Duplex Rewards Optimization for Test-Time Composed Image Retrieval

Haoliang Zhou[1], **Feifei Zhang**[1*], **Changsheng Xu**[2, 3, 4]

[1]Tianjin University of Technology
[2]Institute of Automation, Chinese Academy of Sciences
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences
[4]Peng Cheng Laboratory
hlzhou@stud.tjut.edu.cn, feifeizhang@email.tjut.edu.cn, csxu@nlpr.ia.ac.cn

Abstract

Composed Image Retrieval (CIR) combines the reference image with text to retrieve the intended target image. Recently, zero-shot CIR has gained significant attention by eliminating the need for labeled triplets required in supervised CIR. However, it inevitably demands additional training corpus, storage, and computational resources, limiting its applicability in real-world scenarios. Inspired by advancements in Test-Time Adaptation (TTA), we propose a Test-Time CIR setting named TT-CIR, which aims to efficiently adapt models to unlabeled test samples while reducing resource consumption. Within the TT-CIR setting, we identify that naively introducing existing TTA methods (*e.g.*, reward-based) into CIR faces two vital challenges: 1) Modification-restricted reward pool, which limits the exploration of semantically relevant candidate rewards; 2) Conservative knowledge feedback, which inhibits the adaptability of rewards to the current data distribution. To address these challenges, we propose a test-time reinforcement learning framework that integrates a Counterfactual-guided Multinomial Sampling (CMS) strategy and a Duplex Rewards Modeling (DRM) module. The CMS explores a candidate reward pool that is visually similar and semantically relevant to the given query, while the DRM generates stable and adaptive duplex rewards to guide model adaptation. Extensive experiments demonstrate the superiority and adaptability of our method over existing approaches.

Introduction

Composed Image Retrieval (CIR) aims to retrieve target images that are visually similar to a reference image while meeting modification requirements specified in a textual description. Unlike traditional image or text-based retrieval methods (Zhang et al. 2024b) that rely solely on single-modal queries, CIR integrates composed vision-language queries as input to express user intent more accurately. Consequently, CIR enables user-specific and more precise image retrieval, garnering increasing attention in areas such as internet search and e-commerce (Li et al. 2025b).

While supervised CIR (Chen et al. 2023; Feng et al. 2025) has achieved remarkable progress, it relies on extensive well-annotated triplets (*i.e.*, reference image, modification text, target image) to train task-specific models, which is labor-intensive. To address this issue, recent research proposes zero-shot CIR (ZS-CIR) based on pre-trained vision-language models (VLMs) (Radford et al. 2021), which can be divided into two categories: Textual inversion methods (*e.g.*, Pic2Word (Saito et al. 2023)) train a mapping module using image-caption corpus, representing the reference image to pseudo-tokens and then concatenating with the modification text for retrieval. However, as depicted in Fig. 1(b), they still require large amounts of data for additional training, making the process time-consuming. Alternatively, Large Language Model (LLM)-based methods, such as CIReVL (Karthik et al. 2024), first employ a captioning model to convert the reference image into a textual description, followed by the LLM to derive a target description for retrieval. Despite the commendable progress, these methods depend on complex collaboration among multiple expert models, which require substantial storage and computational resources, as shown in Fig. 1(b). Therefore, it is urgent to develop a more efficient adaptation manner for CIR systems to achieve effective and precise retrieval.

In recent years, Test-Time Adaptation (TTA) (Wang et al.

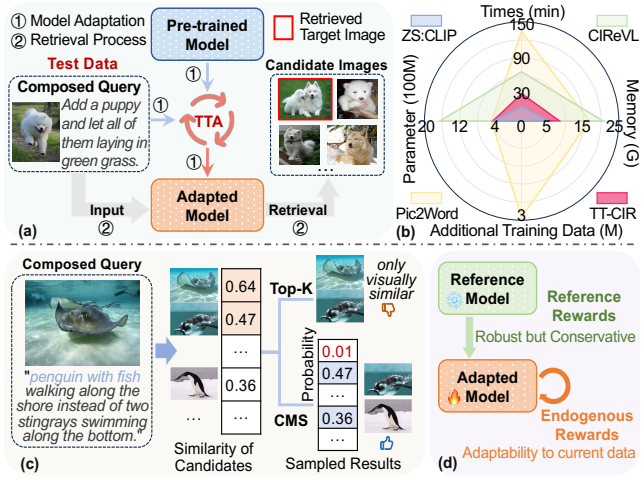

Figure 1: (a) The TT-CIR setting; (b) Computational overhead between ZS-CIR methods and our TT-CIR on CIRR dataset; (c) Modification-restricted of reward pool, and (d) Conservative knowledge feedback in reward-based methods.

*Corresponding author.

2021) has gained significant attention for its ability to effectively adapt pre-trained models to unlabeled data before making predictions during testing. TTA recalibrates models on-the-fly using only the current data point with minimal overhead, making it particularly suitable for real-world applications (Shu et al. 2022; Lee et al. 2024; Wen et al. 2023; Li et al. 2025a). Inspired by the success of TTA approaches, as illustrated in Fig. 1(a), we explore a novel setting, Test-Time Composed Image Retrieval (**TT-CIR**) — aiming to efficiently fine-tune CIR models using only unlabeled test data during testing. As shown in Fig. 1(b), compared to the above ZS-CIR methods, TT-CIR allows the model to achieve improved retrieval results with lower resource requirements and reduced computational overhead.

