PRESERVE-AND-COMPOSE TRAINING FOR COMPOSED IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Composed image retrieval (CIR) aims to retrieve images that satisfy a user-specified modification while preserving relevant visual content from a reference image. Collecting target images for this purpose is costly, motivating zero-shot CIR methods that instead use target captions as supervision. However, target captions may omit source details that should be preserved. We therefore propose PACT, *Preserve-and-Compose Training*, which complements target-caption supervision with visual evidence from the source image. PACT learns from image–text–text (ITT) triplets without target images or gallery updates, aligning composed queries with target captions while preserving source evidence through visual supervision. We further introduce Chord scoring, which combines target similarity with source-relative directional agreement in the frozen image space. Results across four ZS-CIR benchmarks show that combining target-caption supervision with source-image evidence leads to strong retrieval performance across datasets, backbone scales, and external galleries.

1 INTRODUCTION

"Show the same person practicing martial arts" is not a complete description of an image. This is the central challenge of composed image retrieval (CIR), which retrieves an image by combining a reference image with a user-specified modification. The instruction specifies an action; the reference image tells the system whom to retrieve. Composed image retrieval (CIR) must combine these complementary inputs: it should change the attributes named in the instruction while retaining relevant visual evidence from the reference. A result that follows the instruction but depicts the wrong person is no more satisfactory than one that preserves the person but ignores the requested action.

Learning this behavior requires supervision about both what should change and what should remain. A matched source–modification–target-image triplet provides a direct example, but finding or generating a target that satisfies both requirements is costly. This motivates learning without explicit target images. Image–caption pairs support visual inversion, and language-based approaches provide inexpensive compositional supervision (Saito et al., 2023; Gu et al., 2024; Byun et al., 2025). More recently, source images have been paired with modification instructions and generated target captions (Li et al., 2024; Li et al., 2026). These alternatives make composition learning possible without collecting a corresponding target image for every instruction.

Target captions nevertheless provide incomplete evidence about the desired result. A caption such as "a person practicing martial arts" can express the requested action without specifying the person's appearance or clothing. Those details may still matter for retrieval, even when neither the instruction nor the generated caption mentions them. The source image contains evidence that can complement this textual target. The challenge is therefore not simply to expose a model to an image and a modification, but to use the available supervision to learn a query that reflects both.

To address this gap, we propose PACT, *Preserve-and-Compose Training*, which combines caption-endpoint learning with source-image and negative-caption supervision. PACT optimizes a lightweight MLP mapper and LoRA parameters in the text encoder, while keeping the image encoder and gallery embeddings fixed. This design learns composition directly in the vision–language embedding space, without requiring target images or modifying the gallery. It also avoids heavy generative models during retrieval, making the model directly applicable to existing galleries.

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.

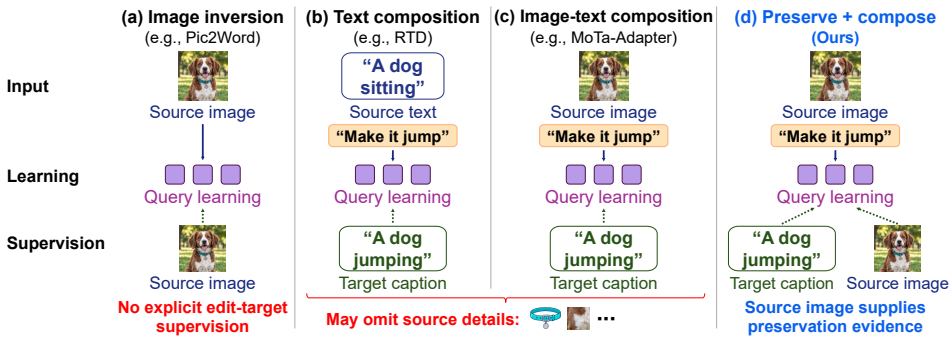

Figure 1: **Supervise both the requested change and the available source evidence.** Training triplets pair a source image with a modification instruction and a target caption, providing a modification target together with a real visual teacher. The diagram characterizes supervision signals, not measured failure rates of all methods in each family.

Figure 1 contrasts the supervision used by different approaches to learn composed queries. Image inversion learns a query representation from the source image without explicit edit-target supervision, while text composition uses a target caption to supervise the desired change. Image–text composition additionally conditions the query on the source image, but still uses the target caption as the primary learning target.

PACT instead assigns complementary roles to the two available sources of supervision. The target caption specifies the desired modification, while the source image provides direct evidence of what should be preserved. Accordingly, each composed query is aligned with its target caption, whereas the shared representation across multiple modifications of the same source is aligned with the source-image embedding. This encourages the query to reflect the requested change without discarding visual information that the target caption leaves unspecified.

We also introduce *Chord scoring* to complement composition learning at retrieval time. Standard cosine similarity measures how well a candidate matches the composed query, but does not account for whether its change from the reference is consistent with the requested modification. Chord augments query–candidate similarity with agreement between the change from the identity query to the modified query and the change from the reference image to each candidate. While related directional matching has been explored previously (Tursun et al., 2026; Wang et al., 2025), Chord combines this change-based agreement with standard similarity and can be applied to existing query models without retraining. We evaluate both its overall retrieval gains and its ability to resolve ambiguous rankings under cosine similarity.

PACT achieves state-of-the-art or competitive results across four CIR benchmarks, with strong gains on CIRR and CIRCO. Beyond benchmark scores, qualitative retrieval over external image and movie galleries shows that PACT better preserves reference-specific visual identity while following the requested modification, compared with a strong prior method. PACT is also highly efficient to train. The ViT-L/14 model requires less than one hour on a single RTX 3090, while keeping the image encoder and gallery representations frozen. Together, these results demonstrate that PACT combines strong retrieval accuracy, visual preservation, and practical training efficiency.

Our contributions are:

- We introduce PACT, a preserve-and-compose training framework that combines target-caption supervision with source-image grounding, enabling CIR without target-image supervision or gallery updates.
- We propose *Chord scoring*, a training-free retrieval rule that jointly considers query–candidate similarity and whether the candidate changes from the reference consistently with the requested modification.
- PACT achieves state-of-the-art or competitive performance across four CIR benchmarks, preserves reference-specific visual identity in large external galleries.

2 Related Work

2.1 CIR without target-image supervision

Supervised CIR typically relies on image–text–image (ITI) triplets, but collecting target images at scale is costly. Zero-shot CIR therefore replaces target-image supervision with more accessible signals. Image–text (IT) methods such as Pic2Word (Saito et al., 2023) and SEARLE (Baldrati et al., 2023) learn visual inversion from image–caption pairs, while Context-I2W (Tang et al., 2024), MoA (Li et al., 2025b), HIT (Li et al., 2025a), PrediCIR (Tang et al., 2025), and FoCo (Zhang et al., 2026b) further improve visual representation or proxy-task design. Language-only approaches include LinCIR (Gu et al., 2024) and RTD (Byun et al., 2025), with RTD explicitly learning from text–text–text (TTT) triplets. Image–text–text (ITT) methods such as MCL (Li et al., 2024) and MoTa-Adapter (Li et al., 2026) retain the source image while replacing the target image with a caption. CIReVL (Karthik et al., 2024) and CoLLM (Huynh et al., 2025) instead employ large language models during retrieval. PACT also learns from ITT triplets, but uses the target caption to supervise the requested composition and the source image as direct visual supervision for what should be preserved.

2.2 Retrieval scoring for composed queries

Most CIR methods rank gallery images by cosine similarity to the composed query, while several approaches exploit geometric structure in the embedding space beyond standard cosine matching. Slerp (Jang et al., 2024) combines image and text representations through spherical interpolation, PDV (Tursun et al., 2026) uses prompt-induced directions to guide query composition, and DIPNEC (Wang et al., 2025) relates modification text to reference-to-candidate visual changes. These approaches incorporate geometric cues into query construction or representation learning. In contrast, we propose *Chord scoring*, a retrieval method for CIR that leaves the learned query representation unchanged and uses reference-relative change only at ranking time, together with standard query–candidate similarity. Chord can therefore be applied to existing CIR models without additional models or retraining, while retaining the same asymptotic retrieval complexity as cosine scoring.

3 Method

PACT learns composed queries from image–text–text (ITT) supervision without requiring target images or modifying gallery representations. Given a source image and a modification instruction, we aim to construct a query that reflects the requested change while retaining visual evidence from the source that may not be expressed in text. Figure 2 summarizes the overall framework. Section 3.1 describes how we construct ITT training data. Section 3.2 introduces the caption-guided composition objective $\mathcal{L}_{\text{compose}}$. Section 3.3 complements this textual supervision with source-image preservation and negative-caption discrimination through $\mathcal{L}_{\text{preserve}}$. Finally, Section 3.4 presents Chord scoring, our retrieval method for incorporating reference-relative change at inference.

3.1 Constructing image–text–text training data

Each *image–text–text (ITT) triplet* consists of a source image I_i, a modification instruction $m_{i,k}$, and a target caption $c_{i,k}^{\text{tgt}}$ describing the desired result. We use CC3M images (Sharma et al., 2018) with their associated source captions, including DreamLIP recaptions (Zheng et al., 2024). This replaces target-image supervision with a textual description of the desired result.

We construct ITT triplets in two ways. First, we create modification–target-caption pairs through rule-based substitution of explicit attributes in source captions. Second, we use Qwen3.5-9B (Qwen Team, 2026) to generate additional pairs directly from source images. For each image-conditioned modification, we also request a negative caption $c_{i,k}^{-}$ that preserves relevant source information but does not satisfy the requested modification. We remove malformed or overlength records, duplicate instructions, no-change or invalid generations, and negative captions that fail validation. The resulting corpus contains 313,240 source images and 856,089 modification triplets, with up to three retained modifications per source. Appendix A provides further details on generation and filtering.

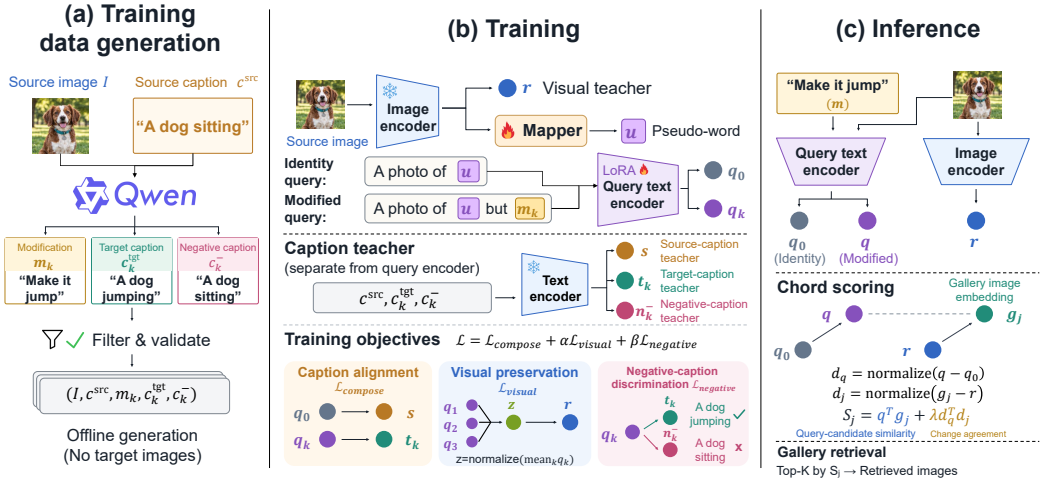

Figure 2: **PACT. (a)** We construct ITT training data from a source image, modification instructions, target captions, and optional negative captions, without target images. **(b)** During training, target captions supervise the composition, source images provide visual supervision for information shared across modifications, and negative captions discourage failed modifications. **(c)** At inference, Chord scoring combines query–candidate similarity with reference-relative change for retrieval.

