FROM GLOBAL TO LOCAL: RETHINKING CLIP FEATURE AGGREGATION FOR PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

CLIP-based person re-identification (ReID) methods typically compress spatial evidence into a single global `[CLS]` representation, which can mix identity-relevant and corrupted regions under occlusion and cross-camera variation. We propose SAGA-ReID, a feature aggregation framework that delays spatial compression by reconstructing each patch through a shared anchor basis before pooling: image patches act as cross-attention queries while learned anchors serve as keys and values, preserving patch-indexed representations rather than first compressing the image into a small set of anchor-indexed features. This limits direct mixing between clean and corrupted patches during refinement and enables selective aggregation afterward. We evaluate this mechanism under synthetic masking and human-distractor occlusion, where SAGA's advantage over global pooling grows with occlusion severity before declining as identity information is exhausted. On standard and occluded ReID benchmarks, SAGA consistently improves over CLIP-ReID, with gains of up to +9.1 Rank-1 on occluded benchmarks. Its reconstructed feature achieves higher Rank-1 than CLIMB-ReID's dedicated Mamba feature despite starting from a weaker backbone, while integrating SAGA with the CLIMB backbone further improves the final fused system on both Rank-1 and mAP. Finally, sensitivity analysis shows that performance is largely invariant to anchor-basis size, indicating that the gains arise from the reconstruction geometry rather than increased anchor capacity.

1 INTRODUCTION

Person re-identification (ReID) aims to match images of the same individual across disjoint camera views and is a fundamental problem in multi-camera perception systems Zheng et al. (2016); Ye et al. (2021). Recent methods increasingly adopt vision–language models, particularly CLIP Radford et al. (2021), to benefit from large-scale image–text pretraining. CLIP-ReID Li et al. (2023), for example, learns identity-discriminative representations while ultimately performing retrieval with a global `[CLS]` feature. Although effective on clean images, this global representation must compress spatial evidence from the entire image into a single vector. Under occlusion or human distractors, identity-relevant and corrupted patches can therefore be mixed before retrieval, precisely when spatial selectivity is most important.

The aggregation problem is not only what information is used, but when it is compressed. Existing alternatives improve spatial modeling in different ways. Part-based methods Sun et al. (2018); He et al. (2021) impose spatial partitions or learned local groupings, while recent CLIP-based approaches such as CLIP-SCGI Han et al. (2024) and PromptSG Yang et al. (2024) use language to condition visual features. A common cross-attention formulation uses a small set of learned or textual queries to attend over image patches. For an anchor query $\mathbf{a}_k$ and visual patches $\{\mathbf{x}_i\}_{i=1}^{N}$, the resulting feature takes the form

$$\tilde{\mathbf{a}}_k = \sum_{i=1}^{N} \alpha_{ki} \mathbf{x}_i,$$

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.

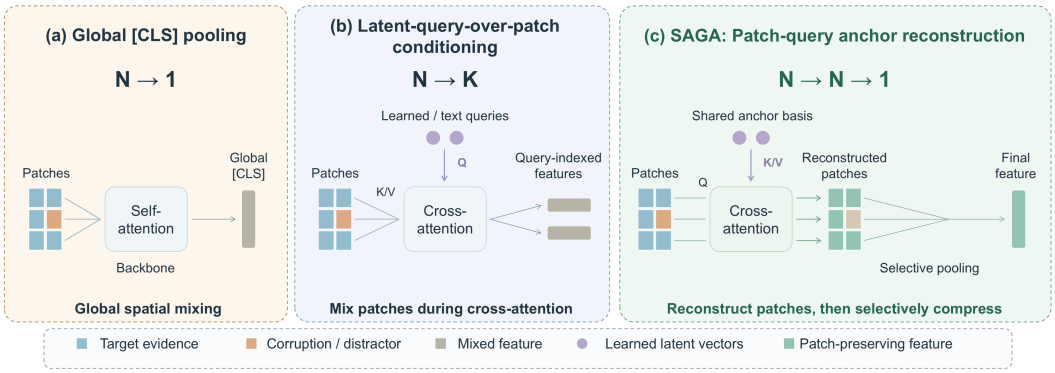


Figure 1: Comparison of feature aggregation geometries. (a) Global [CLS] retrieval mixes spatial evidence into a single representation. (b) Latent-query-over-patch conditioning produces K query-indexed features, mixing information from multiple patches within cross-attention. (c) SAGA reverses the information flow: each of the N patches queries a shared anchor basis, preserving N patch-indexed reconstructed features until selective pooling. Thus, SAGA delays cross-patch mixing until the final aggregation stage.

so multiple spatial regions are mixed into each anchor-indexed output during cross-attention itself. When both target and distractor patches receive high attention, their information is already entangled before subsequent aggregation. CLIMB-ReID Yu et al. (2025) instead applies sequential Mamba-based patch aggregation, but its dedicated Mamba feature remains weaker than its global feature on our single-image occluded evaluation. These observations motivate a different question: can spatial compression be delayed until after each patch has been independently refined?

Patch-wise reconstruction before compression. We propose Structured Anchor Guided Aggregation for ReID (**SAGA-ReID**), which reverses the usual cross-attention roles. Instead of anchors querying image patches, each visual patch queries a shared anchor basis:

$$\tilde{\mathbf{x}}_i = \sum_{k=1}^{K} \beta_{ik} \mathbf{a}_k.$$

This seemingly simple polarity change alters the aggregation geometry. Each patch is refined through the same anchor basis without direct patch-to-patch mixing, so the model retains N patch-indexed representations rather than immediately compressing the image into K anchor-indexed features. Corrupted patches therefore do not directly contribute to the reconstruction of clean patches; spatial mixing is postponed until a subsequent selective pooling step. We parameterize the default anchor basis through the frozen CLIP text space as a structured optimization prior, but the reconstruction mechanism itself does not depend on explicit textual semantics.

This formulation leads to a concrete empirical prediction. If premature spatial compression is a source of fragility, SAGA should provide its largest advantage when part of the image is corrupted but sufficient identity evidence remains, and that advantage should diminish once too much information is lost. We test this prediction under two controlled failure modes: synthetic masking, where identity evidence is removed, and human distractor occlusion, where an overlapping person introduces misleading identity evidence. Across both settings, SAGA exhibits the predicted increase and eventual decline in advantage as occlusion severity grows.

