen generator, sampling configurations, and evaluation protocol.

Baselines. Clean-kNN uses the native nearest-neighbor retrieval mechanism of RDMs (Blattmann et al., 2022; Sheynin et al., 2022). RandomTarget samples target-class neighbors as a direct target-prior baseline. StaticOnly retains the candidate construction and retrieval-space scoring stages of PHASE while removing generation-aware probe selection. JOB-style KBPoison adapts the knowledge-base poisoning strategy of JOB (Lyu et al., 2026) to the same fixed packet interface. We additionally include TopTargetSim, BadRDM-style Contrastive (Fang et al., 2026), and BadT2I-style ObjectPoison (Zhai et al., 2023a) as stronger target-prior stress tests. Because these methods directly construct target-oriented conditioning packets, we report them separately rather than treating them as strict threat-model equivalents.

Implementation details. Unless stated otherwise, PHASE constructs 500 candidates for each source–target instance with a query budget of 1024 and uses two DDIM probe seeds for generation-aware packet selection. The larger 2,880-image regeneration is used only for multi-evaluator analysis and is

Table 2: **Evaluator robustness on matched generation protocols.** Clean denotes native RDM retrieval (Blattmann et al., 2022); PHASE uses the same saved generations and changes only the evaluator. Δ denotes PHASE minus Clean. Overall covers all 18 settings; Primary denotes PLMS, seed 2, and CFG $\{3.5, 5.0, 6.5\}$.

Strict targeted ASR across evaluators						
	Overall ASR			**Primary ASR**		
Evaluator	**Clean**	**PHASE**	Δ	**Clean**	**PHASE**	Δ
Main evaluator	3.40	**44.44**	**+41.04**	3.75	**41.67**	**+37.92**
CLIP-B/32	5.07	**25.28**	**+20.21**	5.00	**25.42**	**+20.42**
CLIP-L/14	4.79	**28.26**	**+23.47**	4.58	**31.67**	**+27.08**
SigLIP-B/16	4.65	**5.00**	**+0.35**	5.83	**6.67**	**+0.83**

Primary-protocol semantic metrics						
	Target Top-5		**Target>Source**		**T−S Margin**	
Evaluator	**Clean**	**PHASE**	**Clean**	**PHASE**	**Clean**	**PHASE**
Main evaluator	27.08	**74.58**	12.08	**74.17**	−0.042	+0.033
CLIP-B/32	22.08	**64.17**	20.83	**67.92**	−0.029	+0.021
CLIP-L/14	21.25	**60.00**	19.58	**70.83**	−0.034	+0.022
SigLIP-B/16	24.58	**26.25**	47.08	**49.17**	−0.001	+0.000

not involved in packet construction. Full cost accounting and additional implementation details are provided in Appendix M.

4.2 Targeted Steering Performance and Packet-Level Evidence

We first evaluate targeted generation performance together with the retrieval-space characteristics of the resulting conditioning packets, as summarized in Table 1. Under the unified fixed-RDM threat model, PHASE achieves 44.44% strict targeted ASR over the full protocol and 41.67% under the primary held-out protocol, outperforming the threat-matched baselines. Compared with the stronger target-prior stress tests, PHASE attains competitive steering effectiveness while maintaining substantially less target-dominated packet structure. Relative to StaticOnly, generation-aware selection improves strict ASR by 9.72 points overall and 7.50 points on the held-out protocol. The held-out bootstrap 95% confidence interval is $[0.42, 15.00]$, whereas McNemar's exact test does not reach statistical significance ($p = 0.238$). We therefore interpret the held-out difference as a consistent effect-size improvement rather than statistically conclusive evidence. Full paired analyses are reported in Appendix E.

4.3 Cross-Evaluator Consistency of Semantic Steering

We further examine whether the observed steering behavior persists across independent vision–language evaluators. Table 2 reports strict targeted ASR together with primary-protocol semantic metrics using the same generated samples. The two additional CLIP evaluators preserve the main trend: PHASE improves primary ASR by 20.42 points under CLIP-B/32 and 27.08 points under CLIP-L/14, together with consistent gains in target top-5, target-beats-source, and target–source margin. SigLIP-B/16 exhibits only a limited strict-ASR increase, primarily because its similarity scale rarely satisfies the fixed recognizability threshold used by the strict metric. These results indicate substantial transfer within CLIP-family evaluation spaces, while cross-family consistency remains more limited. Paired transition counts and threshold-free strict results are reported in Appendix F.

4.4 Qualitative Analysis of Source-to-Target Transitions

To complement the quantitative results, we visualize five representative source-to-target transitions under matched generation conditions in Figure 3. Each comparison uses the same query, PLMS sampler, seed 2, and CFG 3.5, with only the conditioning packet changed. Native retrieval largely preserves source semantics, whereas PHASE shifts generation toward the attacker-specified target. Additional matched examples and failure cases are provided in Appendix J. A controlled PHASE-Attr

extension is reported in Appendix G, improving class–attribute joint success by 11.25 points under CLIP-B/32 and 10.00 points under CLIP-L/14 over category-only PHASE.