Mainstream TTA techniques seek to optimize normalization layers (Wang et al. 2021) or learnable prompt tokens (Shu et al. 2022; Liu et al. 2024) with unsupervised objectives, such as entropy minimization (EM). While EM enhances model predictions on unseen data by reducing the corresponding entropy, it tends to make the model blindly confident and unable to self-correct when predictions are incorrect (Zhao et al. 2024). To mitigate overconfidence, RLCF (Zhao et al. 2024) introduces a reward-based TTA framework for VLMs, employing a powerful reference model, *e.g.*, CLIP (Radford et al. 2021), to generate reference rewards that guide model tuning. In this way, model predictions are enhanced by adjustments from stable external rewards, rather than working in isolation. Although achieving commendable progress, naively adopting this approach for TT-CIR remains challenging due to two critical obstacles: **(1) Modification-restricted of reward pool** — As a prerequisite for stable reward estimation, constructing an appropriate candidate reward pool is indispensable. To achieve this, RLCF samples the top-k targets that are most semantically similar to the query image. However, this approach tends to select target images that are visually similar to the reference image due to homogeneity, while ignoring the instruction of modification text, as shown in Fig. 1(c). Consequently, this results in a semantically irrelevant reward pool, hindering the effectiveness of subsequent reward estimation. **(2) Conservative knowledge feedback** — While the reference model provides stable reward signals, it also builds an insurmountable knowledge barrier. As noted in research (Zhou et al. 2024), the reference model prioritizes the inherent knowledge present in the training data while ignoring the actual input information. Specifically, the frozen reference model generates inflexible and conservative rewards for feedback, making it struggle to adapt to the current input data, as depicted in Fig. 1(d). Moreover, inherent biases within the reference model may be transferred to the adapted model, potentially hindering its adaptability during test time.

To address the above challenges, we propose a Test-Time Reinforcement Learning with Duplex Rewards (**TT-RLDR**) approach for TT-CIR, which aims to improve the adaptability of pre-trained VLMs to unseen queries during testing. Specifically, TT-RLDR comprises two main components: Initially, a **Counterfactual-guided Multinomial Sampling (CMS)** strategy is designed to construct the reward pool by mining candidate target images that are both visually similar and semantically related to the composed query. In CMS, the counterfactual-guided consistency constraint is deployed to adjust the sampling probabilities, exploring a more accurate set of candidate rewards. Additionally, we propose a **Duplex Rewards Modeling (DRM)** module that generates both reference rewards (RR) and endogenous rewards (ER) to guide model optimization. Among them, RR serves as a high-quality and stable reward signal, generated by a powerful reference model to prevent the model from deviating from the correct optimization. Meanwhile, ER reflects the adapted model's confidence in current data distribution, providing task-specific rectifications to RR.

The major contributions of this work are as follows: (1) We introduce TT-CIR, a novel test-time CIR setting that aims to efficiently adapt pre-trained VLMs to unlabeled data during inference. We also propose a test-time reinforcement learning framework to prevent the model from becoming blindly confident during optimization. (2) A Counterfactual-guided Multinomial Sampling (CMS) strategy is designed to precisely explore the candidate reward pool. Besides, we propose a Duplex Rewards Modeling (DRM) module to generate rewards that balance stability and adaptability. (3) We perform extensive experiments to evaluate our method against several well-established ZS-CIR and TTA methods, demonstrating the effectiveness of our approach.

Related Work

Composed Image Retrieval

Composed image retrieval (CIR) enables users to search images according to the given multimodal queries (Bai et al. 2024). Existing prevalent CIR methods employ VLMs (Radford et al. 2021; Li et al. 2022) as foundational encoders for task-specific training (Baldrati et al. 2022; Chen et al. 2024; Levy et al. 2024). However, their superior performance heavily relies on extensive well-annotated triplets, which are labor-intensive to collect. To mitigate the need of training dataset annotation, zero-shot CIR has recently gained significant attention. Currently, two prominent directions exist: Textual inversion methods (Tang et al. 2024; Suo et al. 2024), *e.g.*, Pic2Word (Saito et al. 2023), train a mapping network using only image-caption corpus, transforming the reference image into a pseudo-token and then combining it with modification. However, it still requires additional data for training and is time-consuming. Another line of methods, such as CIReVL (Karthik et al. 2024), adopts captioning models to convert the reference image into a textual description, and then edits this description according to the modification text by a large language model. Despite their success, cascading multiple off-the-shelf experts leads to component incompatibility and high resource consumption, limiting its practical application. To address this, we propose a test-time CIR paradigm that promptly adapts pre-trained VLMs to current test samples, achieving precise retrieval with reduced resource and computational overhead.