3.2 Caption-guided composition: $\mathcal{L}_{\text{compose}}$

Consider an ITT triplet (I, m, c^{tgt}) with source caption c^{src}. We additionally use an identity entry $(I, \varnothing, c^{\text{src}})$, allowing the same query model to represent the source before and after modification.

Let E_I and E_T^0 denote the frozen CLIP image and text encoders (Radford et al., 2021), whose outputs lie in a shared d-dimensional embedding space. We define

$$r = E_I(I), \qquad s = E_T^0(c^{\text{src}}), \qquad t = E_T^0(c^{\text{tgt}}). \tag{1}$$

Here, r, s, and t represent the source image, source caption, and target caption, respectively.

Inspired by Pic2Word (Saito et al., 2023), a lightweight MLP mapper M_ϕ converts the source-image feature into an image-dependent pseudo-word embedding $u = M_\phi(E_I(I))$. We define $\mathcal{P}(u, m)$ as the prompt embedding sequence corresponding to a template such as "a photo of $[u]$ but $[m]$", where $[u]$ is replaced by u and $[m]$ contains the token embeddings of the modification instruction. The resulting query representation is

$$q = f_{\theta,\phi}(I, m) = E_T^\theta(\mathcal{P}(u, m)), \tag{2}$$

where E_T^θ is the CLIP text encoder adapted with attention LoRA (Hu et al., 2022). For the identity query $q_0 = f_{\theta,\phi}(I, \varnothing)$, the modification text is omitted. The pseudo-word carries information from the source image, while the modification text specifies how that content should change. The image encoder and teacher text encoder remain frozen; only the mapper and LoRA parameters are learned.

We align identity queries q_0 with source-caption embeddings s, and modified queries q with target-caption embeddings t. For a minibatch of B such pairs, we stack the query and caption embeddings into $Q, T \in \mathbb{R}^{B\times d}$, respectively, with matching rows as positives and all other in-batch pairs as negatives. We minimize

$$\mathcal{L}_{\text{compose}} = \mathcal{C}_{\tau_c}(Q, T), \tag{3}$$

where

$$\mathcal{C}_{\tau_c}(Q, T) = -\frac{1}{2B}\sum_{a=1}^{B}\left[\log\frac{\exp(Q_a^\top T_a/\tau_c)}{\sum_{b=1}^{B}\exp(Q_a^\top T_b/\tau_c)} + \log\frac{\exp(T_a^\top Q_a/\tau_c)}{\sum_{b=1}^{B}\exp(T_a^\top Q_b/\tau_c)}\right]. \tag{4}$$

This objective learns the source state and desired composition from captions. However, captions may omit source details, motivating the preservation objective in Section 3.3.

3.3 PRESERVATION OBJECTIVE: $\mathcal{L}_{\text{preserve}}$

For a source image I_i with retained modifications $\{m_{i,k}\}_{k=1}^{K_i}$, we denote the corresponding queries by $q_{i,k} = f_{\theta,\phi}(I_i, m_{i,k})$ and the source-image representation by $r_i = E_I(I_i)$.

Visual preservation. Target captions may omit visual details that should remain after modification. Rather than aligning each modified query directly with the source image, which could interfere with the requested change, we supervise the information shared across multiple modifications of the same source.

For $\mathcal{L}_{\text{visual}}$, we form a preservation batch in which some source images contribute multiple modified queries. Let $\mathcal{K}_i$ denote the modifications of source I_i appearing in the batch. For $|\mathcal{K}_i| \geq 2$, we define their shared representation as

$$z_i = \text{normalize}\left(\frac{1}{|\mathcal{K}_i|}\sum_{k\in\mathcal{K}_i} q_{i,k}\right). \tag{5}$$

The identity query is excluded, and sources appearing only once do not contribute to this loss.

Let $r_i = E_I(I_i)$ denote the frozen visual teacher for source I_i, and let $\mathcal{S}$ be the source images represented in the preservation batch. We contrastively align z_i with its source-image teacher:

$$\mathcal{L}_{\text{visual}} = -\,\text{mean}_{i:\,|\mathcal{K}_i|\geq 2} \log \frac{\exp(z_i^\top r_i/\tau_v)}{\sum_{j\in\mathcal{S}} \exp(z_i^\top r_j/\tau_v)}. \tag{6}$$

Thus, the source image supervises visual evidence shared across different modifications, including details that may be absent from any individual target caption. For clarity, Equation 6 shows one visual teacher per source; the implementation retains repeated and near-identical positives in the preservation batch, as detailed in Appendix B.2.

Negative-caption discrimination. Preserving source evidence should not prevent the requested modification. For a modification $m_{i,k}$ with target caption $c_{i,k}^{\text{tgt}}$ and valid negative caption $c_{i,k}^{-}$, define

$$t_{i,k} = E_T^0(c_{i,k}^{\text{tgt}}), \qquad n_{i,k}^{-} = E_T^0(c_{i,k}^{-}). \tag{7}$$

The negative caption preserves relevant source information but fails to satisfy the requested modification. Let $\mathcal{V}$ denote the set of entries with negative captions; we minimize

$$\mathcal{L}_{\text{negative}} = -\frac{1}{|\mathcal{V}|}\sum_{(i,k)\in\mathcal{V}} \log \frac{\exp(q_{i,k}^\top t_{i,k}/\tau_n)}{\exp(q_{i,k}^\top t_{i,k}/\tau_n) + \exp(q_{i,k}^\top n_{i,k}^{-}/\tau_n)}. \tag{8}$$

The two terms play complementary roles: $\mathcal{L}_{\text{visual}}$ preserves source evidence, while $\mathcal{L}_{\text{negative}}$ discourages failure to make the requested change. We define

$$\mathcal{L}_{\text{preserve}} = \alpha\mathcal{L}_{\text{visual}} + \beta\mathcal{L}_{\text{negative}}, \tag{9}$$

and optimize

$$\boxed{\mathcal{L}_{\text{PACT}} = \mathcal{L}_{\text{compose}} + \mathcal{L}_{\text{preserve}}}. \tag{10}$$

No explicit transition-alignment loss is required.

3.4 INFERENCE WITH CHORD SCORING

Standard CIR retrieval ranks gallery images by cosine similarity to the composed query. While this measures how well a candidate matches the desired result, it does not consider whether the candidate differs from the reference consistently with the requested modification. We therefore introduce *Chord scoring*, which incorporates this reference-relative information at ranking time.

Given a reference image I, modification m, and gallery image G_j, we compute

$$r = E_I(I), \qquad q = f_{\theta,\phi}(I, m), \qquad q_0 = f_{\theta,\phi}(I, \varnothing), \qquad g_j = E_I(G_j), \tag{11}$$

where $f_{\theta,\phi}$ denotes the query encoder formed by M_ϕ and E_T^θ. Chord scoring combines standard query–candidate similarity with agreement between two changes in the shared embedding space:

$$S_j = \cos(q, g_j) + \lambda \cos(q - q_0,\ g_j - r). \tag{12}$$

The first term measures similarity between the composed query and candidate, while the second measures whether the candidate changes from the reference consistently with the query change. Because these differences connect pairs of unit embeddings, they form chords of the embedding sphere, motivating the name *Chord scoring*. The cosine term compares the directions of these two changes, independent of their magnitudes.

The weight $\lambda \geq 0$ controls the contribution of this reference-relative term and is selected on validation data for each query model. Chord leaves both the learned query representation and gallery embeddings unchanged, allowing it to be applied without an additional model or retraining. For a gallery of N d-dimensional embeddings, Chord retains the same $O(Nd)$ asymptotic complexity as cosine scoring, with only an additional candidate-wise cosine computation.

4 Experiments

4.1 Evaluation protocol

We evaluate PACT on CIRR (Liu et al., 2021), CIRCO (Baldrati et al., 2023), FashionIQ (Wu et al., 2021), and GeneCIS (Vaze et al., 2023). Unless otherwise stated, all comparisons use ViT-L/14, and PACT results are averaged over five independent runs with different random seeds. We report official test results on CIRR and CIRCO and validation results on FashionIQ and GeneCIS, with FashionIQ recall averaged equally across categories. Baselines follow the retrieval procedures of their source evaluations, whereas PACT uses Chord scoring; accordingly, the main benchmark tables compare complete retrieval systems rather than training objectives in isolation. For any model evaluated with Chord scoring, we select a single weight λ on validation using the mean of CIRR R@1, CIRCO mAP@5, and FashionIQ category-average R@10, and fix it for test and transfer evaluation. Section 3.4 separately isolates the effect of Chord scoring on models for which checkpoints are available. Appendix C provides full evaluation details.

4.2 Benchmark performance

Table 1: **CIRR/CIRCO test and FashionIQ validation, ViT-L/14.** Published baselines and official PACT test results; FashionIQ reports category-average recall.