On standard and occluded ReID benchmarks, SAGA consistently improves over CLIP-ReID, with gains of up to +9.1 Rank-1 on occluded benchmarks. Its reconstructed feature also outperforms CLIMB-ReID's dedicated Mamba aggregation feature in Rank-1 despite starting from a weaker backbone, and applying SAGA to the CLIMB-trained backbone yields further gains. Importantly, performance remains nearly unchanged over a wide range of anchor counts, showing that the benefit is not explained by increased anchor capacity or a particular choice of K.

Contributions.

- **A mechanistic view of aggregation under occlusion.** We identify premature spatial compression as a failure mode of global and query-over-patch aggregation: corrupted and identity-relevant regions can be mixed before their reliability can be assessed.
- **Patch-wise reconstruction before pooling.** SAGA reverses the conventional cross-attention roles so that image patches query an external anchor basis, preserving patch-indexed representations until selective aggregation. Controlled polarity and anchor-count analyses separate this reconstruction geometry from anchor capacity.
- **Controlled and benchmark evidence under challenging conditions.** Across synthetic masking, human distractor occlusion, and standard ReID benchmarks, SAGA provides its largest gains precisely where global aggregation becomes unreliable; its reconstructed feature achieves higher Rank-1 than a dedicated Mamba-based patch aggregation feature from a stronger backbone, and SAGA further improves the CLIMB-ReID system when applied to the same backbone.

2 Related Work

Person re-identification. Early ReID methods focused on supervised metric learning with convolutional backbones Hermans et al. (2017). PCB Sun et al. (2018) introduced part-based pooling to preserve local structure, and TransReID He et al. (2021) extended local feature modeling to transformer architectures. Pose-guided and attribute-aware methods Miao et al. (2019); Zhu et al. (2020) further improve robustness under occlusion using explicit keypoint, parsing, or semantic supervision. These approaches demonstrate the value of spatially selective representations, but generally rely on handcrafted partitions, learned visual groupings, or additional supervision. SAGA instead preserves patch-indexed features through anchor-based reconstruction and delays spatial compression until the final aggregation stage.

CLIP-based ReID. CLIP-ReID Li et al. (2023) adapts CLIP for ReID by aligning visual representations with identity-conditioned text prompts while performing retrieval with a global `[CLS]` feature. Subsequent methods such as CLIP-SCGI Han et al. (2024) and PromptSG Yang et al. (2024) incorporate richer language guidance and achieve strong performance on clean benchmarks, but primarily use text as a conditioning signal rather than as an external basis for patch-wise reconstruction. CLIMB-ReID Yu et al. (2025) introduces Mamba-based sequential patch aggregation and temporal modeling. Our analysis complements these works by isolating aggregation behavior under controlled occlusion: the dedicated Mamba feature remains weaker than its global feature on single-image occluded evaluation, while SAGA's reconstructed feature retains stronger robustness. This motivates studying not only what information is used for aggregation, but also when spatial evidence is mixed and compressed.

Cross-attention aggregation and latent bottlenecks. Cross-attention architectures such as Perceiver Jaegle et al. (2021) and slot attention Locatello et al. (2020) use a small set of latent queries to attend over spatial inputs, producing a compact set of query-indexed representations. This design is effective for information compression, but multiple image regions are mixed within each query output during cross-attention. SAGA reverses this information flow: image patches act as queries while anchors serve as keys and values, so the model retains one reconstructed representation per patch before selective pooling. The resulting difference is therefore not simply a swap of attention inputs, but a change in the stage at which spatial evidence is compressed. Our reconstruction-versus-conditioning control and anchor-count sensitivity analysis further show that SAGA's gains are associated with this patch-preserving aggregation geometry rather than increased anchor capacity or a particular choice of K. Our ablations further separate this reconstruction mechanism from the specific text-space parameterization of the anchor basis.

3 Method

SAGA-ReID is built around a simple principle: *preserve patch-indexed representations during refinement and postpone spatial compression until the final aggregation stage*. The method contains three components: (1) a shared structured anchor basis, (2) image-conditioned domain anchors, and (3) a patch-query refinement module followed by selective pooling (Fig. 2).

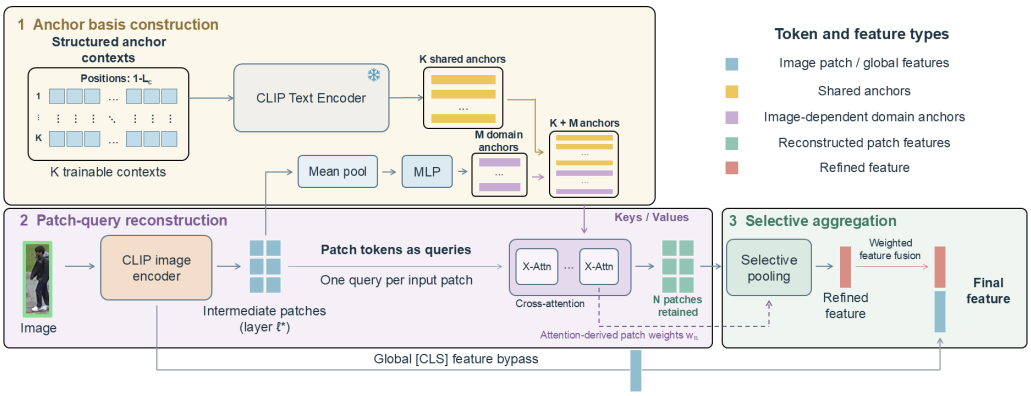


Figure 2: Overview of SAGA-ReID. **(1) Anchor basis construction:** K shared anchors are generated from learnable contexts through the frozen CLIP text encoder and a learnable projection, and augmented with M image-conditioned domain anchors. **(2) Patch-query reconstruction:** intermediate image patches act as queries while the $K + M$ anchors provide keys and values, producing N reconstructed patch features and preserving the patch index through cross-attention. **(3) Selective aggregation:** attention-derived patch weights pool the reconstructed features into a refined representation, which is fused with the global `[CLS]` feature for retrieval.