Test-Time Adaptation

Test-time adaptation (TTA) has proven effective in adapting trained models to unseen out-of-distribution samples

during test time, particularly with potential distribution shifts (Liang, He, and Tan 2025; Lee et al. 2024). To achieve this goal, early researches focus on *Normalization Calibration* (NC), which updates the parameters of the normalization layer using batched test data (Wang et al. 2021; Mirza et al. 2022) or augmented views from a single test sample (Zhang, Levine, and Finn 2022). Subsequently, researchers utilize *Entropy Minimization* (EM) to increase the model's confidence in its predictions, thereby reducing generalization error (Shu et al. 2022; Niu et al. 2023). Among them, CMF (Lee and Chang 2024) leverages the Kalman filter to strike a balance between model adaptability and information retention. Recent studies involve *Consistency Regularization* (CR), which employs a robust teacher model to ensure consistency between teacher and student predictions (Döbler, Marsden, and Yang 2023) or uses the pseudo-label generated by the teacher (Wang et al. 2022). However, the above approaches tend to become blindly confident in incorrect predictions and are unable to resolve this dilemma independently. In contrast, we introduce a reinforcement learning framework with duplex rewards, which provides feedback to the model being optimized and prevents it from getting stuck in a self-optimization loop.

Reinforcement Learning with Rewards

Reinforcement learning (RL) has emerged as a promising approach in improving large language models (LLMs) post-training — particularly demonstrated by the success of RL from human feedback (RLHF) that aligns model responses with human preferences (Ouyang et al. 2022; Rafailov et al. 2023). Among them, RL with verifiable rewards has garnered attention for enhancing the reasoning capabilities of LLMs through customized rewards (Shao et al. 2024; Wang et al. 2025). RL has also been widely applied in the multimodal community. For example, CaptionReward (Cho et al. 2022) investigates CLIPScore (Hessel et al. 2021) as a reward function for image captioning, empirically demonstrating that CLIPScore is an effective reward function. More recently, RLCF (Zhao et al. 2024) employs CLIPScore as a reward function to provide feedback to the adapted model during testing. It utilizes top-k sampling with average baseline strategies to convert non-negative rewards into advantages, further encouraging or penalizing the model's decision-making behavior. In CIR, however, the trivial top-k sampling tends to select negative samples that are visually similar but semantically irrelevant. Besides, the generated rewards rely on the frozen reference model, limiting its ability to adapt to current data. In contrast, we propose a counterfactual-based multinomial sampling strategy to accurately identify candidate reward samples, while introducing endogenous rewards to provide task-specific feedback corrections.

Proposed Method

The overall architecture of our proposed method is illustrated in Fig. 2. This framework employs a pre-trained VLM with bimodal encoders to process the reference image and modification text, as well as all candidate images, followed by a fusion module to combine the composed query. We then design a test-time reinforcement learning framework with duplex rewards to effectively adapt the VLM to unlabeled test queries, which comprises two pivotal components: the Counterfactual-guided Multinomial Sampling (CMS) strategy and the Duplex Rewards Modeling (DRM) module. Specifically, the CMS aims to explore a reward pool with candidates that are visually similar and semantically relevant to the given query. Building upon this, the DRM derives reference rewards from a frozen reference model for stable optimization, while generating endogenous rewards using the adapted model to provide task-specific rectifications.

Preliminary

Problem Formulation. Let $\mathcal{F}(\cdot;\theta)$ denote the pre-trained VLM with parameter θ, such as CLIP (Radford et al. 2021), which consists bimodal encoders, *i.e.*, an image encoder $\mathcal{F}_{\mathrm{V}}$ and a text encoder $\mathcal{F}_{\mathrm{T}}$. The objective of TT-CIR is to adapt $\mathcal{F}_{\mathrm{V}}$ and $\mathcal{F}_{\mathrm{T}}$ on downstream unseen test samples $\mathcal{D}_{\text{query}}=\{q_i^{\mathrm{I}}, q_i^{\mathrm{T}}\}_{i=1}^{N_q}$, where q^{I} and q^{T} represent the reference image and modification text, respectively. Then the adapted model aims to identify target images from the candidate set $\mathcal{D}_{\text{target}}=\{u_i\}_{i=1}^{N_t}$ consisting of N_t images, which are visually similar to q^{I} while meeting the modifications specified in q^{T}. Following TTA methods in image classification (Shu et al. 2022; Zhang et al. 2024a), adaptation is performed with a single test point (one composed query at a time), mimicking real-world online retrieval behavior.

Contrastive Language-Image Pre-training (CLIP). As a classical VLM, CLIP employs an image encoder $\mathcal{F}_{\mathrm{V}}$ and a text encoder $\mathcal{F}_{\mathrm{T}}$ to produce the visual embedding $\mathcal{F}_{\mathrm{V}}(x)$ and textual embedding $\mathcal{F}_{\mathrm{T}}(t)$ for the given image x and text t, respectively. Then, the contrastive loss (Chen et al. 2020) is utilized to encourage similarity between embeddings of the image-text pair, aligning them in a shared representation space. Once pre-trained, the matching score between the image and text can be measured using cosine similarity as:

$$\mathrm{Sim}(x,t)=\frac{(\mathcal{F}_{\mathrm{V}}(x))^{\top}\mathcal{F}_{\mathrm{T}}(t)}{\|\mathcal{F}_{\mathrm{V}}(x)\|\,\|\mathcal{F}_{\mathrm{T}}(t)\|}. \tag{1}$$