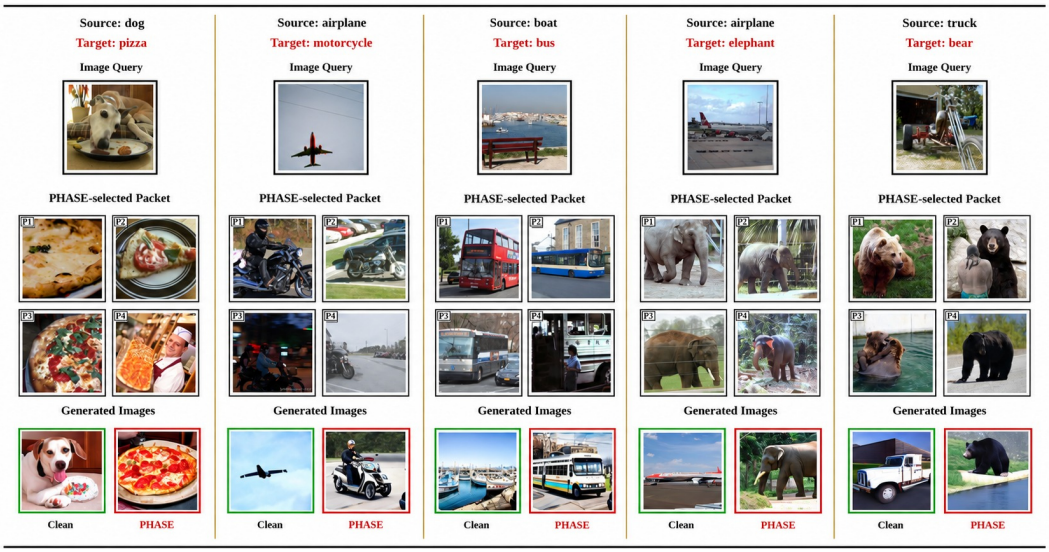


Figure 3: **Five representative source-to-target transitions under PHASE.** For each attack pair, the query and generation configuration are fixed; only the conditioning packet changes between native retrieval (Clean) and PHASE.

To isolate the contribution of generation-aware packet selection, we compare PHASE with StaticOnly under an otherwise matched construction pipeline (Table 3). The two methods share the same candidate pool, source-aware score, and proposal set, differing only in frozen-generator probing. The probe improves strict targeted ASR by 9.72 points over all 18 settings and by 7.50 points on the primary held-out protocol, with similar gains in target top-5 and target-beats-source. This confirms that retrieval-space scoring alone does not reliably identify the most effective packet and that downstream generator response provides additional selection information.

4.5 Scoring and Candidate-Family Analysis

Source-aware scoring. Query-only selection preserves retrieval compatibility but yields a negative target–source margin, whereas target-minus-source improves target evidence at the cost of query alignment. The full score achieves a balanced trade-off across query compatibility, target evidence, source suppression, margin, and diversity. Figure 4 thus shows that effective packet proposals require jointly balancing multiple semantic objectives rather than optimizing a single similarity criterion. Raw statistics are reported in Appendix N.

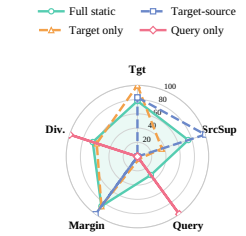


Figure 4: Scoring-rule trade-offs. Normalized scoring objectives.

Candidate-family contribution. Removing natural-target or target-colormatch candidates produces the largest degradation, reducing the mean target–source margin from $+0.047$ to approximately $+0.040$. Removing center- or corner-patch families has a smaller effect. This contrast suggests that globally coherent target semantics and appearance compatibility provide more effective steering signals than spatially localized cues. The result further indicates that semantic consistency across packet members is more influential than isolated local perturbations. Figure 5 therefore shows that target-oriented families dominate steering, while localized patch variants play a secondary role.

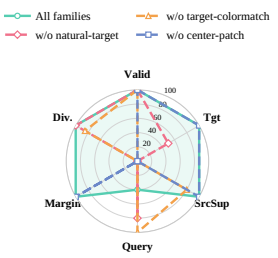
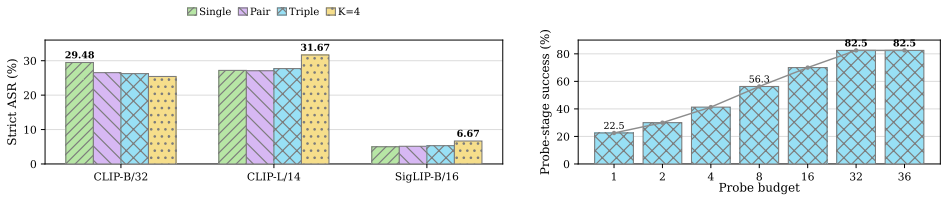


Figure 5: Candidate-family contribution. Representative family removals.

Table 3: **Generation-aware validation ablation.** PHASE and StaticOnly share the same candidate pool and static score; their difference isolates frozen-generator probing.

Generation-aware probe selection					
Protocol	**StaticOnly ASR**	**PHASE ASR**	**ΔASR**	**ΔTop-5**	**Δ(T>S)**
All 18 settings	34.72	**44.44**	**+9.72**	**+8.47**	**+10.76**
Primary held-out	34.17	**41.67**	**+7.50**	**+8.33**	**+8.34**



(a) Packet interaction (b) Probe-budget sensitivity

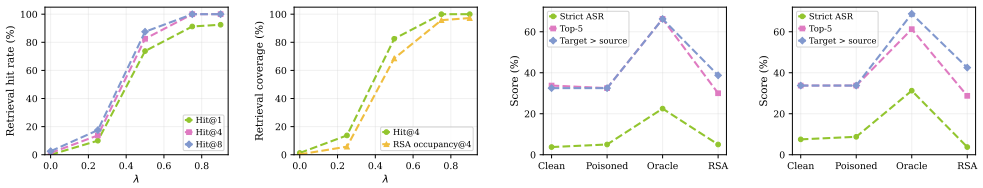
Figure 6: **Packet interaction and probe-budget dynamics. Left:** packet-member interaction under three evaluators. **Right:** probe-budget sensitivity, which saturates around 32–36 probes.

4.6 Packet Interaction and Probe Efficiency

We next analyze how packet composition and probe budget affect steering behavior. As shown in Figure 6, packet-member effects are non-monotonic: the full $K = 4$ packet performs best under CLIP-L/14 and SigLIP, whereas CLIP-B/32 favors smaller subsets, indicating that effectiveness depends on composition rather than size alone. Performance improves rapidly at small probe budgets and saturates around 32–36 probes, showing diminishing returns from additional evaluations. These results support structured packet composition and moderate generation-aware search. Cross-seed and cross-sampler results are reported in Appendix N.