Method	CIRR test			CIRCO test			FashionIQ val	
	R@1	R@5	R@10	mAP@5	mAP@10	mAP@25	Avg R@10	Avg R@50
Pic2Word (CVPR 2023)	23.90	51.70	65.30	8.72	9.51	10.64	24.70	43.70
SEARLE-XL (ICCV 2023)	24.24	52.48	66.29	11.68	12.73	14.33	25.56	46.23
Context-I2W (AAAI 2024)	25.60	55.10	68.50	13.00	14.60	16.10	27.80	48.90
CIReVL (GPT-3.5) (ICLR 2024)	24.55	52.31	64.92	18.57	19.01	20.89	28.55	48.57
LinCIR (CVPR 2024)	25.04	53.25	66.68	12.59	13.58	15.00	26.28	46.49
PrediCIR (CVPR 2025)	27.20	57.00	70.20	15.70	17.10	18.60	30.10	52.30
MoA (SIGIR 2025)	27.10	56.50	69.20	15.30	17.10	18.50	30.10	51.30
HIT* (ICCV 2025)	27.90	57.60	70.50	15.50	16.70	18.90	30.30	51.00
RTD (ICCV 2025)	26.63	56.17	68.96	17.11	18.11	20.06	30.24	51.08
MoTa-Adapter (AAAI 2026)	28.02	58.62	71.06	18.22	19.46	–	28.76	49.53
De-MINDS (AAAI 2026)	30.00	59.70	**74.50**	21.00	22.40	22.70	32.20	**54.50**
FlowCIR (ECCV 2026)	26.20	56.10	68.60	14.90	15.70	17.30	29.70	48.90
PACT (ours)	**31.24**	**60.39**	71.66	**23.59**	**24.52**	**26.89**	**32.43**	52.29

Composed image retrieval. PACT achieves strong performance across all three benchmarks in Table 1. The largest gains are observed on CIRCO, where PACT achieves the best results among the compared methods across all reported metrics. At mAP@5, PACT reaches 23.59, exceeding the strongest baseline, De-MINDS, by 2.59 points and CIReVL by 5.02 points; at mAP@25, the margin over De-MINDS increases to 4.19 points. On CIRR, PACT achieves the highest R@1 (31.24) and R@5 (60.39), while De-MINDS remains higher at R@10. On FashionIQ, PACT achieves the highest category-average R@10 (32.43), while De-MINDS remains higher at R@50 (54.50 versus 52.29).

Table 2: **GeneCIS validation, ViT-L/14.** Local and published baseline results; – denotes unreported metrics.

Method	Focus			Change			Attribute			Object		
	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3
Pic2Word (CVPR 2023)	13.36	24.84	34.66	13.95	24.67	34.51	14.93	26.27	35.87	12.37	23.24	33.29
SEARLE-XL (ICCV 2023)	14.29	25.39	36.09	14.36	25.00	33.67	16.81	27.56	37.59	11.84	22.83	32.17
CIReVL (GPT-3.5) (ICLR 2024)	15.90	26.80	36.25	15.80	27.45	36.40	16.95	28.90	38.60	14.75	25.35	34.05
LinCIR (CVPR 2024)	14.55	26.18	35.71	15.08	26.77	36.53	16.37	29.17	38.61	13.27	23.78	33.62
RTD (ICCV 2025)	15.38	26.17	36.61	15.86	27.38	36.74	17.47	29.42	39.91	13.78	24.13	33.44
MoTa-Adapter (AAAI 2026)	Reported GeneCIS R@1 summary: 16.7; task marginals unreported											
DiffComp (CVPR 2026)	17.95	**29.73**	–	15.95	26.68	–	19.00	30.39	–	14.89	**26.02**	–
PACT (ours)	**18.15**	28.54	**38.96**	**16.96**	**28.73**	**38.77**	**20.06**	**31.90**	**42.18**	**15.06**	25.38	**35.55**

Transfer to conditional image similarity. GeneCIS evaluates conditional similarity under conditions that focus on or change objects and attributes. PACT achieves the highest R@1 across all four reported marginals in Table 2. PACT consistently outperforms the locally evaluated Pic2Word, LinCIR, and RTD models across the displayed recall metrics. Compared with DiffComp, PACT is higher on all four R@1 metrics, while DiffComp is higher on Focus R@2 (29.73 versus 28.54) and Object R@2 (26.02 versus 25.38). The Chord weight is transferred from CIRR, CIRCO, and FashionIQ without tuning on GeneCIS, indicating that the learned retrieval behavior generalizes beyond the standard CIR benchmarks.

4.3 Ablation of training objectives and retrieval scoring

Table 3: **Ablation of PACT training objectives and retrieval scoring.** Validation performance with ViT-L/14. The mean averages CIRR R@1, CIRCO mAP@5, and FashionIQ R@10. The cosine comparison uses the same checkpoints as the corresponding Chord results.

Training configuration	CIRR R@1	CIRCO mAP@5	FIQ R@10	Mean
Chord scoring				
PACT (full)	**32.50**	**24.18**	**32.43**	**29.70**
$\mathcal{L}_{\text{compose}}$ only	29.47	23.80	29.19	27.49
$\mathcal{L}_{\text{preserve}}$ only	28.58	16.12	29.45	24.72
$\mathcal{L}_{\text{compose}} + \alpha\mathcal{L}_{\text{visual}}$	30.40	23.64	30.47	28.17
$\mathcal{L}_{\text{compose}} + \beta\mathcal{L}_{\text{negative}}$	31.26	22.83	29.23	27.77
Cosine scoring (matched checkpoints)				
PACT (full)	**25.76**	**18.68**	**28.29**	**24.24**
$\mathcal{L}_{\text{compose}}$ only	23.10	16.92	24.55	21.52

Composition and preservation provide complementary supervision. With Chord scoring, $\mathcal{L}_{\text{compose}}$ alone achieves a three-benchmark mean of 27.49, outperforming $\mathcal{L}_{\text{preserve}}$ alone (24.72), showing that preservation cannot replace supervision for the requested change. Adding either $\mathcal{L}_{\text{visual}}$ or $\mathcal{L}_{\text{negative}}$ gives a partial gain, while combining both yields the best mean of 29.70 and improves all three benchmarks over composition-only training. Neither auxiliary term alone recovers the full improvement, supporting their complementary roles in preserving source evidence and enforcing the requested modification. Thus, composition provides the primary learning signal, with preservation providing additional supervision.

PACT gains persist without Chord. To separate training from retrieval, we evaluate the same checkpoints with standard cosine similarity. PACT improves the mean over composition-only training by 2.72 points with cosine (21.52→24.24) and by 2.21 points with Chord (27.49→29.70). The gains under both retrieval rules show that the benefit of PACT is not specific to Chord. The improvement is also consistent across CIRR, CIRCO, and FashionIQ under both scoring rules. Chord adds 5.46 points to PACT and 5.97 to composition-only training. Thus, PACT improves the learned query representation, while Chord provides an additional inference-time gain. Per-dataset scoring comparisons are provided in Appendix D.

4.4 Effectiveness and behavior of Chord scoring

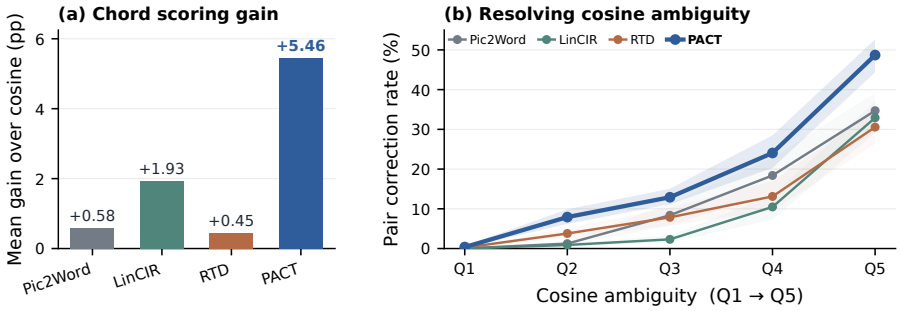

Figure 3: **Chord scoring improves retrieval and resolves ambiguous cosine rankings. (a)** Mean improvement over cosine across CIRR R@1, CIRCO mAP@5, and FashionIQ R@10, using one validation-selected λ per model. **(b)** Correction rate for positive–candidate pairs misordered by cosine similarity at a common $\lambda = 0.34$. Q1–Q5 correspond to decreasing cosine deficits; shading denotes 95% reference-cluster bootstrap intervals. Appendix D provides full details.

Chord scoring generalizes across query models. Figure 3a applies Chord post hoc to Pic2Word, LinCIR, RTD, and PACT without retraining. Relative to cosine retrieval, it improves the three-benchmark mean by 0.58, 1.93, 0.45, and 5.46 points, respectively. Thus, reference-relative scoring benefits multiple query models, with the largest gain observed for PACT.

Reference-relative change resolves cosine ambiguity. For a positive p and the highest-cosine non-positive candidate n, let

$$\delta = \cos(q, g_n) - \cos(q, g_p) > 0, \qquad D_j = \cos(q - q_0, g_j - r).$$

Then

$$S_p - S_n = -\delta + \lambda(D_p - D_n). \tag{13}$$

Chord reverses the cosine ordering when the change-agreement term outweighs the cosine deficit. Figure 3b shows higher correction rates for smaller deficits, indicating that reference-relative change is most useful when cosine rankings are ambiguous. This trend is consistent across all four query models, suggesting that the directional term acts as a complementary signal rather than replacing query–candidate similarity. In particular, it can break near-ties by favoring candidates whose change from the reference better agrees with the requested query change.

4.5 Model size and efficiency

Strong retrieval with a compact inference model. PACT-L/14 achieves 23.59 CIRCO test mAP@5 with approximately 443M inference parameters. It outperforms CoLLM-L/14 (20.30), CIReVL-L/14 (18.57), LinCIR-G/14 (19.71), and RTD-G/14 (21.08), while nearly matching PrediCIR-G/14 (23.70) with at least 5.7× fewer parameters. Notably, PACT-L/14 remains competitive with methods using substantially larger backbones or additional language-model components, despite retaining a compact CLIP-based inference pipeline. PACT also outperforms MCL variants using 2.975–7.068B parameters, which achieve 14.55–17.67 mAP@5 (Li et al., 2024). Across these comparisons, increasing inference model size does not by itself translate into stronger retrieval performance. These results show that PACT achieves strong retrieval without requiring a large language model at inference.

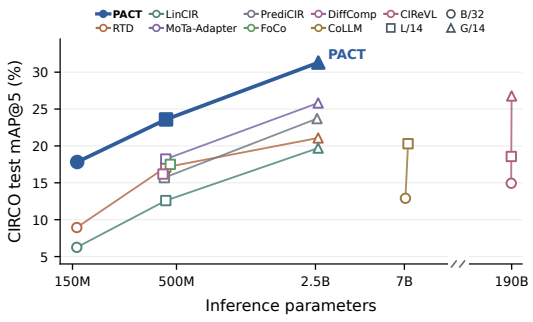

Figure 4: **Performance versus model size.** PACT remains strong with a compact inference model.