3.1 Preliminaries

Let $\mathcal{I}$ denote a person image. A CLIP ViT-B/16 image encoder ϕ processes $\mathcal{I}$ through L transformer blocks, producing patch tokens $\mathbf{T}^{(\ell)} \in \mathbb{R}^{N\times D}$ at layer ℓ, where N is the number of patches and $D = 768$. The final block also produces a global `[CLS]` feature $\mathbf{v} \in \mathbb{R}^D$ and a projected CLIP embedding $\mathbf{p} = \phi_{\text{proj}}(\mathbf{v}) \in \mathbb{R}^{512}$. Standard CLIP-ReID retrieval uses the global feature $\mathbf{v}$. SAGA instead extracts intermediate patch tokens $\mathbf{T} = \mathbf{T}^{(\ell^*)}$, $\ell^* < L$, and refines them before spatial aggregation.

3.2 Patch Reconstruction Before Compression

The central distinction in SAGA is the direction of cross-attention. In a conventional latent-query formulation, K queries attend over N visual patches:

$$\tilde{\mathbf{q}}_k = \sum_{n=1}^{N} \alpha_{kn}\, \mathbf{V}_x \mathbf{t}_n, \tag{1}$$

where α_{kn} is normalized over spatial locations. Each output therefore mixes information from multiple image patches during cross-attention, reducing N spatial representations to K query-indexed features.

SAGA reverses this information flow. Each visual patch acts as a query over an external anchor basis $\mathbf{A}^+$:

$$\mathbf{r}_n = \mathbf{t}_n + \sum_{j=1}^{K+M} \beta_{nj}\, \mathbf{V}_a \mathbf{a}_j^+, \tag{2}$$

where β_{nj} is normalized over anchors. The output remains indexed by the original patch n, so N input patches produce N refined patch representations. Thus, cross-attention itself does not directly mix one visual patch into another; each patch is re-expressed through the same shared basis. Image-level coupling is introduced only through the compact domain anchors described below, while the patch index is preserved until final pooling.

This distinction motivates SAGA's behavior under localized corruption. A distractor patch can affect its own reconstructed representation, but it is not directly averaged into every other patch during refinement. The model can therefore assess the reliability of reconstructed patches before compressing them into a single retrieval feature.

3.3 Structured Anchor Basis

We represent the shared reconstruction basis using K anchors $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_K] \in \mathbb{R}^{K \times D}$. The default SAGA parameterization generates each anchor from a learnable context sequence $\mathbf{c}_k \in \mathbb{R}^{L_c \times 512}$ through the frozen CLIP text encoder ψ and learnable projection $\mathbf{W}_{\text{proj}}$:

$$\mathbf{a}_k = \mathbf{W}_{\text{proj}}\, \psi([\mathbf{e}_{\text{sos}}, \mathbf{c}_k, \mathbf{e}_{\text{suf}}]). \tag{3}$$

Here, $\mathbf{e}_{\text{sos}}$ and $\mathbf{e}_{\text{suf}}$ are fixed prompt embeddings. The K context sequences are initialized from the CLIP-ReID prompt initialization and then optimized independently; they are shared across identities and K is unrelated to the number of training identities.

We use CLIP text space as a structured parameterization rather than assuming that individual anchors correspond to explicit semantic body parts. To encourage a diverse basis, we apply

$$\mathcal{L}_{\text{div}} = \frac{\lambda_{\text{div}}}{K(K-1)} \sum_{i \neq j} \left(\hat{\mathbf{a}}_i^\top \hat{\mathbf{a}}_j\right)^2, \tag{4}$$

where $\hat{\mathbf{a}}_k = \mathbf{a}_k / \|\mathbf{a}_k\|_2$. As shown in Sec. 4.5, freely learned anchors recover most of the gain, while the structured parameterization provides a smaller additional benefit and more stable optimization. This supports viewing text space as a regularizing prior rather than the source of the reconstruction mechanism.

3.4 Image-Conditioned Domain Anchors

The shared anchors are constant across images. To provide compact instance-specific context, we additionally generate M domain anchors from the mean patch representation $\bar{\mathbf{t}} = \frac{1}{N}\sum_n \mathbf{t}_n$:

$$\mathbf{D} = \text{reshape}\left(\mathbf{W}_2 \sigma(\mathbf{W}_1 \bar{\mathbf{t}})\right) \in \mathbb{R}^{M \times D}. \tag{5}$$

The full basis is

$$\mathbf{A}^+ = [\mathbf{A}; \mathbf{D}] \in \mathbb{R}^{(K+M) \times D}. \tag{6}$$

These domain anchors introduce a low-dimensional image-conditioned signal while preserving the patch-indexed output structure of the refinement module.

3.5 Anchor-Guided Refinement and Selective Pooling

Given patch tokens $\mathbf{T}$ and the full anchor basis $\mathbf{A}^+$, we apply n stacked cross-attention blocks:

$$\mathbf{T}^{(i)} = \mathbf{T}^{(i-1)} + \text{MHA}\Big(\text{LN}(\mathbf{T}^{(i-1)}), \text{LN}(\mathbf{A}^+), \text{LN}(\mathbf{A}^+)\Big), \tag{7}$$

$$\mathbf{T}^{(i)} = \mathbf{T}^{(i)} + \text{FFN}\Big(\text{LN}(\mathbf{T}^{(i)})\Big). \tag{8}$$

Thus, visual patches provide the queries while anchors provide keys and values. The cross-attention blocks are initialized from CLIP layer ℓ^*.

After the final layer, the attention matrix $\mathbf{W} \in \mathbb{R}^{N \times (K+M)}$ measures patch-to-anchor alignment. We derive a reliability weight for each patch from its strongest anchor alignment:

$$w_n = \max_j \mathbf{W}_{n,j}, \qquad \tilde{w}_n = \frac{w_n}{\sum_{n'} w_{n'}}. \tag{9}$$

Only at this stage are the N refined patches compressed into a single representation:

$$\mathbf{f}_{\text{ref}} = \sum_{n=1}^{N} \tilde{w}_n\, \mathbf{T}^{(n)}_{\text{refined}}. \tag{10}$$

The pooled feature is projected and batch-normalized to form the refined ReID embedding. This ordering—patch-wise reconstruction followed by selective pooling—is the key difference from query-over-patch conditioning, where spatial mixing occurs inside cross-attention itself.