Feature Extraction and Composed Query Fusion

Given a test composed query $\{q^{\mathrm{I}}, q^{\mathrm{T}}\}\in\mathcal{D}_{\text{query}}$, we obtain visual and textual embeddings by $\mathrm{Q}^{\mathrm{I}}=\mathcal{F}_{\mathrm{V}}(q^{\mathrm{I}})$ and $\mathrm{Q}^{\mathrm{T}}=\mathcal{F}_{\mathrm{T}}(q^{\mathrm{T}})$. Meanwhile, the embeddings of candidate target images of $\mathcal{D}_{\text{target}}$ can be extracted by $\mathrm{U}_i=\mathcal{F}_{\mathrm{V}}(u_i)$ (i=1, ..., N^t). Then we need to combine the visual and textual embeddings into the composed embedding. In supervised CIR, researchers typically fuse visual and textual embeddings by training MLP combiners or cross-attention layers with abundant annotated triples. During test time, however, the absence of ground-truth supervision makes learning an effective fusion module always unreliable. Thus, an alternative fusion manner that employs interpolation between visual and textual embeddings can be considered. Specifically, the composed embedding is obtained by normalized linear interpolation as:

$$\mathrm{Q}=\alpha\cdot\frac{\mathrm{Q}^{\mathrm{T}}}{\|\mathrm{Q}^{\mathrm{T}}\|}+(1-\alpha)\cdot\frac{\mathrm{Q}^{\mathrm{I}}}{\|\mathrm{Q}^{\mathrm{I}}\|}, \tag{2}$$

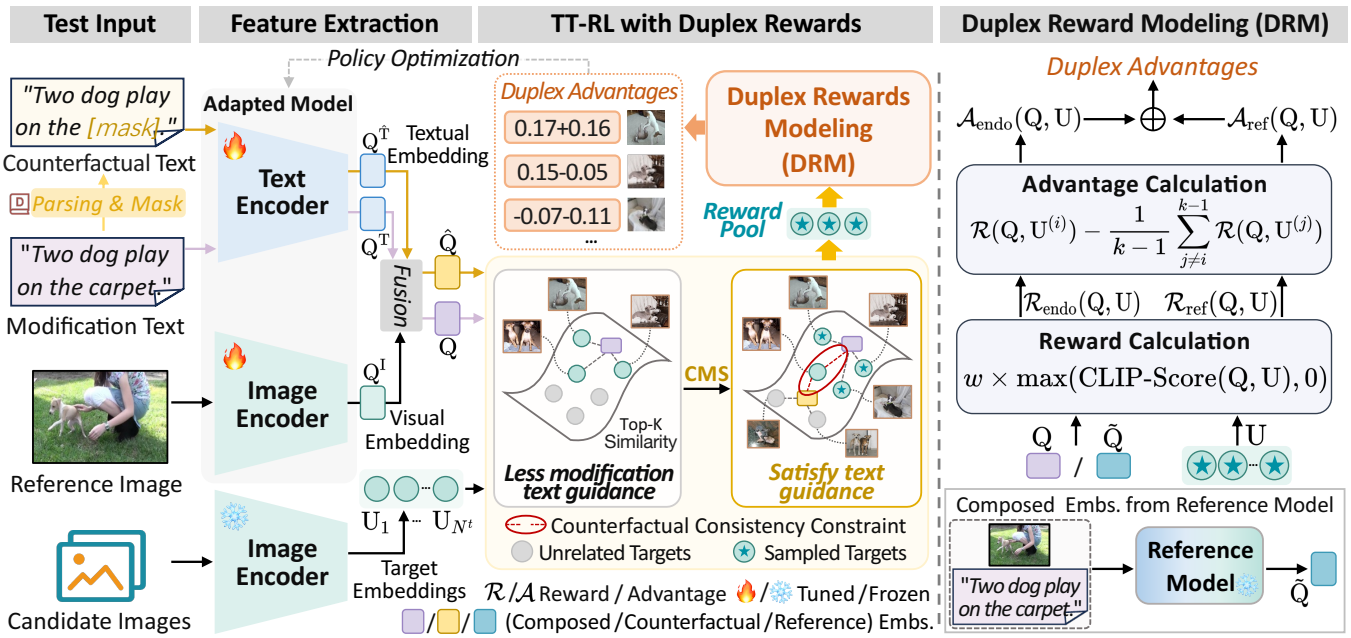

Figure 2: The overall architecture of our proposed TT-RLDR, which comprises two key components: a Counterfactual-guided Multinomial Sampling (CMS) strategy and a Duplex Rewards Modeling (DRM) module. "Embs" stands for "Embeddings".

where α is the interpolation weight that determines the contribution of visual and textual embeddings.