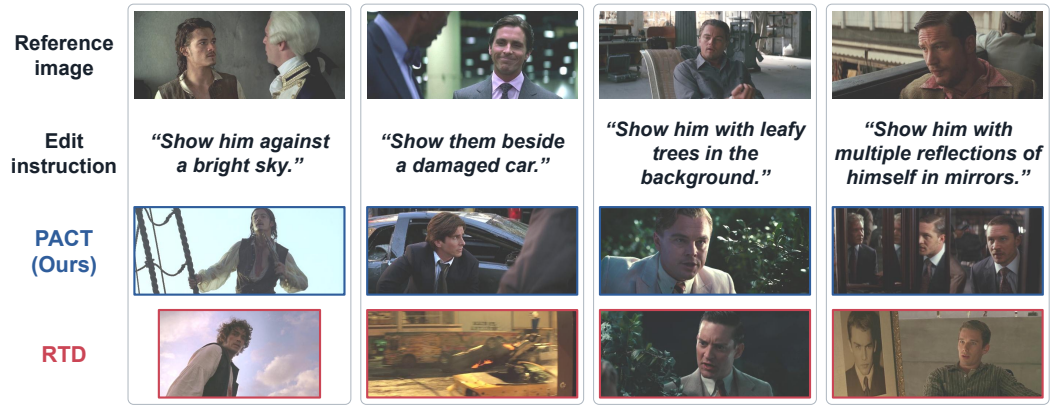

Figure 5: **Reference-dependent retrieval on MovieNet.** Selected top-1 disagreements between PACT and RTD. PACT better preserves the referenced identity while following scene- and background-level modifications.

Training efficiency. Using a batch size of 512, PACT with a ViT-L/14 backbone completes a 333-step identity warmup followed by 1,000 joint-training steps in only ∼**52 min** on a single **RTX 3090**, with **8.19 GiB** peak allocated GPU memory. This efficiency follows from keeping the image encoder frozen and adapting only lightweight query-side components, without target-image supervision or gallery re-encoding. Thus, PACT can substantially adapt a large CLIP retrieval model with modest training compute and memory. The measurement excludes evaluation, checkpoint saving, and offline data preparation.

4.6 Reference preservation beyond the benchmarks

Standard CIR benchmarks do not directly measure whether reference-specific visual identity is preserved after an edit. We therefore construct a MovieNet (Huang et al., 2020) evaluation set where edits change the scene, background, or action while the referenced person remains unchanged. We measure whether the Top-1 retrieval contains the same annotated identity as the reference; evaluation details are provided in Appendix E.

Table 4: MovieNet Top-1 identity (%).

Method	Cosine	Chord
RTD	64.0	64.5
$\mathcal{L}_{\text{compose}}$	58.0	62.0
$+\alpha\mathcal{L}_{\text{visual}}$	78.5	81.5
PACT (full)	**80.5**	**82.5**

Adding $\mathcal{L}_{\text{visual}}$ to composition substantially improves identity preservation under both cosine and Chord retrieval. Full PACT retains this gain and also exceeds RTD, supporting the intended role of $\mathcal{L}_{\text{preserve}}$ in preserving reference-specific visual evidence. Figure 5 shows complementary qualitative examples from a separate MovieNet gallery (Appendix F). PACT more consistently retains the reference identity while following the edit, illustrating the preservation behavior measured above.

5 Conclusion

We introduced PACT, a preserve-and-compose training framework for composed image retrieval without target-image supervision or gallery updates. PACT aligns composed queries with target captions while using the source image to preserve visual evidence that may be omitted from text, with negative-caption discrimination further encouraging the requested modification. At inference, Chord scoring complements query–candidate similarity with reference-relative change agreement and improves multiple query models without retraining. PACT achieves strong performance across CIRR, CIRCO, FashionIQ, and GeneCIS, with particularly strong results on CIRCO. Qualitative retrieval on external movie galleries further illustrates that PACT can follow scene-level edits while retaining reference-specific visual identity. Overall, the results support jointly modeling what should change and what should be preserved for effective composed image retrieval.

AI USE STATEMENT

We used Qwen models to generate training instructions and captions, as detailed in Appendix A. We also used Qwen3.5-27B to generate and validate MovieNet evaluation queries and assess modification fulfillment with model identities hidden. Identity preservation was evaluated using official MovieNet annotations. Additionally, we used generative AI tools to brainstorm research ideas and assist with experimental design.

ETHICS STATEMENT

Our experiments use existing datasets and pretrained models, whose biases and errors may affect generated supervision and retrieval results. Automated filtering and evaluation do not eliminate these limitations. Person-related retrieval also raises privacy concerns if applied to tracking individuals without consent. Our MovieNet experiments evaluate character preservation within movie frames and do not establish suitability for real-world person identification.

REPRODUCIBILITY STATEMENT

Section 3 describes the model, training objectives, and Chord scoring. Appendices A and B detail data construction, filtering, and optimization. The appendix also documents evaluation protocols, baseline sources, random seeds, scoring-parameter selection, variation across training runs, and MovieNet evaluation procedures.

REFERENCES

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *ICCV*, pp. 15338–15347, 2023. URL https://github.com/miccunifi/SEARLE.

Jaeseok Byun, Seokhyeon Jeong, Wonjae Kim, Sanghyuk Chun, and Taesup Moon. An efficient post-hoc framework for reducing task discrepancy of text encoders for composed image retrieval. In *ICCV*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Byun_An_Efficient_Post-hoc_Framework_for_Reducing_Task_Discrepancy_of_Text_ICCV_2025_paper.html.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *CVPR*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Cherti_Reproducible_Scaling_Laws_for_Contrastive_Language-Image_Learning_CVPR_2023_paper.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. URL https://arxiv.org/abs/2010.11929.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. Scaling Deep Contrastive Learning Batch Size under Memory Limited Setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP*, 2021. URL https://aclanthology.org/2021.repl4nlp-1.31/.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *CVPR*, 2024. URL https://arxiv.org/abs/2312.01998.

Zhenqi He, Ziqi Jiang, Yuanpei Liu, Yanghao Wang, Teng Wang, and Long Chen. FlowCIR: Semantic Transport via Flow Matching for Zero-Shot Composed Image Retrieval. In *ECCV*, 2026. URL https://arxiv.org/abs/2607.02284.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. MovieNet: A Holistic Dataset for Movie Understanding. In *ECCV*, 2020. URL https://arxiv.org/abs/2007.10937.

Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Raffay Hamid, Trishul Chilimbi, and Abhinav Shrivastava. Collm: A large language model for composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, and Ser-Nam Lim. Spherical Linear Interpolation and Text-Anchoring for Zero-shot Composed Image Retrieval. ECCV, 2024. URL https://arxiv.org/abs/2405.00571.

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.09291.

Haiwen Li, Delong Liu, Zhaohui Hou, Zeliang Ma, Fei Su, and Zhicheng Zhao. Modality and Task Adaptation for Enhanced Zero-shot Composed Image Retrieval. AAAI, 2026. URL https://ojs.aaai.org/index.php/AAAI/article/view/37534.

Wei Li, Hehe Fan, Yongkang Wong, Yi Yang, and Mohan Kankanhalli. Improving context understanding in multimodal large language models via multimodal composition learning. In *International Conference on Machine Learning*, pp. 27732–27751, 2024.

Zhe Li, Lei Zhang, Zheren Fu, Kun Zhang, and Zhendong Mao. Hierarchy-Aware Pseudo Word Learning with Text Adaptation for Zero-Shot Composed Image Retrieval. ICCV, 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Li_Hierarchy-Aware_Pseudo_Word_Learning_with_Text_Adaptation_for_Zero-Shot_Composed_ICCV_2025_paper.pdf.

Zhe Li, Lei Zhang, Kun Zhang, Weidong Chen, Yongdong Zhang, and Zhendong Mao. Rethinking pseudo word learning in zero-shot composed image retrieval: From an object-aware perspective. In *SIGIR*, pp. 833–843, 2025b. doi: 10.1145/3726302.3730074.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, 2021. URL https://arxiv.org/abs/2108.04024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Qwen Team. Qwen3 Technical Report. arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Qwen Team. Qwen3.5: Towards native multimodal agents. Official model release, 2026. URL https://qwen.ai/blog?id=qwen3.5.

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 *ICML*, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2Word: Mapping Pictures to Words for Zero-shot Composed Image Retrieval. CVPR, 2023. URL https://arxiv.org/abs/2302.03084.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS Datasets and Benchmarks*, 2022. URL https://arxiv.org/abs/2210.08402.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238/.

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. Context-I2W: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *AAAI*, 2024. URL https://arxiv.org/abs/2309.16137.

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Gaopeng Gou, and Qi Wu. Missing Target-Relevant Information Prediction with World Model for Accurate Zero-Shot Composed Image Retrieval. CVPR, 2025. URL https://arxiv.org/abs/2503.17109.

Yuanmin Tang, Jing Yu, Keke Gai, Gang Xiong, Gaopeng Gou, Meikang Qiu, and Qi Wu. Manipulation Intention Understanding for Zero-Shot Composed Image Retrieval. AAAI, 2026. URL https://ojs.aaai.org/index.php/AAAI/article/view/37907.

Osman Tursun, Sinan Kalkan, Simon Denman, and Clinton Fookes. PDV: Prompt Directional Vectors for Zero-shot Composed Image Retrieval. WACV, 2026. URL https://arxiv.org/abs/2502.07215.

Sagar Vaze, Nicolas Carion, and Ishan Misra. GeneCIS: A benchmark for general conditional image similarity. In *CVPR*, 2023. URL https://arxiv.org/abs/2306.07969.

Yifan Wang, Wuliang Huang, and Chun Yuan. Aligning composed query with image via discriminative perception from negative correspondences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8078–8086, 2025. doi: 10.1609/aaai.v39i8.32871.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *CVPR*, pp. 11307–11317, 2021. URL https://arxiv.org/abs/1905.12794.

Jingjing Zhang, Lei Zhang, Zheren Fu, Bo Hu, and Zhendong Mao. Self-guided Semantic Inspection for Zero-Shot Composed Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33881–33890, June 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_Self-guided_Semantic_Inspection_for_Zero-Shot_Composed_Image_Retrieval_CVPR_2026_paper.html.

Jingjing Zhang, Lei Zhang, Zheren Fu, and Zhendong Mao. Learning to compose: Revisiting proxy task design for zero-shot composed image retrieval. In *ECCV*, 2026b. URL https://arxiv.org/abs/2607.00374.

Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/2666_ECCV_2024_paper.php.

A Training data construction

A.1 Source images and supervision

We construct image–text–text triplets from CC3M images (Sharma et al., 2018) and their associated source captions, including DreamLIP recaptions (Zheng et al., 2024). Each triplet contains a source image, a modification instruction, and a target caption. The source image provides visual supervision, while the target caption describes the requested composition. Target images are not required.

The training corpus contains 313,240 source images and 856,089 modification triplets. We additionally form one identity entry per source, pairing the image with its source caption and an empty modification. The resulting composition-training pool contains 1,169,329 entries. Identity targets use the source captions associated with the images. Data generation and feature preparation are performed offline.

A.2 Modification generation

Caption-based construction combines lexical substitutions of explicit attributes, including color, material, pattern, and count, with language-model generation. Qwen3-8B (Qwen Team, 2025) phrases exact-span modifications as natural instructions, while Qwen3.5-27B (Qwen Team, 2026) proposes instruction–target-caption pairs from source captions. Together, these routes provide one caption-based modification record per source. Image-conditioned generation uses Qwen3.5-9B to produce further modification–target-caption pairs directly from source images, along with negative captions. This route contributes 542,849 eligible modification records, giving 856,089 triplets in total.