3.6 Training and Inference

Training starts from a CLIP-ReID initialization with the image backbone frozen. We optimize the SAGA contexts, domain-anchor generator, refinement module, and output heads using cached intermediate tokens. The objective combines label-smoothed identity classification, batch-hard triplet loss, CLIP image-to-text alignment, and anchor decorrelation:

$$\mathcal{L} = \mathcal{L}_{\text{id}} + \lambda_{\text{tri}}\mathcal{L}_{\text{tri}} + \lambda_{\text{i2t}}\mathcal{L}_{\text{i2t}} + \mathcal{L}_{\text{div}}. \tag{11}$$

At inference, the normalized refined feature and global `[CLS]` feature are concatenated with fixed weights, $\mathbf{f}_{\text{fuse}} = [\sqrt{w_r}\hat{\mathbf{f}}_{\text{ref}}; \sqrt{w_i}\hat{\mathbf{v}}]$. For CLIP-ReID-backbone experiments, we use $w_r = 10, w_i = 1$, selected once on a held-out ReID validation dataset and fixed across benchmarks. The separately calibrated CLIMB-backbone ratio is reported in Table 2. Full optimization details are provided in the appendix.

4 Experiments

Our experiments test three questions motivated by the aggregation mechanism in Sec. 3.2: (1) does preserving patch-indexed representations improve robustness when only part of the spatial evidence is corrupted; (2) does this behavior hold for both missing evidence and semantically misleading human distractors; and (3) are the gains explained by the reconstruction geometry rather than anchor capacity or a stronger backbone? We first study these questions under controlled corruption and then evaluate on standard ReID benchmarks.

4.1 Robustness under Controlled Masking

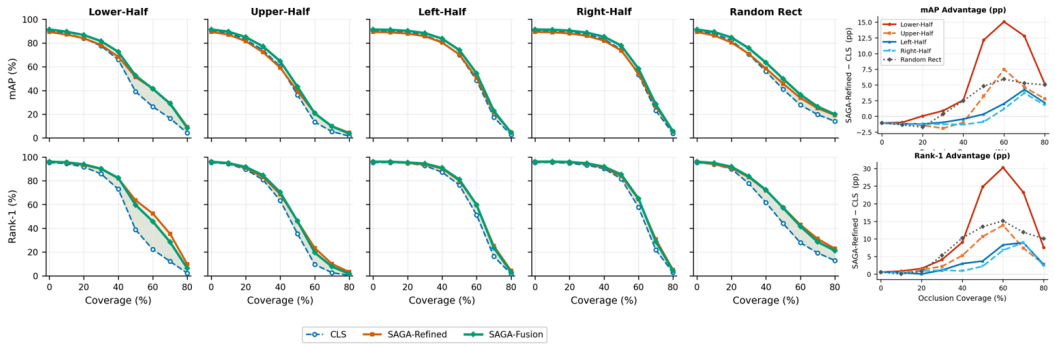


Figure 3: Robustness under controlled masking on Market-1501. **Left:** mAP and Rank-1 under five masking geometries at increasing coverage. **Right:** SAGA-Refined advantage over the global `[CLS]` representation. The advantage grows as global aggregation becomes increasingly corrupted and declines once too little identity evidence remains.

We evaluate Market-1501 with synthetically masked query images under five geometries—lower-half, upper-half, left-half, right-half, and random rectangle—at coverage levels from 0% to 80%. Gallery images remain unmodified. Because the backbone, training data, and gallery are fixed, this experiment isolates how the deployed representations respond to localized loss of spatial evidence.

Figure 3 shows a consistent peak-then-decline pattern across all five masking geometries. SAGA's advantage increases as masking begins to corrupt the global representation, reaches its maximum at intermediate-to-high coverage, and decreases once insufficient identity evidence remains for either representation. The largest gain occurs under lower-half masking, reaching **+30 Rank-1 pp** and **+15 mAP pp** at 60% coverage. The differing gain magnitudes across geometries are expected because each mask preserves different identity cues; the common trend is that delaying spatial compression is most beneficial when useful and corrupted evidence coexist.

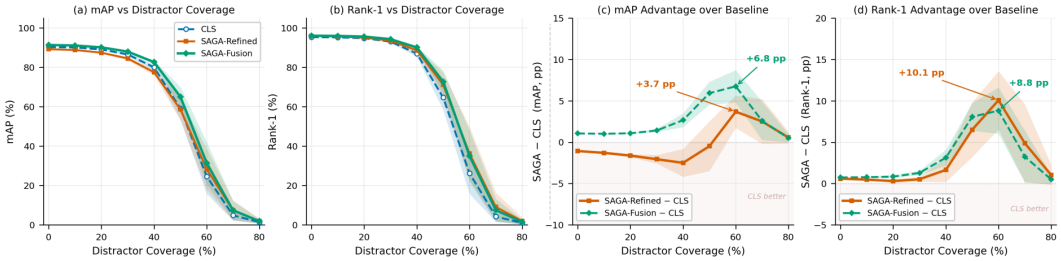


Figure 4: Robustness under human distractor occlusion on Market-1501. A segmented person from the disjoint MSMT17 dataset is composited into query images at increasing coverage while galleries remain clean. SAGA-Fusion reaches a mean advantage of **+10.1 Rank-1 pp** and **+6.8 mAP pp** near 60% coverage. Shaded regions denote ± 1 standard deviation across distractor identities and insertion angles.

4.2 Robustness under Human Distractor Occlusion

Masking removes identity evidence, whereas real occlusion can introduce *misleading* evidence from another person. We therefore composite segmented MSMT17 identities onto Market-1501 queries at controlled coverage levels. This setting is particularly relevant to the proposed mechanism because target and distractor regions are both person-like and can receive strong responses from a global or query-over-patch aggregator.

The reconstructed branch is slightly weaker than the global feature at very low distractor coverage, where the target remains largely intact, but its relative advantage increases as distractor coverage grows. The fused representation remains beneficial throughout the tested range and peaks near 60% coverage. Together with the masking experiment, this shows that the benefit is not limited to absent image regions: the same qualitative behavior appears when corruption contains semantically plausible but identity-incorrect evidence.