TT-RL with Duplex Rewards

Test-Time Reinforcement Learning with Rewards. Reinforcement learning (RL) has recently emerged as a powerful paradigm for enabling models to self-improve on tasks with objective outcomes and verifiable rewards. Previous reward-based optimization methods have used labeled preference data to train a reward model (Ouyang et al. 2022) or applied labels to measure the reward (Cho et al. 2022). Since ground truth is inaccessible during test time, recent research employs off-the-shelf models to generate rewards (Zhao et al. 2024) or applies majority voting to aggregate rewards (Zuo et al. 2025), showing promising progress in vision-language tasks. Inspired by this, we integrate RL with rewards to improve the CIR model's capability to swiftly adapt the unlabeled test data streams, a process we term TT-RLR. Specifically, during test time, our goal is to fine-tune the CIR model $\mathcal{F}(\mathrm{u}|\{q^{\mathrm{I}}, q^{\mathrm{T}}\};\theta)$ by learning the parameterized policy $\pi_\theta(\mathrm{U}|\{\mathrm{Q}^{\mathrm{I}}, \mathrm{Q}^{\mathrm{T}}\})$ to maximize the following objective:

$$\max_{\theta} \ \mathbb{E}_{\mathrm{Q}\sim\mathcal{D},\mathrm{U}\sim\pi_\theta(\cdot|\mathrm{Q})}[\mathcal{R}(\mathrm{Q},\mathrm{U})], \tag{3}$$

where the reward function $\mathcal{R}(\mathrm{Q},\mathrm{U})$ is used to assess whether the model's candidate retrieval U for the query Q is correct.

Since the discrete nature of query-target matching, the above objective is not differentiable. In RL, log-derivative trick and the subsequent REINFORCE estimator (Williams 1992; Ahmadian et al. 2024; Zhao et al. 2024) are typical applied to calculate gradient of the non-differentiable reward function for a given input query:

$$\nabla_\theta \mathbb{E}[\mathcal{R}(\mathrm{Q},\mathrm{U})] = \mathbb{E}_{\mathrm{U}\sim\pi_\theta(\cdot|\mathrm{Q})}\left[\mathcal{R}(\mathrm{Q},\mathrm{U}) \cdot \nabla_\theta \log \pi_\theta(\mathrm{U}|\mathrm{Q})\right]. \tag{4}$$

Counterfactual-guided Multinomial Sampling. In CIR, the input (composed query Q) and output (candidate target images U) are semantically related in the representation space. Therefore, we can use CLIP to evaluate the similarity between the input and output, since the ground truth target image is inaccessible during test time. Then the objective can be optimized by maximizing the similarity between these query-target pairs. Specifically, we apply CLIP-Score (Hessel et al. 2021) as the reward function:

$$\mathcal{R}(\mathrm{Q},\mathrm{U}) = w \times \max(\mathrm{Sim}(\mathrm{Q},\mathrm{U}), 0), \tag{5}$$

where, w=2.5 is a constant, and $\mathrm{Sim}(\cdot,\cdot)$ denotes the cosine similarity as described in Eq.1.

To eliminate irrelevant samples and enhance computational efficiency, reward calculations are not conducted over the entire target gallery. To achieve this, we design a Counterfactual-guided Multinomial Sampling (CMS) strategy to explore a subset of representative candidate targets for structuring the reward pool. Specifically, as shown in Fig. 2, we first calculate the fact matching score based on the similarity for each query, *i.e.*, $S_{\mathrm{fact}}(\{\mathrm{Q}^{\mathrm{I}}, \mathrm{Q}^{\mathrm{T}}\}, \mathrm{U})$=$\mathrm{Sim}(\mathrm{Q},\mathrm{U})$. Subsequently, for each original modification text Q^{T}, objects (nouns), relations (verbs, adverb), and attributes (adjectives) are identified using the NLP parsing technique. We then randomly choose several words from these concepts and replace them with the [MASK] token to form a counterfactual caption $\mathrm{Q}^{\hat{\mathrm{T}}}$. After that, we fuse Q^{I} and $\mathrm{Q}^{\hat{\mathrm{T}}}$ to obtain the counterfactual composed embedding $\hat{\mathrm{Q}}$ by Eq. 2. The corresponding

matching score is computed as $S_{\text{cf}}(\hat{\text{Q}}, \text{U})$. Finally, we sample K candidate targets using the *torch.multinomial* sampling function with a refined matching probability distribution:

$$\{\text{U}_i\}_{i=1}^{K} \sim \text{Multinomial}(\mathbb{1}_{\text{cf}} \cdot Softmax(S_{\text{fact}}), K). \quad (6)$$

Here, $\mathbb{1}_{\text{cf}}$ is the counterfactual-guided consistency constraint, which equals 1 when $S_{\text{fact}}/S_{\text{cf}} > 1$, and approaches 0 otherwise. The inherent motivation is that if the semantic integrity of Q^{T} is compromised but the similarity score remains unchanged, it suggests that the pairs may be overly dependent on visual content while ignoring intended modifications. Therefore, these negative sample pairs are assigned a lower probability during sampling. By doing so, we obtain K candidate rewards $\{\mathcal{R}(\text{Q}, \text{U}_i)\}_{i=1}^{K}$, which are as visually consistent and semantically relevant as possible, thereby enhancing the subsequent reward calculation.