A.3 Filtering and negative captions

Caption-based modifications undergo consistency and token-length checks. Image-conditioned proposals contain a source description, a modification, and a target caption. We remove malformed outputs, duplicate instructions, no-change modifications, overlength text, and records explicitly rejected by checks of source support, modification applicability, or target-caption consistency. The generated source description is used for these checks; it does not replace the source caption used as the identity target.

The negative-caption objective uses captions that preserve source information but fail to satisfy the modification. A negative is eligible only when its validity flag is true, it is nonempty, it differs from the target caption, and it fits the text-token limit. Negatives that instead violate the intended preservation constraint are excluded. These automatic filters establish eligibility for training; they do not constitute human verification of every retained triplet.

B Implementation details

B.1 Backbones and optimization

PACT uses pretrained vision–language encoders with ViT backbones (Dosovitskiy et al., 2021). ViT-B/32 and ViT-L/14 initialize from OpenAI CLIP (Radford et al., 2021). ViT-G/14 denotes OpenCLIP ViT-bigG/14 (Cherti et al., 2023), pretrained on the LAION-2B English subset (Schuhmann et al., 2022); the pretrained checkpoint is `CLIP-ViT-bigG-14-laion2B-39B-b160k`. The image encoder and teacher text encoder remain frozen. Learning adapts the query-side mapper and text-encoder attention LoRA parameters.

Training comprises 333 identity-warmup updates followed by 1,000 joint updates. During identity warmup, the text encoder is frozen and identity queries are aligned with source captions. The warmup uses symmetric contrastive cross-entropy with temperature 0.07 and masks off-diagonal source-caption similarities above 0.999. Joint training uses the composition objective in Equation 3 together with the preservation objective in Equation 9.

We use AdamW (Loshchilov & Hutter, 2019) with FP32 optimizer moments and BF16 trainable weights. Learning rates increase linearly over the first 100 updates of each stage and remain constant

thereafter. Gradient replay (Gao et al., 2021) processes microbatches of 32 while retaining the full contrastive batch. Table 5 summarizes the ViT-L/14 configuration.

Table 5: **ViT-L/14 training configuration.**

Setting	Value
Identity / joint updates	333 / 1,000
Composition / preservation batch size	512 / 512
Replay microbatch size	32
Warmup mapper learning rate	10^{-4}
Joint mapper / text LoRA learning rate	3×10^{-5} / 10^{-5}
Weight decay	0.001
Text attention LoRA rank	16
Composition temperature τ_c	0.05
Visual / negative temperature τ_v, τ_n	0.03 / 0.07
Preservation weights α, β	1 / 1
Trainable weights / optimizer moments	BF16 / FP32

B.2 Composition and preservation sampling

Each composition batch contains 512 entries drawn independently and uniformly with replacement from the union of identity and modification records. Identity frequency follows the composition of this pool; the sampler imposes neither a fixed identity quota nor identity–modification pairing within a batch. In joint training, Equation 4 treats matched query–caption positions as positives and all off-diagonal positions as negatives, without an additional same-source or duplicate-caption mask.

An independently sampled preservation batch contains 256 caption-based and 256 image-conditioned modification records. Image-conditioned records are sampled with replacement. Of the caption-based records, 128 are selected to share sources with image-conditioned records, and 128 are sampled independently. This construction provides groups of modified queries for the visual objective. The composition and preservation losses are combined before the optimizer update. Because sampling is with replacement, the fixed update schedule does not correspond to an epoch over the eligible training pool.

B.3 Visual supervision with repeated positives

Equation 6 presents the visual objective with one teacher per source. Here we specify its implementation when the preservation batch contains repeated sources. Index the B_v sampled entries by b, with source and modification indices (i_b, k_b), where $k_b \geq 1$. Let

$$\mathcal{J}_i = \{b : i_b = i\}, \qquad \mathcal{G} = \{i : |\mathcal{J}_i| \geq 2\}.$$

For each eligible source, the pooled query is

$$z_i = \text{normalize}\left(\frac{1}{|\mathcal{J}_i|}\sum_{b \in \mathcal{J}_i} q_{i_b,k_b}\right). \tag{14}$$

The cardinality $|\mathcal{J}_i|$ counts sampled occurrences, which may include repeated draws of a modification. Identity queries are excluded. The teacher bank retains the source-image embedding for every sampled entry, including repeated occurrences. With unit-normalized teacher embeddings, the positive positions for source i are

$$\mathcal{P}_i = \{b : i_b = i \text{ or } r_i^\top r_{i_b} > 0.999\}. \tag{15}$$

The implemented objective is

$$\mathcal{L}_{\text{visual}} = -\frac{1}{|\mathcal{G}|}\sum_{i \in \mathcal{G}} \log \frac{\sum_{b \in \mathcal{P}_i} \exp(z_i^\top r_{i_b}/\tau_v)}{\sum_{b=1}^{B_v} \exp(z_i^\top r_{i_b}/\tau_v)}. \tag{16}$$

Thus, repeated or near-identical source images contribute positive mass rather than being treated as negatives. Sources represented once remain in the teacher bank but do not contribute pooled queries. The visual loss is zero when $\mathcal{G}$ is empty.

C Evaluation protocols

C.1 Benchmarks and metrics

CIRR (Liu et al., 2021) and CIRCO (Baldrati et al., 2023) use official test evaluation, while FashionIQ (Wu et al., 2021) and GeneCIS (Vaze et al., 2023) use validation evaluation. CIRR excludes the reference image and uses global top-50 and subset top-3 predictions. CIRCO uses its 123,403-image gallery and top-50 predictions. In our local CIRCO evaluation, the reference image remains in the candidate gallery. FashionIQ retrieval is restricted to the query category. We average query representations from the two caption orders before ranking, and report the unweighted mean of Dress, Toptee, and Shirt recall.

GeneCIS evaluates conditional similarity under instructions that focus on or change objects and attributes. Table 2 reports recall at ranks 1, 2, and 3 for the Focus, Change, Attribute, and Object summaries.

C.2 Sources of GeneCIS baseline results

The GeneCIS comparison combines local evaluation of available checkpoints with published results, as summarized in Table 6. We evaluate Pic2Word, SEARLE-XL, and LinCIR using their released checkpoints. For RTD, we evaluate a checkpoint reproduced using the official code. CIReVL summaries are computed from the task-wise scores reported in its paper. MoTa-Adapter provides an R@1 summary of 16.7, without the task marginals shown for the other methods. DiffComp uses the GeneCIS results reported in its paper. Missing metrics remain unreported; published rows are not described as local reproductions.

Table 6: **Sources of GeneCIS baseline results.**

Method	Evaluation source
Pic2Word (Saito et al., 2023)	Local evaluation of the authors' released checkpoint
SEARLE-XL (Baldrati et al., 2023)	Local evaluation of the authors' released checkpoint
LinCIR (Gu et al., 2024)	Local evaluation of a released checkpoint
RTD (Byun et al., 2025)	Local evaluation of a checkpoint reproduced with official code
CIReVL (Karthik et al., 2024)	Summaries computed from published task-wise scores
MoTa-Adapter (Li et al., 2026)	Published R@1 summary (16.7)
DiffComp (Zhang et al., 2026a)	Published GeneCIS results

C.3 Checkpoint and scoring-weight selection

We select checkpoints and the Chord coefficient using the unweighted validation mean of CIRR R@1, CIRCO mAP@5, and FashionIQ category-average R@10. The coefficient is selected from $\lambda \in \{0, 0.02, \ldots, 1\}$. For each PACT training seed, we select one coefficient that maximizes this validation mean and use it across datasets; test scores do not enter selection. GeneCIS is excluded from selection and uses the same coefficient without further tuning.

The benchmark tables compare complete retrieval systems, with GeneCIS baseline sources specified in Table 6. The Chord analysis instead compares scoring rules while holding each query model and its gallery embeddings fixed. For the training ablations, each configuration selects its coefficient on the same validation grid. The cosine and Chord rows in Table 3 use matched checkpoints, separating the effect of scoring from changes in the trained model.

C.4 Variability across training seeds

Table 7 reports variation across five independent training runs for each backbone. All entries are validation results with Chord scoring. We report sample standard deviations across five seeds, in percentage points. For FashionIQ, we first average the three category scores within each run and then compute the standard deviation across runs. The GeneCIS entry similarly uses the four-task mean for each run. These standard deviations measure training-seed variation and are distinct from the query-sampling bootstrap intervals in Appendix D.

Table 7: **Standard deviation across five training seeds.** Validation results with Chord scoring. Values are sample standard deviations, in percentage points.

Benchmark and metric	ViT-B/32	ViT-L/14	ViT-G/14
CIRR R@1	0.268	0.386	0.319
CIRCO mAP@5	0.344	1.072	0.592
FashionIQ category-average R@10	0.115	0.340	0.327
FashionIQ category-average R@50	0.298	0.476	0.395
GeneCIS four-task mean R@1	0.144	0.384	0.434

Across the metrics and backbones in Table 7, standard deviations range from 0.115 to 1.072 percentage points. CIRCO mAP@5 has the largest standard deviation for each backbone, while the FashionIQ category-average R@10 standard deviation ranges from 0.115 to 0.340 points.

C.5 Additional backbone results

Tables 10 and 11 report PACT performance with ViT-B/32 and ViT-G/14 alongside published baselines; Table 8 provides additional retrieval metrics. PACT uses Chord scoring. We use seeds 1007, 1013, 1019, 1021, and 1031 and report mean $\pm$ sample standard deviation across five seeds. CIRR and CIRCO results are evaluated on their test sets, and FashionIQ results on its validation set. For FashionIQ, each seed's score is the equally weighted mean across the three clothing categories; the reported mean and standard deviation are then computed across these five seed-level scores. GeneCIS summaries likewise average the four tasks within each seed before aggregation across seeds. The Chord coefficient is selected separately for each seed using only the three-benchmark validation objective described above.

Table 8: **Additional PACT retrieval metrics across five training seeds.** Chord scoring; means are in percent and sample standard deviations are in percentage points.

Benchmark	Metric	ViT-B/32	ViT-G/14
CIRR test	R@50	92.34 ± 0.19	95.68 ± 0.05
	Subset R@1	63.46 ± 0.65	70.46 ± 0.43
	Subset R@2	82.42 ± 0.38	87.22 ± 0.35
	Subset R@3	91.24 ± 0.16	93.61 ± 0.17
CIRCO test	mAP@50	19.94 ± 0.23	35.70 ± 0.75

C.6 Category-wise FashionIQ results

Table 9 reports category-wise FashionIQ validation recall with ViT-L/14 and Chord scoring. Each entry is the mean and sample standard deviation across five training seeds. The category-average row summarizes the equally weighted category mean computed within each run.