4.3 Comparison with State-of-the-Art

Table 1: Comparison with state-of-the-art methods on standard and occluded ReID benchmarks. Methods are grouped by input modality and supervision setting. †Results reproduced using authors' released code and configurations. ‡No public implementation available; results are from the original paper and could not be reproduced under a consistent experimental protocol.

Method	Backbone	Market-1501		MSMT17		Occluded-ReID		Occluded-Market	
		R1	mAP	R1	mAP	R1	mAP	R1	mAP
Image-based ReID Methods									
PCB Sun et al. (2018)	ResNet50	93.8	81.6	–	–	41.3	38.9	66.0	49.4
BOT Luo et al. (2019)	ResNet50	94.5	85.9	–	–	–	–	–	–
TransReID He et al. (2021)	ViT-B/16	95.2	88.9	85.3	67.4	–	–	78.2	64.7
CLIP-ReID Li et al. (2023)	ViT-B/16	95.4	90.5	89.7	75.8	90.2	88.2	79.5	68.7
PRO-FD Cui et al. (2024)	ViT-B/16	95.6	90.8	–	–	92.3	90.3	–	–
FCFormer Wang et al. (2024a)	ViT-B/16	–	–	–	–	84.9	86.2	–	–
PADE Wang et al. (2024b)	ViT-B/16	95.8	89.8	–	–	83.7	79.9	–	–
SCING Xie et al. (2025)	ViT-B/16	96.2	91.0	–	–	93.8	90.9	80.3	69.2
PromptSG‡ Yang et al. (2024)	ViT-B/16	97.0	94.6	92.6	87.2	–	–	–	–
Video-based ReID Methods									
CLIMB-ReID† Yu et al. (2025)	ViT-B/16	96.3	92.0	89.6	76.9	87.9	85.9	88.6	78.7
Methods with External Supervision									
CLIP-SCGI‡ Han et al. (2024)	ViT-B/16	**97.6**	**96.0**	**92.9**	**88.2**	–	–	–	–
Ours (Image-based, No External Supervision)									
SAGA-ReID (Ours)	ViT-B/16	96.20 ± 0.08	91.53 ± 0.09	90.40 ± 0.08	77.1 ± 0.12	**94.23 ± 0.25**	**91.90 ± 0.22**	88.60 ± 0.08	76.43 ± 0.25
SAGA+CLIMB (Ours)	ViT-B/16	96.6	92.7	90.0	78.4	89.9	88.4	**90.1**	**81.5**

Table 1 evaluates whether the behavior observed under controlled corruption transfers to standard ReID benchmarks.

Occluded benchmarks. On Occluded-ReID and Occluded-Market, SAGA-ReID achieves **94.2/91.9** and **88.6/76.4** Rank-1/mAP, respectively. Relative to CLIP-ReID, the gain reaches **+9.1 Rank-1** on Occluded-Market, consistent with the controlled experiments showing the largest benefit when global aggregation is unreliable. Applying SAGA to the reproduced CLIMB-ReID back-

bone further improves performance to **90.1/81.5**, showing that SAGA remains complementary to a stronger backbone and a dedicated sequential aggregation architecture.

Standard and cross-camera benchmarks. SAGA-ReID also improves over CLIP-ReID on Market-1501 (**96.2/91.5** Rank-1/mAP) and MSMT17, where the gains are **+0.7/+1.3** Rank-1/mAP. The smaller improvements on clean benchmarks are consistent with the controlled low-corruption regime, where the global representation already retains most useful identity evidence. PromptSG and CLIP-SCGI report stronger clean-benchmark performance, but our focus is the robustness of feature aggregation under spatial corruption rather than state-of-the-art accuracy on every clean benchmark.

4.4 Representation and Backbone Analysis

Table 1 reports mean performance over three runs, while the component and ablation analyses below use one representative run from the same experimental setting.

Table 2: Analysis of feature representations at different stages of the proposed pipeline, and comparison with CLIMB-ReID feature representations. Results show that (1) structured anchor refinement improves robustness under occlusion, (2) combining global and refined features yields the best performance, and (3) applying SAGA to the CLIMB backbone further improves CLIMB-ReID's combined representation in both Rank-1 and mAP. Bold indicates best result within each backbone group.

	CLIP-ReID backbone		*CLIMB-ReID backbone*[†]	
Feature Type	R1	mAP	R1	mAP
Occluded-Market				
Global (`[CLS]`)	79.5	68.7	88.9	80.3
Mamba feature	–	–	84.6	71.7
CLIMB combined (reproduced)	–	–	88.6	78.7
SAGA feature	87.8	70.0	87.6	74.8
SAGA fusion	**89.2**	**75.6**	**90.1**	**81.5**

Fusion weights: $^{\text{CLIP}}(w_r, w_i) = (10, 1)$; $^{\text{CLIMB}}(w_r, w_i) = (1, 2)$.

[†] On this single-image occluded evaluation, the Mamba feature underperforms [CLS], suggesting that sequential patch aggregation alone does not guarantee improved occlusion robustness.

Global versus reconstructed representations. The reconstructed feature provides a substantially more robust representation than the CLIP-ReID `[CLS]` feature on Occluded-Market, while their fusion achieves the strongest result. This complementarity is expected: reconstruction preserves selective local evidence, whereas `[CLS]` retains global information that remains useful when corruption is limited.

Comparison with dedicated patch aggregation. CLIMB-ReID provides an important alternative aggregation baseline through its Mamba-based patch feature. On Occluded-Market, the reproduced Mamba feature achieves 84.6 Rank-1, compared with 87.8 for SAGA's reconstructed feature, despite SAGA starting from the weaker CLIP-ReID backbone. Applying SAGA to the CLIMB-trained backbone yields further improvement. These results indicate that the observed robustness is not explained solely by stronger backbone fine-tuning or by introducing an additional patch-aggregation module.

4.5 Mechanistic Analysis

Reconstruction versus conditioning. Reversing the attention direction changes the representation being produced: patch-query reconstruction preserves N patch-indexed outputs, whereas anchor-query conditioning produces K anchor-indexed outputs and mixes spatial evidence within cross-attention. Because these output structures differ inherently, an identical downstream readout is not naturally defined. We therefore compare SAGA against a strong reversed-polarity conditioning variant using the same anchor initialization and domain anchors together with learned aggregation.