Policy gradient with REINFORCE Leave-One-Out. Since the rewards are always non-negative, meaning the reward model inherently encourages all sampled retrieval behaviors. However, we expect the reward model to provide positive feedback for retrievals that align with both textual semantic guidance and visual references, while penalizing actions that are visually similar but semantically irrelevant. To achieve this, we employ an *advantage* function to reduce the variance of the estimator in Eq.4, by subtracting a *baseline* that has high covariance with the stochastic gradient, *i.e.*, $\mathcal{A}(\text{Q}, \text{U})$=$\mathcal{R}(\text{Q}, \text{U})$-*baseline*. Inspired by REINFORCE Leave-One-Out (RLOO) estimator (Kool, van Hoof, and Welling 2019), each reward for the query in the candidate pool can serve as a *baseline* for all other pairs:

$$\mathcal{A}(\text{Q}, \text{U}_i) = \mathcal{R}(\text{Q}, \text{U}_i) - \frac{1}{k-1}\sum_{j \neq i}^{k-1} \mathcal{R}(\text{Q}, \text{U}_j). \quad (7)$$

By doing so, the reward model is expected to provide positive feedback to retrieval behaviors with high reward score while penalizing undesirable outcomes. Then, policy updates can be performed on an average of gradient estimates for each sample, from $\text{U}_i \sim \pi_\theta(\cdot|\text{Q})$ (i=1, ..., K), resulting in a variance-reduced multi-sample Monte-Carlo estimation (Mnih and Rezende 2016), and Eq.4 is rewritten:

$$\begin{aligned} &\nabla_\theta \mathbb{E}_{\text{U} \sim \pi_\theta(\cdot|\text{Q})} \mathcal{R}(\text{Q}, \text{U}) \\ &= \frac{1}{k} \sum_{i=1}^{k} \mathcal{A}(\text{Q}, \text{U}_i) \cdot \nabla_\theta \log \pi_\theta(\text{U}_i|\text{Q}) \ \text{ for } \text{U}_i \sim \pi_\theta(\cdot|\text{Q}). \end{aligned} \quad (8)$$

Duplex Rewards Modeling. While TT-RLR has shown promising empirical results, it relies on the reference model's internal worldview. Specifically, the generated reference rewards (RR) $\mathcal{R}_{\text{ref}}$ are inherently limited by the knowledge boundaries of the reference model, which constrains the optimized model's ability to effectively adapt to new test data. This limitation is particularly pronounced in scenarios where the test data significantly diverges from the training distribution of the reference model. Furthermore, the inherent biases of the reference model may inadvertently propagate to the optimized model, leading to suboptimal performance. During test time, in fact, the CIR model can produce its own reward signals, which we term endogenous rewards (ER) $\mathcal{R}_{\text{endo}}$. Compared to conservative RR, ER is generated by the model itself during self-motivation and optimization, exhibiting better adaptability for the current data distribution. To this end, we propose a Duplex Rewards Modeling (DRM) module, which incorporates ER to provide task-specific corrections for RR. Note that ER and RR share the same candidate reward pool and reward generation process (Eqs. 5 to 7), while employing different reward models: a powerful off-the-shelf reference model for RR, and the currently adapted model for ER. Finally, the optimization objective in Eq.3 can be re-formulated as follows:

$$\max_\theta \mathbb{E}_{\text{U} \sim \pi_\theta(\cdot|\text{Q})}[\mathcal{R}_{\text{ref}}(\tilde{\text{Q}}, \text{U}) + \mathcal{R}_{\text{endo}}(\text{Q}, \text{U})]. \quad (9)$$

Note that $\tilde{\text{Q}}$ denotes the composed embedding extracted by the reference model, while Q is extracted by the adapted model, as shown in Fig. 2. Correspondingly, reference advantage $\mathcal{A}$ in Eq.8 is replaced by the duplex advantage, *i.e.*, $\mathcal{A}_{\text{ref}}$+$\mathcal{A}_{\text{endo}}$. By doing so, we balance both external knowledge and internal assessments, enhancing the model's adaptability to the specific characteristics of the current test data.

Experiments

Experimental Setup

Datasets and Metrics. We utilize three commonly used datasets in CIR: CIRR (Liu et al. 2021), FashionIQ (Wu et al. 2021), and COCO (Lin et al. 2014). FashionIQ focuses specifically on fashion-related retrieval, while the remaining datasets are designed for retrieval in open-domain scenes. Following the original benchmarks, we adopt Recall@k (R@k) as the evaluation metric for all datasets.

Implementation Details. For VLMs, we load weights from the official CLIP (Radford et al. 2021) for CLIP-ViT-B/16 and L/14, and OpenCLIP (Ilharco et al. 2021) for H/14 and G/14. For simplicity, we notate each as C-B16, C-L14, C-H14, and C-G14, respectively. For the retrieval model, the Layer Normalization layers are trainable, while the remaining layers are fixed. In contrast, all parameters of the reference reward model are fully frozen. We optimize the model for 1 step based on an individual test composed query, using the AdamW optimizer with a learning rate of 0.0005. The sampling factor K is set to 16 for CIRR and COCO, while 8 for FashionIQ. Besides, we set α=0.7 for FashionIQ and 0.8 on the remaining datasets, respectively.