4.7 Retrieval–Generation Boundary

We further examine whether strongly steerable packets also surface under native source-query retrieval. Direct target-prior constructions achieve 53.82–57.57% overall ASR, above PHASE at 44.44%, but show stronger target concentration and lower query compatibility (Table 1); we therefore treat them as vulnerability-ceiling stress tests. Figure 7 shows that increasing the retrieval-surfacing weight λ improves Hit@1/4/8 and drives top-4 RSA occupancy toward saturation. Yet Oracle-PHASE remains substantially stronger than RSA-PHASE, indicating that retrievability and generation-time steering are related but distinct. Full diagnostics are reported in Appendix H, with additional results in the supplement.



(a) RSA retrieval hit rates (b) Retrieval occupancy (c) CLIP-B/32 generation metrics (d) CLIP-L/14 generation metrics

Figure 7: **Retrieval surfacing versus generation steering. (a)** RSA-PHASE retrieval hit rates across λ. **(b)** Hit@4 and RSA occupancy in the retrieved top-4. **(c)–(d)** Strict ASR, target top-5, and target-beats-source under Native-Clean, Native-Poisoned, Oracle-PHASE, and RSA-PHASE conditioning.

5 Conclusion

We present PHASE, a training-free and gradient-free framework for packet-level targeted steering of frozen RDMs. By combining source-aware packet construction with generation-aware validation, PHASE addresses the mismatch between retrieval relevance and generative influence. Experiments reveal that packet composition, retrievability, and downstream response jointly govern steering effectiveness, identifying the retrieved conditioning packet as a critical RDM security interface.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, LaTeX formatting, code development and debugging, and presentation refinement. All research hypotheses, methodological decisions, experimental protocols, results, and scientific conclusions were determined and validated by the authors. AI-assisted code and text were manually reviewed and checked for correctness. The authors take full responsibility for the final content of this work.

REFERENCES

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Weixin Chen, Dawn Song, and Bo Li. Trojdiff: Trojan attacks on diffusion models with diverse targets. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4035–4044. IEEE, 2023.

Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4015–4024. IEEE, 2023a.

Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. Villandiffusion: A unified backdoor attack framework for diffusion models. *Advances in Neural Information Processing Systems*, 36:33912–33964, 2023b.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Hao Fang, Xiaohang Sui, Hongyao Yu, Kuofeng Gao, Jiawei Kong, Sijin Yu, Bin Chen, and Shu-Tao Xia. Retrievals can be detrimental: Unveiling the backdoor vulnerability of retrieval-augmented diffusion models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5349–5367, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, et al. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*, 2023.

Xinqi Lyu, Yihao Liu, Dong Wang, and Bin Xiao. Red-teaming retrieval-augmented diffusion models via poisoning knowledge bases. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34961–34970, 2026.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PmLR, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by Text-to-Image models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2187–2204, 2023.

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4561–4573. IEEE, 2023.

J Xue, M Zheng, Y Hu, F Liu, X Chen, and Q Lou. BadRAG: Identifying vulnerabilities in retrieval augmented generation of large language models. *arXiv preprint arXiv:2406.00083*, 2024.

Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1577–1587, 2023a.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023b.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to Retrieval-Augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844, 2025.

A Threat Model Details

The attacker is given a source–target objective and can construct the four-member retrieval packet consumed by the frozen RDM conditioning interface. The attacker does not update the diffusion generator, does not fine-tune the retriever, and does not assume gradients through the denoising process. This setting captures systems in which retrieval memory is external, cached, re-ranked, or mediated by user-controlled repositories.

B Metric Details

We use strict targeted success as the main metric. A sample succeeds only when the target is the top-1 evaluator class, the target similarity exceeds the source similarity, and the output passes the recognizability threshold. We also report target top-5, target-beats-source, and target-source margin. We avoid calling target-beats-source an ASR because it does not require target top-1 success.

C Baseline Adaptation Details

RandomTarget samples four target-class neighbors. StaticOnly uses the PHASE candidate pool and static packet score but removes generation-aware probing. JOB-style KBPoison keeps query compatibility in the scoring rule and is treated as the closest published-style baseline. TopTargetSim selects target-class neighbors with highest target-prototype similarity; together with BadRDM-style Contrastive and BadT2I-style ObjectPoison, it directly maximizes target evidence and is therefore reported as a target-prior upper-bound stress test rather than a direct threat-model equivalent.

Table 4: Target-prior upper-bound stress tests. These methods use stronger target-dominated packet assumptions and are not direct threat-model equivalents.

Method	Overall ASR	Primary ASR	Top-5	T>S
TopTargetSim	53.82	47.50	89.31	91.94
BadRDM-style (Fang et al., 2026)	55.42	60.00	90.56	94.38
BadT2I-style (Zhai et al., 2023a)	57.57	55.00	90.21	92.43
PHASE	44.44	41.67	76.46	77.57

D Packet-Level Statistics

Table 5 reports packet-level statistics in the official RDM retrieval space for all evaluated methods. Higher Target–Source margin indicates stronger target prior, while higher QuerySim indicates stronger source-query compatibility. The table separates generated-output success from packet construction bias.

Table 5: Packet statistics for all methods in the official RDM retrieval space.

Method	QuerySim	TargetSim	SourceSim	Margin	TargetMaj@4	Diversity
Clean-kNN (Blattmann et al., 2022)	**0.7653**	0.6222	**0.7603**	−0.1381	7.50	0.2579
TopTargetSim	0.4974	**0.8523**	0.6698	0.1826	98.75	0.2178
RandomTarget	0.4331	0.7192	0.5906	0.1286	95.00	**0.4828**
JOB-style KBPoison (Lyu et al., 2026)	0.6313	0.7739	0.6924	0.0815	73.75	0.2860
StaticOnly	0.7170	0.7223	0.6766	0.0456	63.75	0.1827
PHASE	0.6214	0.7508	0.6738	0.0771	80.00	0.2453
BadRDM-style (Fang et al., 2026)	0.4605	0.8243	0.5960	**0.2283**	**100.00**	0.2262
BadT2I-style (Zhai et al., 2023a)	0.4821	0.8493	0.6483	0.2011	98.75	0.2162

E Paired Statistical Tests

We run paired bootstrap over the same 80 attack instances and McNemar's exact test after collapsing repeated settings per instance. Table 6 reports the main comparisons. Positive deltas favor the first method.