Table 3: Sensitivity to anchor-basis size K on Market-1501.

K	8	16	24	32	48	128
mAP	91.74	91.75	91.67	**91.81**	91.64	91.56
R1	**96.26**	96.23	96.14	96.11	96.08	95.93

The reversed variant decreases standalone performance from 87.8/70.0 to 85.1/66.4 Rank-1/mAP on Occluded-Market. More importantly, its fusion gain is substantially smaller (+0.6 Rank-1, +2.4 mAP) than that of patch-query reconstruction (+1.4 Rank-1, +5.6 mAP). This supports the interpretation that preserving patch-indexed representations before aggregation produces information that is more complementary to the global feature.

Sensitivity to anchor-basis size. To test whether the gain is caused simply by anchor capacity, we vary K over a $16\times$ range while keeping the SAGA direction and all other settings fixed:

Across $K = 8$–128, variation is only 0.25 mAP and 0.33 Rank-1, showing that the gain is neither tied to a narrowly selected K nor increased anchor capacity. Once a sufficient external basis exists, its exact cardinality has little effect, supporting reconstruction geometry rather than basis size as the important factor.

Anchor parameterization. Freely learned anchors already improve Rank-1 from 79.5 to 86.9, showing that explicit text semantics are not required for the core gain. Structured anchors provide a further fused mAP improvement from 74.5 to 75.6 and were more stable during development; we therefore treat text-space parameterization as a secondary optimization prior rather than the reconstruction mechanism itself.

Qualitative attention visualizations in Appendix 6.1 further show that SAGA remains more target-focused under human-distractor occlusion and that individual anchors exhibit distinct spatial response patterns.

Table 4: Ablation study on Occluded-Market. We compare free vs structured anchors and progressively add domain anchors and feature fusion.

Variant	FAR	SAR	PR	DA	FF	R1	mAP
Structured anchor-based model							
Baseline (CLIP-ReID)						79.5	68.7
+ Structured anchors (SAR)		✓				87.4	68.6
+ Domain anchors (DA)		✓		✓		87.8	70.0
+ Feature fusion (FF)		✓		✓	✓	**89.2**	**75.6**
Polarity reversal (PR)							
+ SAR + DA + PR		✓	✓	✓		85.1	66.4
+ SAR + DA + PR + FF		✓	✓	✓	✓	85.7	68.8
Free anchor baseline (control)							
+ Free anchors (FAR)	✓					86.9	68.5
+ Free anchors + DA + FF	✓			✓	✓	88.9	74.5

5 Conclusion

We introduced SAGA-ReID, which preserves patch-indexed representations through anchor-guided reconstruction and delays spatial compression until selective pooling. Controlled masking and human-distractor experiments show increasing advantages as global aggregation is corrupted, with gains of up to **+9.1 Rank-1** under occlusion. On the CLIMB backbone, replacing Mamba aggregation with SAGA improves both Rank-1 and mAP. Polarity and K-sensitivity analyses further attribute the gains to patch-preserving reconstruction rather than anchor count or capacity. These results support delayed spatial compression as an effective strategy for robust ReID.

AI USE STATEMENT

Generative AI tools were used to assist with implementation of the proposed method, refinement of the conceptual framing and hypotheses, feedback on research methodology and experimental design, interpretation of experimental results, and editing of the manuscript for clarity and readability. Generative AI was not used to generate synthetic datasets, formulate or prove theoretical results, translate content, clean or reformat datasets, or perform qualitative or thematic data analysis. AI-assisted code and research suggestions were reviewed and validated by the authors through manual inspection and experimental testing. The authors take full responsibility for the methodology, implementation, experiments, claims, and conclusions presented in this work.

REPRODUCIBILITY STATEMENT

The architecture and training procedure are described in Sec. 3, with optimization settings, computational costs, ablation implementation details, and controlled-occlusion construction provided in the Appendix. Code sufficient to reproduce the reported experiments is included in the anonymous supplementary material.

ETHICS STATEMENT

Person re-identification can support legitimate multi-camera perception applications but also carries privacy risks when used for tracking without consent. This work introduces no new data collection or deployed surveillance system and evaluates only existing research benchmarks. Deployment should consider applicable privacy, consent, and data-governance requirements.

REFERENCES

Can Cui, Siteng Huang, Wenxuan Song, Pengxiang Ding, Min Zhang, and Donglin Wang. Profd: Prompt-guided feature disentangling for occluded person re-identification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1583–1592, 2024.

Qianru Han, Xinwei He, Zhi Liu, Sannyuya Liu, Ying Zhang, and Jinhai Xiang. CLIP-SCGI: Synthesized caption-guided inversion for person re-identification. *arXiv preprint arXiv:2410.09382*, 2024. URL https://arxiv.org/abs/2410.09382.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15013–15022, 2021.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. URL https://arxiv.org/abs/1703.07737.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Siyuan Li, Li Sun, and Qingli Li. Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11525–11538. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8511df98c02ab60aea1b2356c013bc0f-Abstract.html.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

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 International Conference on Machine Learning (ICML)*, 2021.

Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Tao Wang, Mengyuan Liu, Hong Liu, Wenhao Li, Miaoju Ban, Tianyu Guo, and Yidi Li. Feature completion transformer for occluded person re-identification. *IEEE Transactions on Multimedia*, 26:8529–8542, 2024a.

Zi Wang, Huaibo Huang, Aihua Zheng, Chenglong Li, and Ran He. Parallel augmentation and dual enhancement for occluded person re-identification. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3590–3594. IEEE, 2024b.

Yunfei Xie, Yuxuan Cheng, Juncheng Wu, Haoyu Zhang, Yuyin Zhou, and Shoudong Han. Scing: Towards more efficient and robust person re-identification through selective cross-modal prompt tuning. *arXiv preprint arXiv:2507.00506*, 2025.

Zexian Yang, Dayan Wu, Chenming Wu, Zheng Lin, Jingzi Gu, and Weiping Wang. A pedestrian is worth one prompt: Towards language guidance person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17343–17353, 2024.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Chenyang Yu, Xuehu Liu, Jiawen Zhu, Yuhao Wang, Pingping Zhang, and Huchuan Lu. Climbreid: A hybrid clip-mamba framework for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9589–9597, 2025.

Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. URL `https://arxiv.org/abs/1610.02984`.

Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, volume 12348 of *Lecture Notes in Computer Science*, pp. 346–363, Cham, 2020. Springer. doi: 10.1007/978-3-030-58580-8_21.

6 Appendix

6.1 Attention Visualization

Figure 5 qualitatively illustrates the behavior observed in the controlled experiments. Under human distractor occlusion, the global `[CLS]` attention shifts toward the distractor, while the aggregate SAGA weighting remains more concentrated on the target. Individual anchors exhibit different spatial response patterns, consistent with the anchor basis providing multiple reconstruction directions rather than a single fixed pooling strategy.

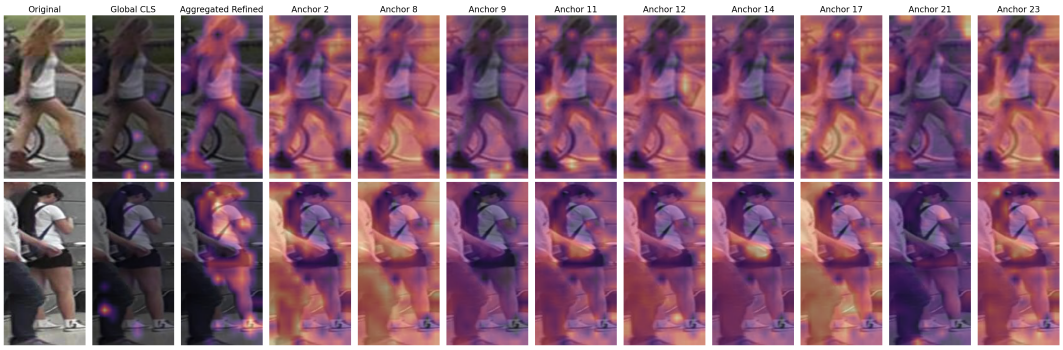


Figure 5: Attention maps on Occluded-Market. Columns show the original image, global [CLS] attention, aggregated SAGA attention, and representative per-anchor responses. Under human distractor occlusion, global attention shifts toward the distractor, whereas SAGA's aggregate weighting remains more concentrated on the target.

6.2 Computational Overhead

SAGA-ReID adds approximately 17M trainable parameters on top of the frozen ViT-B/16 backbone (≈86M) — a Structured Refinement Module (two cross-attention layers, ≈14.2M), a Domain Anchor Generator (two-layer MLP, ≈2.4M), and lightweight projection heads (≈0.8M) — compared to full end-to-end fine-tuning of the backbone in CLIP-ReID. Inference overhead is approximately 13.7% of backbone cost: passing $N{=}210$ patch tokens (from a $256{\times}128$ image with stride 12) through two cross-attention layers over $K{+}M{=}27$ anchors adds ≈5.1 GFLOPs per image against a backbone cost of ≈37.5 GFLOPs. The additional SAGA training stage completes in approximately 38 minutes on Market-1501, following approximately 60 minutes for the CLIP-ReID initialization, as backbone tokens are pre-extracted and cached, so each iteration involves only a forward pass through the lightweight visual module. We trained on an RTX 4080 GPU with 16 GB memory.

6.3 Implementation Details

Stages 1 and 2 in config correspond to the CLIP-ReID training. Stage 3 is our SAGA training. This configuration is used for all of our SAGA-ReID experiments.

6.4 Nature of text-space grounding.

We augmented each anchor's context with a fixed text suffix (e.g., "upper body of a person") to test whether explicit semantic steering improves performance. It did not — performance was equivalent to fully learnable anchors and attention maps did not align with the specified concepts. We attribute this to operating at layer $\ell^* = 9$, where patch tokens have diverged from CLIP's joint embedding space. Text-space initialization therefore contributes through optimization structure and anchor diversity rather than enforced semantic correspondence — consistent with the per-anchor analysis in Section 6.1, which shows behaviorally distinct but not rigidly semantic patterns.

6.5 Fusion Weight Sensitivity

Performance rises sharply from the CLIP-only baseline ($w_r \to 0$), peaks at an intermediate ratio, and gradually declines as the backbone signal is suppressed — confirming complementarity between the two components. The larger optimal ratio on Occluded-ReID ($9.9\times$ vs. $5.9\times$ on Market-1501) reflects the increased importance of structured reconstruction under occlusion. Our fixed $10\times$ configuration achieves near-optimal performance across all datasets without dataset-specific tuning.

Config Param	Value	Brief Description
dataset	occ_market	Dataset used for training/evaluation.
img_size	[256, 128]	Input image size (H, W).
stride_size	[12, 12]	ViT patch stride.
pixel_mean	[0.5, 0.5, 0.5]	Input normalization mean.
pixel_std	[0.5, 0.5, 0.5]	Input normalization std.
stage1_epochs	100	Stage-1 epochs.
stage1_lr	3.5e-4	Stage-1 learning rate.
stage1_warmup	5	Stage-1 warmup epochs.
stage2_epochs	80	Stage-2 epochs.
stage2_lr	5e-6	Stage-2 learning rate.
stage2_milestones	[40, 70]	Stage-2 LR decay milestones.
stage2_warmup_iters	10	Stage-2 warmup iterations.
lambda_id	0.25	ID classification loss weight.
lambda_triplet	1.0	Triplet loss weight.
lambda_i2t	1.0	Image-to-text alignment loss weight.
margin	0.3	Triplet margin.
s3_token_layer	9	ViT layer for token extraction in Stage-3.
s3_anchor_part_ctx	12	Learnable context tokens per structured anchor, initialized from CLIP-ReID prompt.
s3a_epochs	20	Stage-3a epochs.
s3a_lr	3.5e-4	Stage-3a learning rate.
s3a_warmup	5	Stage-3a warmup epochs.
s3b_epochs	80	Stage-3b epochs.
s3b_lr	2e-4	Stage-3b learning rate.
s3b_warmup	5	Stage-3b warmup epochs.
s3_num_anchors	24	Number of shared reconstruction anchors.
s3_num_domain_anchors	3	Number of dynamic domain anchors.
s3_refine_layers	2	Structured refinement module depth.
s3_embed_dim	768	Refined feature dimension.
s3_dropout	0.1	Dropout in Stage-3 visual module.
s3_min_lr_ratio	0.5	Minimum LR ratio for Stage-3 cosine schedule.
fusion_w_refined	10.0	Refined feature weight in fusion evaluation.
fusion_w_imgfeat	1.0	Global image feature weight in fusion evaluation.
batch_size	64	Training batch size.
num_instance	4	Images per identity for sampler.
weight_decay	1e-4	Optimizer weight decay.