Table 9: **FashionIQ validation across five training seeds.** ViT-L/14 with Chord scoring; mean $\pm$ sample standard deviation, in percentage points.

Category	R@10	R@50
Dress	27.86 ± 0.49	49.70 ± 0.42
Shirt	34.15 ± 0.53	52.25 ± 0.20
Toptee	35.28 ± 0.30	54.94 ± 0.89
Category average	32.43 ± 0.34	52.29 ± 0.48

C.7 Comparison with published results across backbones

Tables 10 and 11 extend the comparison in Table 1 to ViT-B/32 and ViT-G/14 with the same metrics. Baselines retain the training and scoring procedures of their published evaluations. PACT uses Chord scoring and reports means and sample standard deviations across five seeds; additional metrics are given in Table 8. We retain all baseline methods from Table 1; – indicates that no published value or local reproduction is available for that backbone and metric. SEARLE denotes the corresponding backbone variant of SEARLE-XL. RTD and MoTa-Adapter in Table 1 correspond to the LinCIR-based variants explicitly named here.

Table 10: **CIRR/CIRCO test and FashionIQ validation, ViT-B/32.** PACT cells show the five-seed mean above ± sample standard deviation (%p). Bold marks the highest mean in each column; – indicates no available result for this backbone and metric.

Method	CIRR test			CIRCO test			FashionIQ val	
	R@1	R@5	R@10	mAP@5	mAP@10	mAP@25	Avg R@10	Avg R@50
Pic2Word (2023)†	13.64	37.45	52.22	2.85	3.24	3.89	11.73	26.30
SEARLE (2023)†	23.71	53.30	66.84	8.90	9.42	10.64	22.64	41.85
Context-I2W (2024)	–	–	–	–	–	–	–	–
CIReVL (2024)	23.94	52.51	66.00	14.94	15.42	17.00	**28.29**	**49.35**
LinCIR (2024)†	18.87	45.66	58.43	6.25	6.74	7.62	18.14	36.20
PrediCIR (2025)	–	–	–	–	–	–	–	–
MoA (2025)	–	–	–	–	–	–	–	–
HIT (2025)	–	–	–	–	–	–	–	–
Pic2Word + RTD (2025)†	23.59	51.76	65.16	6.39	6.66	7.64	22.50	42.86
SEARLE + RTD (2025)†	26.29	56.41	69.74	11.26	12.11	13.63	24.70	45.40
LinCIR + RTD (2025)†	24.82	53.47	66.87	8.94	9.35	10.57	22.79	43.68
LinCIR + MoTa-Adapter (2026)	–	–	–	–	–	–	–	–
De-MINDS (2026)	–	–	–	–	–	–	–	–
FlowCIR (2026)	25.50	56.50	69.80	13.10	13.40	14.60	23.40	43.20
PACT (ours)	**30.09** ±0.27	**61.27** ±0.29	**74.33** ±0.17	**16.74** ±0.34	**17.42** ±0.15	**19.16** ±0.25	27.15 ±0.11	47.05 ±0.30

†Results reported in RTD, supplementary Tables S6–S7 (Byun et al., 2025), including its evaluations of the base methods. CIReVL and FlowCIR results are from their respective Tables 1–2.

Table 11: **CIRR/CIRCO test and FashionIQ validation, ViT-G/14.** PACT cells show the five-seed mean above ± sample standard deviation (%p). Bold marks the highest mean in each column; – indicates no available result for this backbone and metric.

Method	CIRR test			CIRCO test			FashionIQ val	
	R@1	R@5	R@10	mAP@5	mAP@10	mAP@25	Avg R@10	Avg R@50
Pic2Word (2023)‡	30.41	58.12	69.23	5.54	5.59	6.68	31.28	51.89
SEARLE (2023)‡	34.80	64.07	75.11	13.20	13.85	15.32	34.81	55.71
Context-I2W (2024)	–	–	–	–	–	–	–	–
CIReVL (2024)	34.65	64.29	75.06	26.77	27.59	29.96	32.19	52.36
LinCIR (2024)‡	35.25	64.72	76.05	19.71	21.01	23.13	45.11	65.69
PrediCIR (2025)	37.00	66.10	77.90	23.70	24.60	25.40	47.20	67.80
MoA (2025)	–	–	–	–	–	–	–	–
HIT (2025)	–	–	–	–	–	–	–	–
Pic2Word + RTD (2025)	–	–	–	–	–	–	–	–
SEARLE + RTD (2025)	–	–	–	–	–	–	–	–
LinCIR + RTD (2025)	36.31	67.47	78.31	21.08	22.29	24.46	46.21	67.26
LinCIR + MoTa-Adapter (2026)	38.39	69.47	80.05	25.82	27.06	–	47.06	67.37
De-MINDS (2026)	40.30	69.60	79.80	30.80	32.20	33.90	**49.50**	**70.50**
FlowCIR (2026)	–	–	–	–	–	–	–	–
PACT (ours)	**40.91** ±0.32	**71.63** ±0.28	**82.68** ±0.21	**31.48** ±0.59	**32.78** ±0.73	**34.92** ±0.75	46.43 ±0.33	66.45 ±0.39

‡Results reported in LinCIR, Tables 3, 5, and 6 (Gu et al., 2024). LinCIR + RTD uses RTD supplementary Tables S8–S9; LinCIR + MoTa-Adapter uses its Table 1. CIReVL, PrediCIR, and De-MINDS use their original reported results.

Among the compared methods, PACT achieves the highest CIRR R@1/5/10 and CIRCO mAP@5/10/25 with both backbones. FashionIQ shows a different pattern: CIReVL leads with ViT-B/32, while De-MINDS leads with ViT-G/14.

C.8 GeneCIS comparison across backbones

We report the available GeneCIS results at their published level of aggregation. We retain the baseline methods from Table 2 and leave unavailable metrics as –. Table 12 uses the four-task averages reported in RTD, supplementary Table S4 (Byun et al., 2025); task-level scores are not reported there. These six baseline rows retain RTD's evaluation setting and are not combined with SEARLE results from other evaluations. PACT's average assigns equal weight to Focus Attribute, Change Attribute, Focus Object, and Change Object and is computed across the five seeds.

Table 12: **GeneCIS four-task averages with ViT-B/32.** RTD-evaluated baselines and a separately sourced CIReVL row. PACT uses Chord scoring and five-seed means. Bold marks column maxima; – indicates no available result for this backbone and metric.

Method	Average R@1	Average R@2	Average R@3
Pic2Word (2023)	11.13	21.08	31.05
SEARLE (2023)	12.19	22.56	32.03
CIReVL (2024)‡	15.85	26.83	36.78
LinCIR (2024)	12.23	21.29	30.80
Pic2Word + RTD (2025)	12.03	21.61	31.09
SEARLE + RTD (2025)	12.82	22.97	32.44
LinCIR + RTD (2025)	12.83	22.83	32.22
MoTa-Adapter (2026)	–	–	–
DiffComp (2026)	–	–	–
PACT (ours)	**17.23**	**28.53**	**38.65**

‡Equal averages of the four ViT-B/32 task scores in CIReVL, Table 3 (Karthik et al., 2024); this row uses CIReVL's evaluation setting. Pic2Word, SEARLE, LinCIR, and their +RTD rows come from RTD Table S4.

Table 13 reports the four individual tasks for ViT-G/14. Pic2Word, SEARLE, and LinCIR results are reported in LinCIR, supplementary Table B.3 (Gu et al., 2024); CIReVL results are reported in its Table 3 (Karthik et al., 2024). The task order follows Table B.3: Focus Attribute, Change Attribute, Focus Object, and Change Object. These are individual tasks, rather than the overlapping marginals in Table 2. PACT entries are five-seed means with Chord scoring. PACT has the highest mean in each displayed task and recall rank among these comparison methods.

Table 13: **GeneCIS task-level results with ViT-G/14.** Published baselines and five-seed PACT means. Bold indicates the highest value in each column among the compared methods; – indicates no available task-level result.

Method	Focus Attribute			Change Attribute			Focus Object			Change Object		
	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3
Pic2Word (2023)	12.45	23.40	33.65	11.74	21.88	30.87	9.90	19.34	27.35	8.57	18.16	26.12
SEARLE (2023)	16.30	29.40	40.70	16.15	27.32	35.46	10.77	18.16	27.91	8.27	15.56	25.77
CIReVL (2024)	20.50	34.00	44.50	16.10	28.60	39.40	14.70	25.20	33.00	18.10	31.20	41.00
LinCIR (2024)	19.05	33.00	42.30	17.57	30.16	38.07	10.10	19.08	28.06	7.91	16.33	25.71
RTD (2025)	–	–	–	–	–	–	–	–	–	–	–	–
MoTa-Adapter (2026)	–	–	–	–	–	–	–	–	–	–	–	–
DiffComp (2026)	–	–	–	–	–	–	–	–	–	–	–	–
PACT (ours)	**22.32**	**34.30**	**45.65**	**19.11**	**32.68**	**42.11**	**20.22**	**29.12**	**37.77**	**19.83**	**32.72**	**42.06**

LinCIR + MoTa-Adapter reports a four-task R@1 summary of 19.1 (Li et al., 2026), but no task-level values; its task cells remain –.

D Additional analysis of Chord scoring

D.1 Retrieval implementation and per-dataset gains

Chord operates on unit-normalized query and gallery representations. Each model constructs its identity query with an empty modification. The reference-relative term is set to zero for a candidate identical to the reference image. Gallery embeddings remain fixed, and scoring uses no inverted-softmax calibration. The reference is excluded on CIRR, retained in local CIRCO evaluation, and FashionIQ candidates are restricted to the relevant category.

Table 14 expands the mean gains in Figure 3a into individual benchmark results. Pic2Word and LinCIR use released mappers; RTD is reproduced with its official implementation. Within each model, cosine and Chord use the same query representations and candidate embeddings.

Table 14: **Cosine and Chord retrieval on validation data.** Entries show cosine → Chord with one selected coefficient per model. PACT corresponds to the full-model rows in Table 3.

Model	λ	CIRR R@1	CIRCO mAP@5	FashionIQ R@10
Pic2Word	0.16	$22.82 \rightarrow 23.44$	$8.10 \rightarrow 9.02$	$25.56 \rightarrow 25.76$
LinCIR	0.34	$25.09 \rightarrow 27.36$	$10.65 \rightarrow 12.90$	$26.05 \rightarrow 27.30$
RTD	0.08	$27.12 \rightarrow 27.77$	$13.23 \rightarrow 14.03$	$29.93 \rightarrow 29.83$
PACT	0.34	$25.76 \rightarrow 32.50$	$18.68 \rightarrow 24.18$	$28.29 \rightarrow 32.43$

PACT improves on all three benchmarks, by 6.74 points on CIRR, 5.50 on CIRCO, and 4.14 on FashionIQ. Chord also increases the aggregate performance of the other query models, although gains need not occur on every dataset: RTD improves on CIRR and CIRCO and decreases by 0.10 points on FashionIQ. These results reflect validation-selected scoring coefficients.