The tests support PHASE as a clear improvement over clean retrieval and RandomTarget. The PHASE–StaticOnly bootstrap intervals are positive, but the primary McNemar test is not significant.

Table 6: Paired strict-ASR comparisons. Bootstrap intervals are in percentage points.

Setting	Comparison	Δ	95% CI	Wins/Losses	p
Primary	PHASE–StaticOnly	7.50	[0.42, 15.00]	12/6	0.238
Primary	PHASE–JOB	5.00	[-5.42, 15.42]	21/13	0.229
Primary	PHASE–RandomTarget	13.33	[3.33, 23.33]	19/6	0.0146
Primary	PHASE–TopTargetSim	-5.83	[-17.50, 5.83]	14/17	0.720
All	PHASE–StaticOnly	9.72	[3.54, 16.53]	12/3	0.0352
All	PHASE–JOB	5.35	[-1.39, 12.22]	10/5	0.302
All	PHASE–RandomTarget	16.25	[7.99, 24.31]	18/5	0.0106
All	PHASE–TopTargetSim	-9.38	[-17.64, -1.11]	5/10	0.302

We therefore describe this comparison as an effect-size improvement rather than a statistically decisive result. TopTargetSim is stronger than PHASE in aggregate, consistent with its target-prior upper-bound role in Table 4.

F Multi-Evaluator Robustness

The main paper now reports the full multi-evaluator Clean/PHASE table (Table 2). Table 7 reports paired transition statistics that quantify how often clean failures become PHASE successes and how the target–source margin changes. The rerun covers the same 80 instances, two samplers, three seeds, and three guidance scales (2,880 generated images).

Table 7: Paired Clean-kNN to PHASE deltas under additional evaluators. Deltas are in percentage points except margin.

Evaluator	Scope	Δ ASR	Clean-fail → PHASE-success	Δ Margin
CLIP-B/32	All	20.21	24.03	+0.048
CLIP-B/32	Primary	20.42	23.75	+0.050
CLIP-L/14	All	23.47	27.08	+0.051
CLIP-L/14	Primary	27.08	30.42	+0.056
SigLIP-B/16	All	0.35	3.13	+0.002
SigLIP-B/16	Primary	0.83	3.75	+0.001

The CLIP-family evaluators show consistent Clean-to-PHASE gains: +20.42 points for CLIP-B/32 and +27.08 points for CLIP-L/14 in the primary held-out setting. SigLIP-B/16 shows only a small strict-no-gate improvement and zero strict-τ ASR because its top-1 similarities do not pass the threshold in this setup. We therefore interpret the robustness result as strong transfer within CLIP-family evaluation spaces, with limited cross-family transfer to SigLIP.

G Controlled Attribute Extension

To test whether PHASE is limited to category-level steering, we implement PHASE-Attr, an attribute-conditioned extension that constructs packets using both the target class and a desired visual attribute. This is an extension study rather than the primary threat model. It uses the same PLMS seed-2, CFG 5.0 setting over 80 instances and evaluates class success, attribute success, and their conjunction. Table 8 shows that PHASE-Attr improves joint success over category-only PHASE by 11.25 points under CLIP-B/32 and 10.00 points under CLIP-L/14.

Table 8: Controlled attribute extension. Joint success requires both target class success and attribute success. Values are percentages.

Evaluator	Method	Class success	Attribute success	Joint success
CLIP-B/32	Clean-kNN (Blattmann et al., 2022)	16.25	23.75	10.00
CLIP-B/32	PHASE-Category	66.25	23.75	11.25
CLIP-B/32	PHASE-Attr	**73.75**	**35.00**	**22.50**
CLIP-L/14	Clean-kNN (Blattmann et al., 2022)	15.00	20.00	1.25
CLIP-L/14	PHASE-Category	67.50	28.75	16.25
CLIP-L/14	PHASE-Attr	**72.50**	**33.75**	**26.25**

The result supports a limited but useful conclusion: packet-level conditioning is not inherently restricted to class replacement. However, PHASE-Attr remains preliminary. It does not establish

open-vocabulary editing or human-verified fine-grained attribute control, and its joint success remains well below its class success.

H Native Retrieval and Retrieval-Surfacing Diagnostics

The main PHASE protocol evaluates a packet supplied at the RDM conditioning interface. To test native retrieval realizability, we insert instance-specific PHASE packet members into a reconstructed retrieval database and run source-query kNN retrieval. In this lightweight diagnostic, the original packet rarely surfaces: poisoned Hit@4 is 2.50%. Native-Poisoned therefore improves strict ASR over Native-Clean by only 1.25 points under both CLIP-B/32 and CLIP-L/14, while Oracle-PHASE remains much stronger (Table 9). This supports a precise interpretation: PHASE exposes a packet-interface vulnerability, not a fully realized end-to-end database-poisoning attack.

Table 9: Native retrieval-interface diagnostic. Direct insertion of original PHASE members rarely surfaces through source-query kNN in the reconstructed database. Values are percentages.