Table 5: Implementation-relevant training configurations for V3 on Occluded-Market.

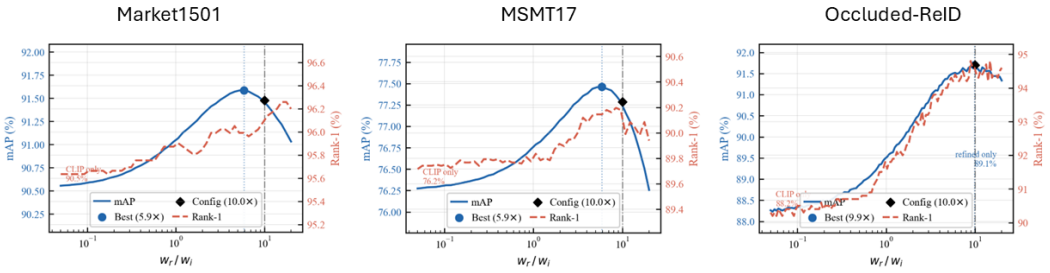


Figure 6: **Effect of fusion weight ratio w_r/w_i on retrieval performance.** mAP (left axis, blue) and Rank-1 (right axis, red dashed) as a function of w_r/w_i with $w_r + w_i = 1$. Circle marks best mAP; diamond marks our fixed $10\times$ configuration.

6.6 Polarity Reversal Ablation Implementation Details

To study whether anchor-driven *suppression* changes aggregation behavior, we introduce a polarity-reversed refinement mode, `anchor_patch_suppress`. In this mode, anchors act as queries and visual patches act as keys/values.

Let $X^{(l)} \in \mathbb{R}^{B\times N\times D}$ be patch tokens at layer l, and $A^{(l)} \in \mathbb{R}^{B\times K\times D}$ be structured anchors. We compute

$$\tilde{A}^{(l+1)},\, W^{a\to t,(l)} = \mathrm{CA}\Big(A^{(l)}, X^{(l)}\Big), \tag{12}$$

where $W^{a\to t,(l)} \in \mathbb{R}^{B\times K\times N}$ are anchor-to-token attention weights. For compatibility with downstream modules, we convert to token-to-anchor form:

$$W^{t\to a,(l)} = \mathrm{Norm}_K\Big((W^{a\to t,(l)})^\top\Big), \quad W^{t\to a,(l)} \in \mathbb{R}^{B\times N\times K}. \tag{13}$$

Instead of reconstructing tokens from anchors, we derive a token saliency gate:

$$s_n^{(l)} = \max_k W_{n,k}^{t\to a,(l)}, \qquad \hat{s}_n^{(l)} = \frac{s_n^{(l)}}{\frac{1}{N}\sum_{m=1}^{N} s_m^{(l)} + \epsilon}, \tag{14}$$

$$g_n^{(l)} = \mathrm{clip}\Big((\hat{s}_n^{(l)})^\gamma,\ g_{\min},\ 1\Big), \qquad X_n^{(l+1)} = g_n^{(l)} X_n^{(l)}. \tag{15}$$

In our implementation, $g_{\min} = 0.15$ and $\gamma = 1.0$ by default.

Aggregation head. For this ablation, we replace the default token-weighted sum with a dedicated learnable attention pooling head:

$$z = \mathrm{MHA}\Big(\mathrm{LN}(q),\, \mathrm{LN}(X^{(L)}),\, \mathrm{LN}(X^{(L)})\Big), \quad q \in \mathbb{R}^{1\times 1\times D}, \tag{16}$$

followed by projection to the embedding space:

$$f = \mathrm{Proj}(z). \tag{17}$$

This provides a strong reversed-polarity conditioning control—anchor-query patch suppression followed by learned aggregation—for comparison with the default patch-query reconstruction and token-weighted pooling.

6.7 Controlled Masking Experiment Details

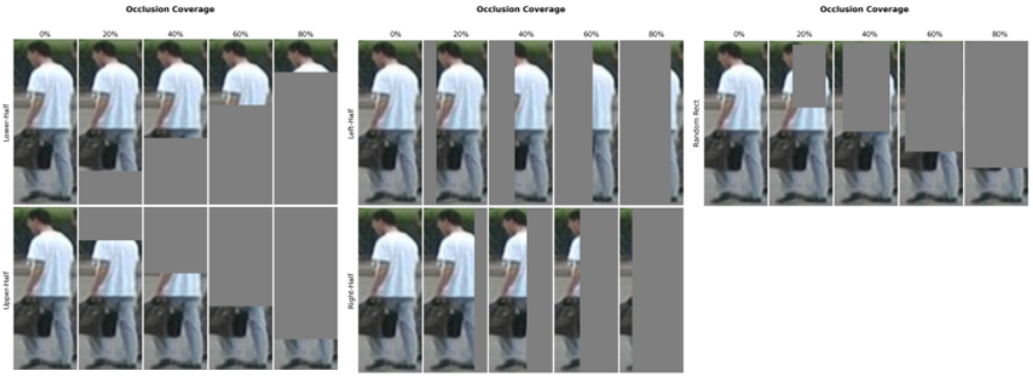


Figure 7: Example of how images are masked for different occlusion types and coverage %.

6.8 Controlled Distractor Experiment Details



Figure 8: Image and mask for the 4 distractors used in our experiments .

Distractor images are drawn from four identities in MSMT17 (disjoint from Market-1501), with foreground masks generated via SAM2 with manual prompting. The segmented foreground is composited directly onto the query image without photometric blending — our goal is not photorealistic multi-person rendering but the introduction of a clear non-target identity signal at controlled coverage levels. To approximate realistic occlusion geometry, distractor insertion is constrained to the lower semicircle around the query subject.



Figure 9: Example of how distractors are composited with query image at various insertion angles and coverage rates.