D.2 Cosine ambiguity and pair correction

For each query, we compare every annotated positive with the highest-cosine candidate that is not annotated as positive. We retain pairs for which

$$\delta = \cos(q, g_n) - \cos(q, g_p) > 10^{-7}.$$

Each query contributes equal total pair weight within its category, and FashionIQ categories receive equal weight. Within each benchmark, we pool the four models' cosine-deficit distributions with equal weight per model and form five weighted quantile bins. The boundaries are shared across models. Q1 contains the largest deficits and Q5 the smallest. These bins measure cosine-ranking ambiguity. On CIRCO, candidates outside the annotated positive set are unannotated, rather than verified irrelevant.

Within each quintile, we stratify the positive's cosine rank into $[2, 6)$, $[6, 11)$, $[11, 21)$, $[21, 51)$, $[51, 101)$, $[101, 501)$, and $[501, \infty)$. A rank cell is retained only if it contains at least ten distinct queries for every model. We use common weights proportional to the minimum query-weighted cell mass across models to standardize the rank distribution. This controls the rank composition of the comparison; the models can still contribute different queries and candidates.

We measure the fraction of misordered pairs corrected by Chord at a common $\lambda = 0.34$, and average the three benchmark rates equally for Figure 3b. The 95% intervals use 1,000 bootstrap replicates clustered by reference image, with bin boundaries, support, and standardization weights held fixed. They quantify query-sampling variation rather than variation across training seeds. Figure 6 reports the individual benchmark curves.

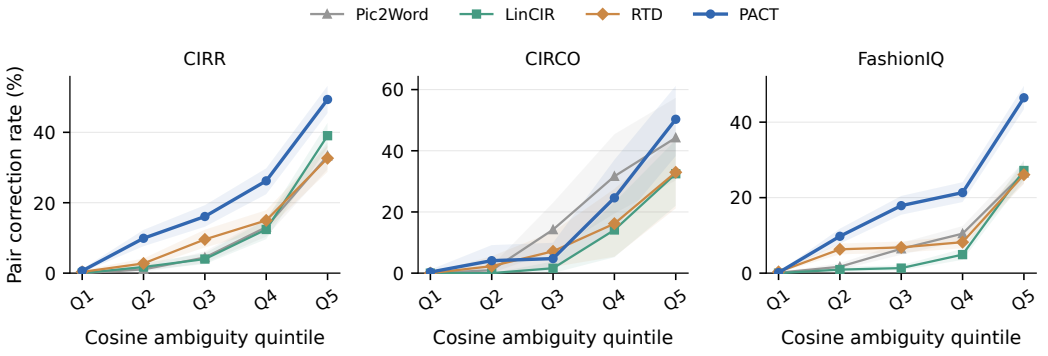

Figure 6: **Chord pair correction by benchmark.** Shared cosine-deficit quintiles and rank standardization follow the procedure described above. All models use $\lambda = 0.34$. Q5 contains the smallest deficits; shading denotes 95% reference-cluster bootstrap intervals.

E Reference preservation on MovieNet

E.1 Evaluation sets

We use MovieNet (Huang et al., 2020) for two complementary evaluations: an annotation-based identity-preservation evaluation and qualitative retrieval over a larger gallery. Table 15 summarizes their respective query and gallery sets. This section describes the quantitative evaluation in Table 4; the qualitative comparison is described in Appendix F.

Table 15: **MovieNet evaluation sets.** Query-film counts refer to the films represented by the references; gallery-film counts refer to the retrieval collections.

	Identity-preservation evaluation	Qualitative comparison
Queries	200	1,151
Distinct reference images	200	90
Films represented in queries	62	4
Gallery images	6,162	100,000
Films represented in gallery	66	104

E.2 Gallery and query construction

From the 100,000-image MovieNet collection, we retain frames with official person annotations for which every annotated person has a single, unambiguous identity assignment. Frames with unresolved or multiple candidate identities are excluded. The resulting gallery contains 6,162 frames from 66 films and includes images with multiple people.

To construct queries, we form source–target pairs from the same film and with the same person identity. Each frame in a pair has one annotated person, the files have different contents, and the shot indices differ by at least 40. Qwen3.5-27B receives both images and generates an instruction that preserves the person while changing the background, scene, or action. Automatic verification checks that the target satisfies the instruction and the source does not already satisfy it. Reference images undergo additional checks for frontal or three-quarter face orientation, clarity, illumination, and occlusion. The final set contains 200 queries with 200 distinct references from 62 films.

At retrieval, we exclude images with file contents identical to the reference and images from the same shot in the same film. We retain Top-5 predictions and evaluate the Top-1 result. Source–target pairs are used to construct evaluation queries and do not provide training supervision for PACT.

E.3 Identity and modification criteria

For query i, let $\hat{G}_i$ be the top-ranked image. Define $A_i = 1$ when the official MovieNet annotations for $\hat{G}_i$ contain the reference person's film-ID/person-ID pair, and $A_i = 0$ otherwise. The identity preservation rate is

$$\text{Identity} = \frac{100}{N} \sum_{i=1}^{N} A_i, \qquad N = 200. \tag{17}$$

Identity is determined from official annotations, without face recognition or language-model identity judgments. A retrieved image containing several people is counted as preserving identity when the reference person's annotation is present. This criterion measures person-identity retention; it does not require preservation of every visual attribute, such as clothing, color, or texture.

Modification fulfillment is assessed separately by Qwen3.5-27B, which receives the reference image, retrieved image, and instruction with the retrieval model's identity hidden. Let M_i be its binary judgment of whether the retrieved image fulfills the modification. We define

$$\text{Modification} = \frac{100}{N} \sum_{i=1}^{N} M_i, \qquad \text{Joint} = \frac{100}{N} \sum_{i=1}^{N} A_i M_i. \tag{18}$$

The joint criterion requires both identity retention and modification fulfillment. Table 4 reports the identity metric.

E.4 Effect of visual supervision

Adding the visual objective to composition-only training increases Top-1 identity from 58.0% to 78.5% with cosine and from 62.0% to 81.5% with Chord, gains of 20.5 and 19.5 percentage points, respectively. Full PACT reaches 80.5% and 82.5%, exceeding RTD by 16.5 and 18.0 points under the corresponding scoring rules. The improvement under both rules supports the role of visual supervision in retaining source-specific identity.

F Qualitative retrieval on MovieNet

F.1 Gallery construction

The qualitative evaluation uses 100,000 images from 104 films. The gallery comprises 9,660 images from the four query films and 90,340 images from 100 additional films. The additional films are selected in a fixed hash order from films with at least 1,000 shots, excluding the query films. Shots are sampled evenly within each selected film, and the middle frame of each sampled shot is used. Unlike the annotation-based collection in Appendix E, this gallery is not restricted to frames with unambiguous person-identity annotations.

F.2 Reference images and modifications

We use 90 reference images from *The Matrix* (1999), *Pirates of the Caribbean: The Curse of the Black Pearl* (2003), *The Dark Knight* (2008), and *Inception* (2010). Descriptions of visual states are written after inspecting clothing, pose, action, and background. Modifications request a different state of the same person, using combinations of these conditions. Additional instructions isolate subsets of the described conditions, such as an action or a background. The resulting set contains 1,151 queries, with multiple modifications associated with the same reference image.

F.3 BASELINE CONFIGURATIONS

The qualitative comparisons use the same 100,000-image MovieNet gallery and ViT-L/14 backbone. Image-only and Instruction-only use the original, frozen CLIP encoders; neither passes through the learned PACT mapper or adapted text encoder. Compose-only is a training-objective ablation of PACT and retains both query modalities. Let $v(I)$ and $t(m)$ denote the L_2-normalized embeddings from the original CLIP image and text encoders, respectively, and let I_g be a gallery image.

Image-only. The reference image is encoded by the original CLIP image encoder, and its normalized embedding is used directly as the retrieval query:

$$s_{\text{image-only}}(I_g) = v(I_{\text{ref}})^\top v(I_g). \tag{19}$$

This cosine baseline receives no modification instruction and uses neither the PACT mapper nor its adapted text encoder. It measures retrieval from source appearance alone.

Instruction-only. The complete, unmodified instruction m is encoded by the original frozen CLIP text encoder and normalized before cosine retrieval:

$$s_{\text{instruction-only}}(I_g) = t(m)^\top v(I_g). \tag{20}$$

The input includes reference-dependent phrases such as "Keep the same character, but ..." exactly as written. No reference image, source-image caption, or image-dependent pseudo-word is supplied. The learned PACT text encoder is not used. Both single-modality baselines use the same gallery image embeddings and exclude the reference image and its byte-identical duplicates from retrieval.

Compose-only. This model uses the same training data, architecture, trainable modules, and optimization schedule as full PACT (Appendix B). It retains the 333-update identity warmup and 1,000-update joint-training stage. The mapper and text-attention LoRA follow the same training procedure; the image encoder and teacher text encoder remain frozen. Only the joint objective changes: it is $\mathcal{L}_{\text{compose}}$ from Equation 3, with the visual and negative-caption preservation terms disabled. Thus, the model still receives identity-warmup supervision and both query modalities. At retrieval time, Compose-only uses Chord with $\lambda = 0.34$, the same scoring rule and coefficient as full PACT. Their comparison therefore holds the query inputs and retrieval scoring fixed while varying the joint-training objective.

Full PACT and RTD. Full PACT uses both the reference image and modification instruction, the complete training objective, and Chord with $\lambda = 0.34$. RTD uses the official-code reproduction checkpoint and cosine scoring. All displayed model results are the original Top-1 retrievals for the corresponding query and checkpoint.

F.4 COMPARISON AND INTERPRETATION

Figures 7–10 show eight selected comparisons. Checks mark matched requests; crosses mark unmet character or modification conditions. Blue identifies PACT independently of correctness. A uniform 20% brightness increase is applied only to the displayed images for visibility, not to retrieval inputs. These selected examples do not estimate overall success rates.



Figure 7: MovieNet Top-1 comparisons S1–S2. S1: composition-only retrieves a different character in a pale outfit. S2: it matches the beside-a-woman relation but loses the reference character. PACT satisfies both requests. "Compose-only" denotes $\mathcal{L}_{\text{compose}}$-only training.