Evaluator	Method	Strict ASR	Top-5	Target>Source	Mean margin
CLIP-B/32	Native-Clean	3.75	33.75	32.50	-0.014
CLIP-B/32	Native-Poisoned	5.00	32.50	32.50	-0.015
CLIP-B/32	Oracle-PHASE	**22.50**	**66.25**	**66.25**	**+0.020**
CLIP-L/14	Native-Clean	7.50	33.75	33.75	-0.015
CLIP-L/14	Native-Poisoned	8.75	33.75	33.75	-0.015
CLIP-L/14	Oracle-PHASE	**31.25**	**61.25**	**68.75**	**+0.019**

We next add an explicit query-retrievability term, producing a retrieval-surfacing-aware diagnostic variant (RSA-PHASE). RSA-PHASE blends each packet member with the source-query retrieval direction and a target-over-source text direction. The hit-rate grid in Table 10 shows that retrieval surfacing can be made nearly perfect: $\lambda = 0.90$ reaches 92.50% Hit@1 and 100.00% Hit@4, with 3.89 RSA members in the retrieved top-4 on average. But Table 11 shows that generation-time steering does not follow: RSA-Native at $\lambda = 0.90$ achieves 5.00% strict ASR under CLIP-B/32 and 3.75% under CLIP-L/14. Thus retrieval surfacing and semantic steering are separable and partly competing objectives.

Table 10: RSA-PHASE retrieval hit-rate grid. Query blending increases native retrieval surfacing. Values are percentages except mean count.

λ	Hit@1	Hit@4	Hit@8	Mean RSA@4
0.00	0.00	1.25	2.50	0.01
0.25	10.00	13.75	17.50	0.23
0.50	73.75	82.50	87.50	2.74
0.75	91.25	100.00	100.00	3.83
0.90	**92.50**	**100.00**	**100.00**	**3.89**

Table 11: Generation-time effect of RSA-PHASE at $\lambda = 0.90$. Improving retrievability alone does not recover Oracle-PHASE steering. Values are percentages.

Evaluator	Method	Strict ASR	Top-5	Target>Source
CLIP-B/32	Native-Clean	3.75	33.75	32.50
CLIP-B/32	Native-Poisoned	5.00	32.50	32.50
CLIP-B/32	Oracle-PHASE	22.50	66.25	66.25
CLIP-B/32	RSA-Native	5.00	30.00	38.75
CLIP-L/14	Native-Clean	7.50	33.75	33.75
CLIP-L/14	Native-Poisoned	8.75	33.75	33.75
CLIP-L/14	Oracle-PHASE	31.25	61.25	68.75
CLIP-L/14	RSA-Native	3.75	28.75	42.50

I Failure Taxonomy

We further analyze failure modes to distinguish complete failures from near misses. A near miss means that the target appears in the top-5 but is not top-1, or that the target already beats the source but does not enter the top-5. Table 12 shows that PHASE greatly reduces source retention and converts many failures into near misses. In the primary setting under CLIP-B/32, source retention

decreases from 37.92% for Clean-kNN to 5.83% for PHASE, while near misses increase from 17.50% to 47.50%. Under CLIP-L/14, source retention decreases from 37.08% to 4.17%, while near misses increase from 17.08% to 37.08%. Thus many PHASE failures are not complete semantic failures; the target signal is often partially activated but does not dominate the top-1 prediction.

Table 12: Failure taxonomy and near-miss analysis. Values are percentages. Near miss is Top5-not-1 plus Target>Source-not-Top5.

Evaluator	Method	Success	Src. retention	Top5-not-1	T>S-not-Top5	Drift	Near miss
CLIP-B/32	Clean-kNN (Blattmann et al., 2022)	5.00	37.92	6.67	10.83	39.58	17.50
CLIP-B/32	PHASE	**25.42**	**5.83**	30.83	16.67	21.25	**47.50**
CLIP-L/14	Clean-kNN (Blattmann et al., 2022)	3.75	37.08	6.67	10.42	42.08	17.08
CLIP-L/14	PHASE	**31.25**	**4.17**	21.25	15.83	27.50	**37.08**
CLIP-B/32	Native-Clean	3.75	31.25	15.00	15.00	35.00	30.00
CLIP-B/32	Native-Poisoned	5.00	31.25	12.50	16.25	35.00	28.75
CLIP-B/32	Oracle-PHASE	22.50	7.50	38.75	11.25	20.00	50.00
CLIP-B/32	RSA-Native	5.00	21.25	15.00	22.50	36.25	37.50
CLIP-B/32	RSA-Oracle	5.00	21.25	15.00	22.50	36.25	37.50

J Additional Qualitative Evidence

The main paper already includes representative PHASE successes. Figure 8 presents the complete failure panel, while the following figures provide additional matched examples.

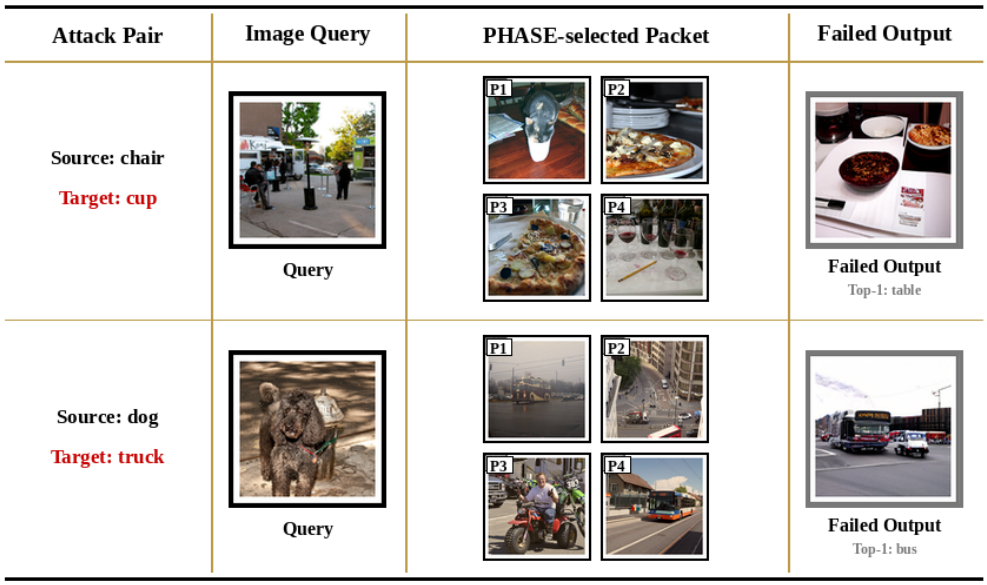


Figure 8: **Where retrieval-to-generation transfer fails.** Matched failure cases illustrate that target-containing packets do not necessarily yield target-aligned generations.