Experimental Results

Comparison Methods. We compare TT-RLDR with recent state-of-the-art (sota) ZS-CIR and TTA methods. For ZS-CIR approaches, we include textual inversion methods such as SEARLE (Baldrati et al. 2023), Pic2Word (Saito et al. 2023), KEDs (Suo et al. 2024); LLM-based method, CIReVL (Karthik et al. 2024); as well as Slerp (Jang et al. 2024) and LinCIR (Gu et al. 2024). The results are taken from the respective original papers. Besides, the baseline "ZS: Image+Text" (ZS: I+T) denotes performing zero-shot retrieval with pre-trained VLMs weights, using the average embeddings of reference image and modification text. For

Method	Backbone	Dress		Toptee		Shirt		Average	
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
ZS: Image+Text	C-B16	14.03	30.74	19.02	33.91	17.03	30.62	16.69	31.76
ZS: Image+Text	C-L14	14.68	30.34	21.47	37.23	20.66	33.76	18.93	33.78
SEARLE†,* (Baldrati et al. 2023)	C-B32	18.54	39.51	25.70	46.46	24.44	41.61	22.89	42.53
Slerp† (Jang et al. 2024)	C-B32	20.53	41.00	26.98	46.77	23.75	40.92	23.75	42.90
Pic2Word†,* (Saito et al. 2023)	C-L14	20.00	40.20	27.90	47.40	26.20	43.60	24.70	43.73
KEDs†,* (Suo et al. 2024)	C-L14	21.70	43.80	29.90	51.90	28.90	48.00	26.83	47.90
LinCIR†,* (Gu et al. 2024)	C-L14	20.90	42.40	28.80	50.20	29.10	46.80	26.27	46.47
CIReVL† (Karthik et al. 2024)	C-L14	24.79	44.76	31.36	53.65	29.49	47.40	28.55	48.60
TENT (Wang et al. 2021)	C-B16	13.54	30.94	26.31	46.00	25.56	41.27	21.80	39.40
CMF (Lee and Chang 2024)	C-B16	17.06	37.23	25.34	44.31	25.32	40.73	22.57	40.76
DeYO (Lee et al. 2024)	C-B16	15.08	35.52	25.40	46.35	25.61	40.48	22.03	40.78
KD (Hinton et al. 2014)	C-B16	19.13	40.16	25.80	44.42	24.14	39.30	23.02	41.29
RLCF (Zhao et al. 2024)	C-B16	20.87	40.79	27.19	47.07	25.25	42.69	24.44	43.52
TT-RLDR (Ours)	C-B16	21.33	42.49	29.63	51.39	28.92	46.82	26.62	46.90
	C-L14	**24.86**	**46.62**	**34.52**	**54.77**	**33.46**	**52.65**	**30.95**	**51.35**

Table 1: Results against ZS-CIR and TTA methods on FashionIQ validation set. † indicates results from the original paper, and $*$ denotes using additional data for training. The best results are marked in underline and **bold** for C-B16 and C-L14, respectively.

TTA techniques, they can be categorized into normalization calibration (TENT (Wang et al. 2021)), entropy minimization (CMF (Lee and Chang 2024), DeYO (Lee et al. 2024)), and reward mechanism (RLCF (Zhao et al. 2024)) approaches. We implement these methods using their official codes and share the same backbone, *i.e.*, C-B16. By default, C-L14 is deployed as the reward model for our methods or the teacher model for knowledge distillation (KD) (Hinton et al. 2014). Additionally, we include a C-L14 backbone with C-H14 reward to facilitate a fair comparison of ZS-CIR methods using the same backbone.

Results Analysis. Our main quantitative experimental results are presented in Tables 1, and 2. In Table 1, we summarize the main results on FashionIQ validation set, requiring accurate localization of specific attributes within fashion images. From the table, we can draw the following conclusions: (1) Across all metrics with different CLIP backbones, our TT-RLDR consistently outperforms all baseline approaches. (2) Compared to textual inversion methods such as Pic2Word, TT-RLDR (C-L14) achieves impressive performance across multiple metrics (averaging improvement of 6.9%), even without any additional training corpus. (3) Similarly, compared to CIReVL, which employs extra captioner and LLM modules to express and reason the composed queries, our approach outperforms it by 2.4% in average R@10 and 2.75% in average R@50. (4) Under similar architectures (C-B16 *v.s.* C-B32), existing TTA methods, which perform unsupervised adaptation using only current test data, exhibiting comparable results to ZS-CIR (SEARLE, Slerp) methods, without the need for external training data or modules. Additionally, our method (C-B16) surpasses all these TTA techniques by 2.18% to 4.82% in average R@10, and by 3.38% to 7.5% in average R@50. These results strongly support TT-RLDR's effectiveness.

We further evaluate TT-RLDR's capabilities on the open-domain datasets CIRR and COCO, as indicated in Table 2. Compared to Pic2Word, our method demonstrates significant improvement, achieving an average performance increase of 5.96% on CIRR and 2.34% on COCO. It is worth noting that even when using fewer parameters, *i.e.*, C-B16, TT-RLDR still consistently surpasses existing ZS-CIR models by a substantial margin on CIRR, including those equipped with the larger C-L14 backbone. Moreover, by comparing the results between our TT-RLDR and current TTA techniques, we observe that our method achieves an average retrieval performance improvement of 3.56% to 5.08% on CIRR and 1.5% to 4.1% on COCO. These findings highlight the robustness of our approach, demonstrating its ability to deliver precise retrieval even in the presence of noisy data and its adaptability across diverse real-world scenarios.