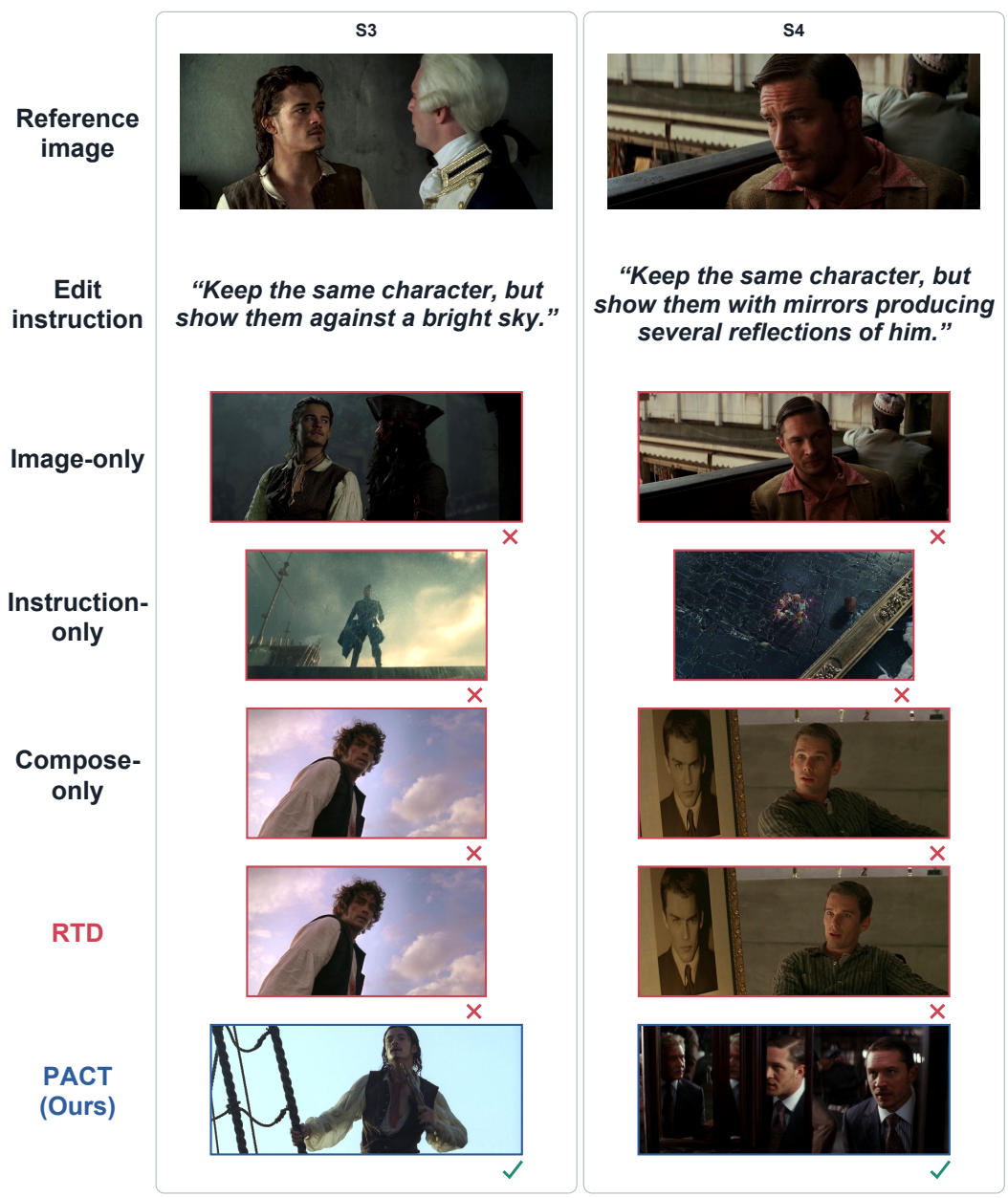

Figure 8: MovieNet Top-1 comparisons S3–S4. S3: composition-only matches the bright sky with a different character. S4: it retrieves a portrait instead of mirror reflections. PACT satisfies both requests. "Compose-only" denotes $\mathcal{L}_{\text{compose}}$-only training.



Figure 9: MovieNet Top-1 comparisons S5–S6. S5: PACT retrieves the requested pale pillow behind the head. S6: both composition-only and PACT retrieve the reference character beside a damaged car. "Compose-only" denotes $\mathcal{L}_{\text{compose}}$-only training.



Figure 10: MovieNet Top-1 comparisons S7–S8. Both composition-only and PACT satisfy the ornate-room (S7) and work-table (S8) requests. "Compose-only" denotes $\mathcal{L}_{\text{compose}}$-only training.

F.5 FAILURE CASES: FINE-GRAINED MODIFICATION CONSTRAINTS

Figure 11 illustrates unmet pose and action constraints. Both models use ViT-L/14: PACT with Chord ($\lambda = 0.34$), RTD with cosine. These selected examples do not estimate failure frequency.

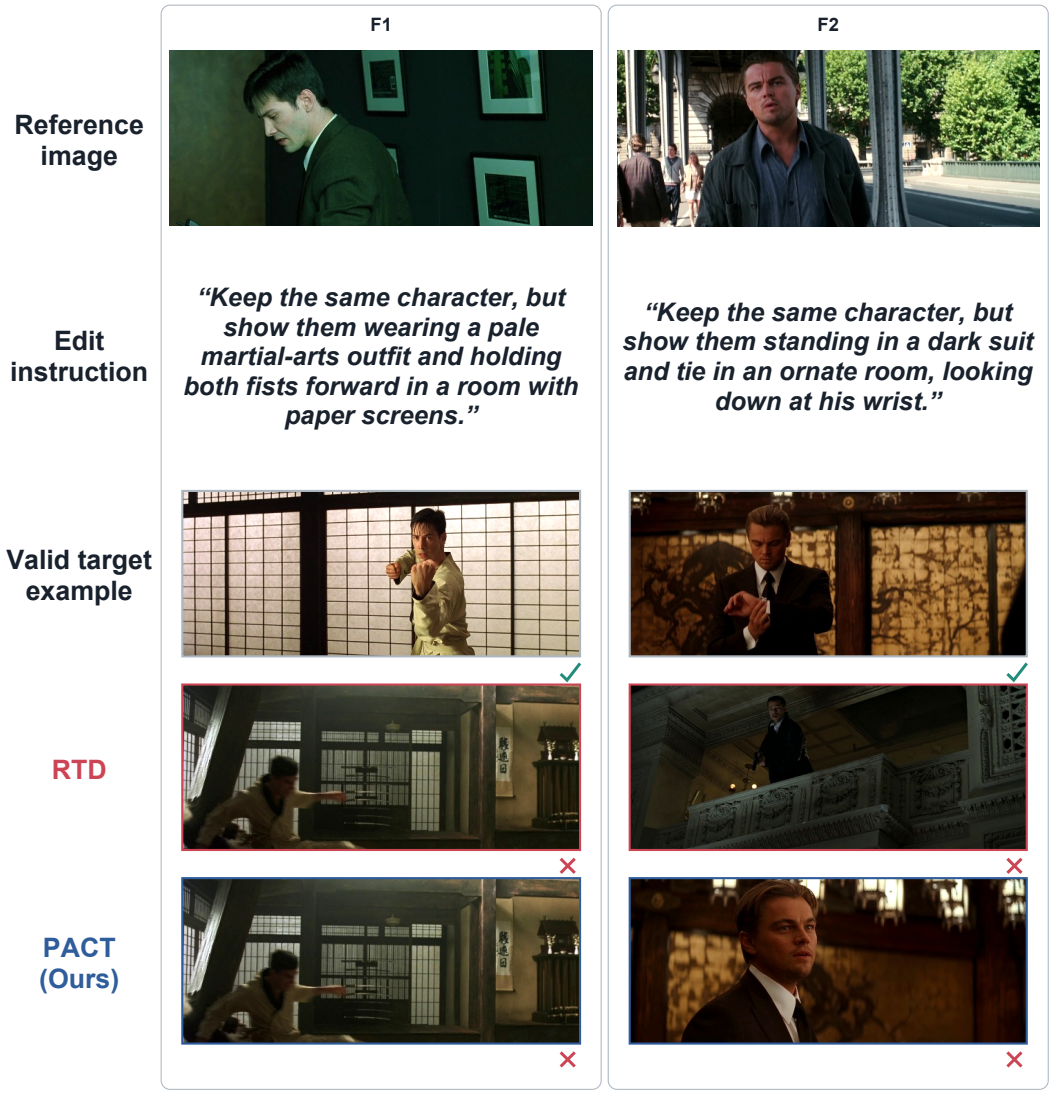

Figure 11: MovieNet failures F1 and F2. In F1, both methods retrieve the same martial-arts frame, but the requested two-fist pose is not visible. In F2, both recover suit and interior cues without the action of looking down at a wrist. Valid targets are illustrative gallery positives, not unique correct frames. All model results are Top-1.

G EFFICIENCY AND MODEL-SIZE ACCOUNTING

G.1 TRAINING-TIME MEASUREMENT

We measure the ViT-L/14 training configuration with batch size 512 on a single RTX 3090. CUDA-synchronized updates for 333 identity-warmup and 1,000 joint-training steps take 52 min 9 s. Peak allocated GPU memory is 8.19 GiB. The timing excludes evaluation, checkpoint saving, offline data generation, and image-feature preparation. These measurements characterize optimization with cached supervision and frozen image features.

G.2 INFERENCE PARAMETER ACCOUNTING

Parameter counts include the image and text encoders and query-side modules used for retrieval; offline generation models and training-only teachers are excluded. PACT-L/14 has approximately 443M inference parameters. Figure 4 uses CIRCO test mAP@5 as the common performance measure. Architecture-derived estimates and component lower bounds are distinguished below because they have different counting scopes.

For MoTa-Adapter (Li et al., 2026), we add the adapter components to CLIP and the LinCIR mapper. The added terms are $KNd_T + Kd_T + 2d_I$, with $K = 4$ and $N = 8$, giving estimated totals of 441.81M for L/14 and 2.57895B for G/14. For MCL (Li et al., 2024), the totals include the causal language model, the CLIP-L/14 visual encoder, visual and retrieval projections, and the fusion module. Tied weights are counted once. The OPT-2.7B, OPT-6.7B, and Llama2-7B variants yield architecture-derived totals of 2.975B, 6.988B, and 7.068B, respectively.

CoLLM (Huynh et al., 2025) uses a nominal 7B language model; adding its visual encoder gives a component lower bound because adapter and projection overhead is not fully specified. Its plotted scores use image-caption pretraining and exclude MTCIR fine-tuning. For PrediCIR (Tang et al., 2025), the lower bound includes CLIP and the attention projection weights of 12 predictor blocks of width 384: $12(4 \times 384^2) = 7{,}077{,}888$ additional parameters. Feed-forward layers and other unspecified components are excluded, yielding lower bounds of 434.69M for L/14 and 2.54664B for G/14.

The parameter count of the GPT-3.5 component in CIReVL (Karthik et al., 2024) is undisclosed. Its horizontal position in Figure 4 uses an illustrative 175B assumption for that component, together with the captioner and CLIP. This position is not a measured model size and cannot establish a quantitative parameter ratio relative to PACT.