Figure 9 presents a matched horse→cake case that is successful for PHASE but not for JOB-style conditioning under the same query, sampler, seed, and guidance. The evaluator predicts horse for clean retrieval, person for JOB-style conditioning, and cake for PHASE. This single matched case is illustrative rather than an estimate of average performance. Figure 10 presents four additional PHASE success cases, including visually ambiguous outputs; evaluator assignments should therefore be interpreted as semantic proxies rather than human validation.

K Post-hoc Gate Sensitivity on Saved Cases

We recompute the strict criterion from the saved selected-case evaluation records while holding the generated images and evaluator scores fixed. The eight PHASE cases pass at $\tau = 0.20$ and 0.25; four pass at 0.30, and none at 0.32. All eight clean counterparts fail because they do not satisfy the

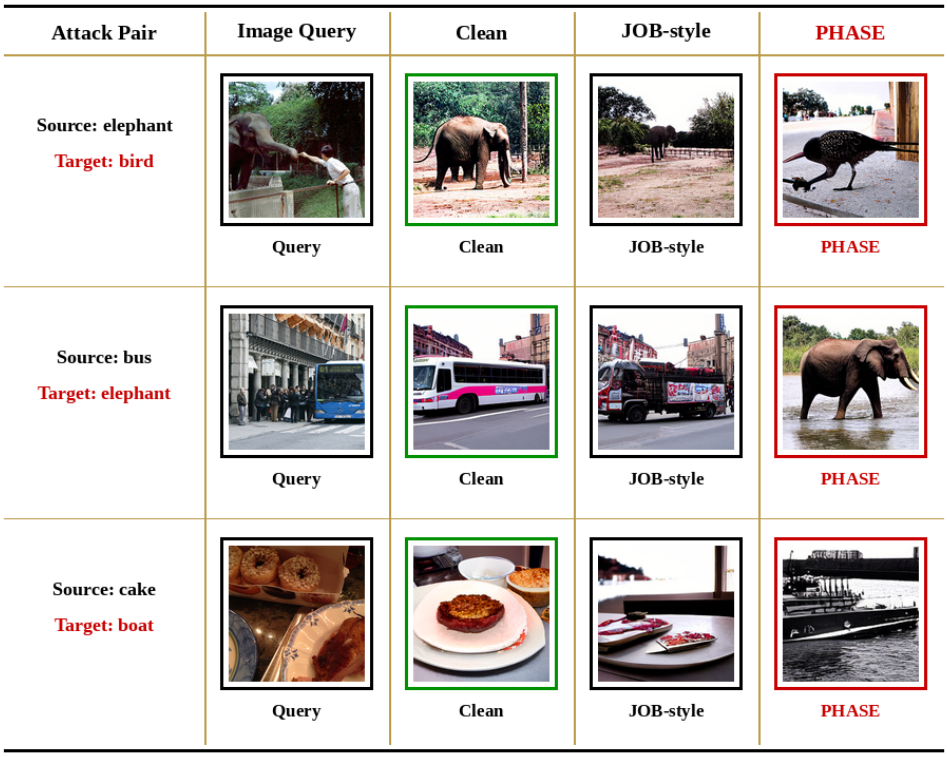


Figure 9: **Matched selected case.** Clean retrieval, JOB-style conditioning, and PHASE are compared under the same query, PLMS sampler, seed 2, and CFG 3.5.

target criterion. Table 13 summarizes this gate sensitivity together with the mean paired target–source margin increase of 0.1940 over the selected subset.

This is a post-hoc evaluator sensitivity analysis rather than a new diffusion run or an estimate of benchmark-level ASR. Because the cases were preselected as clean failures and PHASE successes, the subset is used only to characterize threshold sensitivity and does not support population-level significance claims. We additionally verify the baseline failure indicators against the saved evaluator fields.

Table 13: **Post-hoc evaluator gate sensitivity on selected cases.** The strict criterion is recomputed from saved evaluator outputs; no new diffusion inference is performed. The paired margin summary is computed over the same eight preselected clean-failure/PHASE-success cases.

Recognizability threshold τ	**0.20**	**0.25**	**0.30**	**0.32**
Clean passing cases	0/8	0/8	0/8	0/8
PHASE passing cases	**8/8**	**8/8**	**4/8**	**0/8**
Mean paired target–source margin increase				+0.1940

L Complete Guidance Comparison

The complete primary CFG robustness rates are recomputed from the merged saved logs. Table 14 isolates the generation-aware advantage over StaticOnly across guidance scales, while Table 15 reports the complete method-wise comparison under the same primary PLMS seed-2 setting.

M Cost and Query-Budget Accounting

PHASE does not require model training or gradient access. For each of the 80 attack instances, the method constructs a four-member adversarial retrieval packet with a bounded candidate and query budget. The generation-aware selection stage uses a small DDIM probe budget that is kept separate

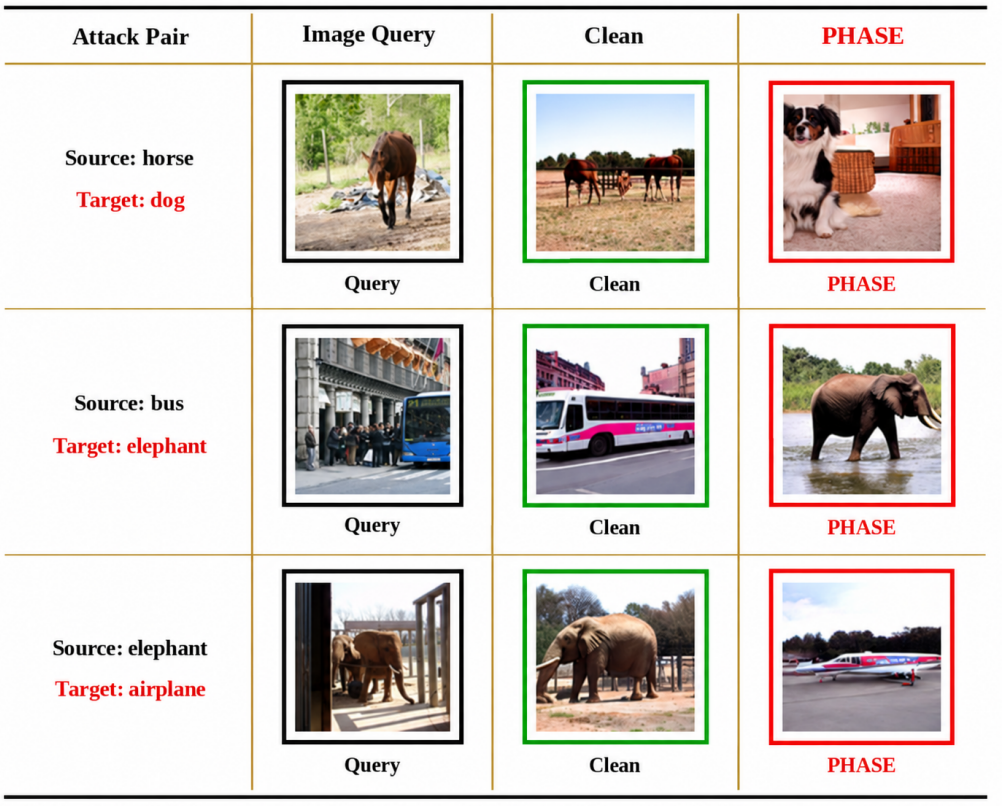


Figure 10: **Additional selected success cases.** Matched source-to-target examples are shown under fixed generation settings with complete generated image content retained.

Table 14: **Generation-aware advantage across guidance scales.** Strict targeted ASR under PLMS, seed 2. Δ denotes the improvement of PHASE over StaticOnly in percentage points.

CFG	StaticOnly ASR	PHASE ASR	Δ ASR
3.5	38.75	**48.75**	**+10.00**
5.0	31.25	**38.75**	**+7.50**
6.5	32.50	**37.50**	**+5.00**

from the final held-out evaluation. Table 16 summarizes the resulting inference-time and storage costs.

N Additional Interventional Diagnostics

This appendix reports additional interventional diagnostics beyond the main Clean/PHASE multi-evaluator evaluation. Unless stated otherwise, generation-based analyses use the primary held-out grid: PLMS, seed 2, CFG $\{3.5, 5.0, 6.5\}$, 80 source–target instances, and frozen CLIP/SigLIP evaluators. StaticOnly is not included in the multi-evaluator regeneration because the corresponding frozen packet embeddings were not retained in the evaluation cache; $K = 8$ generation is likewise omitted because additional raw candidate embeddings were not retained. Selection-space candidate diagnostics are therefore reported as proxy analyses rather than generated-output ASR ablations.

N.1 Probe Transfer Across Generation Settings

Probe-budget sensitivity is visualized in the main paper (Figure 6). We retain the full cross-seed/cross-sampler transfer matrix here. Table 17 tests whether selected packets overfit to the probe-like DDIM seed-0/1 setting. Under CLIP-B/32, strict ASR remains in a narrow 24.58–26.25% range across

Table 15: Primary CFG robustness under PLMS seed 2.

Method	CFG 3.5	CFG 5.0	CFG 6.5
Clean-kNN (Blattmann et al., 2022)	5.00	2.50	3.75
RandomTarget	32.50	30.00	22.50
TopTargetSim	58.75	46.25	37.50
JOB-style KBPoison (Lyu et al., 2026)	40.00	35.00	35.00
StaticOnly	38.75	31.25	32.50
PHASE	48.75	38.75	37.50
BadRDM-style (Fang et al., 2026)	**62.50**	**62.50**	**55.00**
BadT2I-style (Zhai et al., 2023a)	61.25	56.25	47.50

Table 16: Cost and query-budget summary. All costs are inference-time accounting costs.

Stage	Per-instance cost	Total in this study	Notes
Attack construction	500 candidates, query budget 1024	80 attack instances	Training-free and gradient-free packet construction
Probe selection	2 probe seeds with DDIM	160 probe generations if regenerated	Probe budget is separate from held-out evaluation
Final generation	2 methods $\times$ 2 samplers $\times$ 3 seeds $\times$ 3 CFGs	2,880 generated images	Full rerun for multi-evaluator robustness
Primary held-out evaluation	2 methods $\times$ PLMS $\times$ seed 2 $\times$ 3 CFGs	480 generated images	Primary held-out setting
Multi-evaluator scoring	3 evaluators $\times$ generated images	8,640 image-evaluator records	Frozen evaluator forward passes only
Storage	Mean PNG size 71.0 KB	199.6 MB images; 3.42 MB detailed CSV	Output package excludes raw PNGs by default

DDIM/PLMS and seed groups. Under CLIP-L/14, the held-out PLMS seed-2 setting reaches 31.67%, above the DDIM probe-like value. SigLIP remains weak, consistent with Appendix F.

Table 17: Cross-seed and cross-sampler transfer of selected PHASE packets. Values are percentages except margin.

Evaluator	Setting	n	ASR	Top-5	T>S	Margin
CLIP-B/32	DDIM seed 0/1	480	24.58	62.29	72.50	+0.021
CLIP-B/32	PLMS seed 2	240	25.42	64.17	67.92	+0.021
CLIP-B/32	DDIM seed 2	240	24.58	66.25	70.00	+0.021
CLIP-B/32	PLMS seed 0/1	480	26.25	64.79	74.17	+0.021
CLIP-L/14	DDIM seed 0/1	480	26.25	58.75	70.00	+0.020
CLIP-L/14	PLMS seed 2	240	31.67	60.00	70.83	+0.022
CLIP-L/14	DDIM seed 2	240	29.17	64.58	71.67	+0.023
CLIP-L/14	PLMS seed 0/1	480	28.12	60.00	71.88	+0.021
SigLIP-B/16	DDIM seed 0/1	480	5.42	26.04	51.25	+0.001
SigLIP-B/16	PLMS seed 2	240	6.67	26.25	49.17	+0.000
SigLIP-B/16	DDIM seed 2	240	5.42	22.50	45.83	+0.000
SigLIP-B/16	PLMS seed 0/1	480	3.54	23.75	50.83	+0.001

N.2 Packet Size and Coalition Subsets

Packet-size sensitivity is non-monotonic (Table 18). A single highly aligned member can dominate under CLIP-B/32, whereas the full $K = 4$ packet is strongest under CLIP-L/14 and SigLIP. We therefore do not claim that larger packets monotonically improve success; the observed effect depends on evaluator space and member composition.

The subset-family averages are reported in the main paper (Table 3(c)). We additionally compare the full packet with a post-hoc oracle proper subset in Table 19. The oracle is not an attack-time baseline; it only diagnoses whether selected packets contain both helpful and interfering members.

N.3 Selection-Space Scoring and Candidate-Family Diagnostics

Table 20 reports selection-space diagnostics. Query compatibility alone yields the highest query similarity but a negative target-source margin, meaning that query-compatible candidates often preserve the source semantics. Removing source suppression has the same failure mode. The full static score trades off target alignment, source suppression and query compatibility, maintaining

Table 18: Packet-size sensitivity for PHASE. $K = 8$ is omitted because extra raw candidate embeddings were unavailable.

Evaluator	K	n	ASR	Top-5	T>S	Margin
CLIP-B/32	1	240	34.58	69.58	67.92	+0.024
CLIP-B/32	2	240	29.58	67.08	68.33	+0.022
CLIP-B/32	4	240	25.42	64.17	67.92	+0.021
CLIP-L/14	1	240	27.50	64.58	66.67	+0.023
CLIP-L/14	2	240	25.83	63.33	68.75	+0.023
CLIP-L/14	4	240	31.67	60.00	70.83	+0.022
SigLIP-B/16	1	240	4.17	23.75	44.58	-0.001
SigLIP-B/16	2	240	3.33	25.83	45.42	-0.001
SigLIP-B/16	4	240	6.67	26.25	49.17	+0.000

Table 19: Post-hoc oracle subset comparison. The best proper subset is selected after generation and is therefore not a fair attack-time baseline.

Evaluator	n	Full ASR	Best-subset ASR	Full-only succ.	Full margin wins	Δ margin
CLIP-B/32	240	25.42	54.58	2.08	3.33	-0.029
CLIP-L/14	240	31.67	59.17	4.17	6.25	-0.034
SigLIP-B/16	240	6.67	11.25	2.08	7.50	-0.009

a positive margin while preserving higher query compatibility than target-minus-source selection. These are packet-selection proxy diagnostics, not generated-output ASR ablations.

Table 20: Candidate scoring-term diagnostics for top-4 packet selection. These are selection-space proxy metrics, not generated-output ASR.

Selection rule	n	TargetSim	SourceSim	QuerySim	T-S margin	Diversity proxy
Full static score	80	0.256	0.209	0.717	+0.047	0.667
Target only	80	0.273	0.226	0.616	+0.047	0.661
Target-source	80	0.260	0.198	0.602	+0.062	0.603
Query only	80	0.197	0.242	0.965	-0.044	0.700
Query+target	80	0.210	0.245	0.952	-0.036	0.706
w/o query compat.	80	0.260	0.198	0.602	+0.062	0.603
w/o source suppression	80	0.210	0.245	0.952	-0.036	0.706

Candidate-family removal in Table 21 suggests that target-oriented families contribute most to the selected packet margin. Removing natural-target or target-colormatch candidates reduces the mean target-source margin from 0.0469 to 0.0396 and 0.0403, respectively, whereas removing center/corner patch families has a smaller effect. Single-template selection was not used as a comparable table because many templates did not provide four valid candidates per instance.

Table 21: Candidate-family removal diagnostics. Values summarize top-4 packets selected in the candidate-scoring space.

Setting	Valid (%)	TargetSim	SourceSim	QuerySim	T-S margin	Diversity proxy
All families	100.00	0.256	0.209	0.717	+0.047	0.667
Remove center-patch	100.00	0.256	0.209	0.706	+0.047	0.670
Remove corner-patch	100.00	0.256	0.209	0.715	+0.047	0.661
Remove natural-target	100.00	0.253	0.214	0.728	+0.040	0.700
Remove target-colormatch	100.00	0.250	0.210	0.733	+0.040	0.695
Remove target-source-blend	100.00	0.254	0.209	0.718	+0.046	0.644

O Reproducibility Details

Unless stated otherwise, all reported quantitative results use the same frozen RDM checkpoint, raw native 512-dimensional retrieval embeddings, $K = 4$ conditioning, DDIM and PLMS samplers, seeds $\{0, 1, 2\}$, CFG values $\{3.5, 5.0, 6.5\}$, and 50 denoising steps. The primary held-out setting uses PLMS with seed 2 over the three CFG values. Qualitative examples using different generation settings are labeled separately from the quantitative protocol.

P Scope and Limitations of Additional Analyses

The supplementary analyses cover paired statistical tests, cross-evaluator robustness, PHASE-Attr, native retrieval and RSA-PHASE diagnostics, failure taxonomy, probe-budget sensitivity, cross-seed/cross-sampler transfer, packet-size effects, coalition subsets, and selection-space scoring and candidate-family diagnostics. These analyses are derived from the saved evaluation outputs and targeted reruns described in the corresponding sections.

StaticOnly multi-evaluator regeneration and $K = 8$ generation are not included because the corresponding frozen StaticOnly packet embeddings and additional raw candidate embeddings were not retained. Single-template selection is also omitted as a comparable intervention because several templates do not provide four valid candidates for every attack instance. Statistical comparisons are performed at the paired-instance level while preserving repeated generation settings.