Ablation Studies

We examine the contributions of core components in TT-RLDR on FashionIQ and CIRR (Table 3). For efficiency, we apply *C-B16 as the default backbone*, equipped with *C-L14 as the reward model* in ablation experiments. **(1) Models '2-4' validate the effect of key modules.** First, removing the CMS strategy and applying top-k sampling with the highest similarity scores (model '2') results in a significant average drop of 2.29% for FashionIQ and 1.81% for CIRR. This indicates that the proposed CMS effectively constructs a reward pool for reward estimation, where candidate targets are not only visually similar to the composed query but also align with its modification semantics. Furthermore, using only reference rewards (RR, model '3') yields suboptimal results, with a decrease of 1.6% and 1.23% compared to our full model (model '1'). This is because the frozen RR provides robust yet conservative feedback signals, lacking adaptability to current test data. Similarly, we observe that employing only endogenous rewards (ER, model '4') leads to a further performance decline of 2.20% and 1.02% from model 3'. We hypothesize that ER relies on the adapted

Method	Back bone	CIRR R@1	CIRR R@10	CIRR R@50	COCO R@1	COCO R@10
ZS: I+T	C-B16	12.44	50.13	77.28	9.84	25.02
ZS: I+T	C-L14	12.84	50.44	78.43	10.02	26.60
SEARLE†,*	C-B32	24.00	66.82	89.78	-	-
Slerp†	C-B32	24.22	66.48	89.59	-	-
Pic2Word†,*	C-L14	23.90	65.30	87.80	11.50	33.40
KEDs†,*	C-L14	26.40	67.20	89.20	12.00	34.90
LinCIR†,*	C-L14	25.00	66.70	-	11.70	34.20
CIReVL†	C-L14	24.55	64.92	86.34	-	-
TENT	C-B16	23.97	67.40	89.17	10.68	27.02
CMF	C-B16	24.04	66.83	88.83	10.33	28.91
DeYO	C-B16	23.03	65.68	87.85	9.89	29.38
KD	C-B16	23.80	65.20	87.90	10.21	30.97
RLCF	C-B16	24.54	67.15	89.41	11.00	31.85
TT-RLDR	C-B16	27.91	71.98	91.89	11.84	34.01
(Ours)	C-L14	**29.71**	**73.02**	**92.16**	**12.96**	**36.62**

Table 2: Comparison results against ZS-CIR (upper part) and TTA (middle part) methods on CIRR and COCO test data.

model's internal perspective, potentially steering the model towards reinforcing its own flaws without the guidance of RR. **(2) Models '5-7' evaluate different backbones.** Compared to C-B16 (model '1'), C-L14 (model '5') exhibits certain advantages, albeit requiring more updated parameters. We also explore other architectures, namely BLIP2 (model '6') and ALBEF (model '7'), and find their performance to be suboptimal, despite their greater computational costs. We speculate that this is due to a significant architectural disparity between these backbones and the reference model C-L14, which hinders rewards alignment. Thus, we chose C-B16 as the default backbone due to its computational efficiency. **(3) Models '8-11' assess various reward model.** TT-RLDR relies on the good quality of the reward models. It can be seen that utilizing the lighter C-B16 (model '8') to provide reference rewards results in a moderate performance decline compared to C-L14 (model '1'). Notably, even with C-B16 as the reward model, our TT-RLDR still surpasses existing TTA methods (in Tables 1 and 2) by a significant margin. When deploying more large-scale reference models (models '9', '10', '11'), the average retrieval performance of TT-RLDR is further improved by 1.53% to 1.92% on FashionIQ and 0.57% to 1.05% on CIRR. Here, CLIP-Ensemble adopts an averaged reward sum of {C-L14, H14, G14}. Thus, we default to applying C-L14 as the reward model for its good balance between retrieval accuracy and resource efficiency.

Impact of Hyper-parameters.We investigate the influence of two essential hyper-parameters in our method, *i.e.*, sampling factor K and interpolation weight α. We first vary the sampling factor K and display the results in Fig. 3(a). As K increases, the results initially rise and then gradually fall. This is because a small K leads to insufficient reward estimation, while a large K introduces more negative samples, generating erroneous rewards. Additionally, as shown in Fig. 3(b), we vary α from 0 to 1, where 0 represents using only the reference image, while 1 denotes using only the modification text. From these results, we observe that increasing the weight given to modification text improves the retrieval performance, up to α=0.7 for FashionIQ and α=0.8 for CIRR. However, the image also plays a significant role, as evidenced by the significant drop in performance when the weight of images is continuously weakened (α=1).

Method	FashionIQ-Avg R@10	FashionIQ-Avg R@50	CIRR R@1	CIRR R@10	CIRR R@50
1. Full model	26.62	46.90	27.91	71.98	91.89
Significance of key modules of TT-RLDR					
2. w/o CMS	24.89	44.06	26.00	69.84	90.53
3. w/o DRM (RR-only)	25.67	44.65	26.21	70.80	91.08
4. w/o DRM (ER-only)	23.42	42.52	25.86	69.67	89.50
Impact of different backbone models					
5. CLIP-ViT-L/14	27.16	47.83	28.13	72.07	91.74
6. BLIP2-ViT-L/16	22.34	42.77	24.46	67.53	88.32
7. ALBEF-ViT-L/16	20.48	40.39	22.96	67.83	90.57
Impact of various reference reward models					
8. CLIP-ViT-B/16	25.15	44.01	26.43	70.27	90.92
9. CLIP-ViT-H/14	28.11	48.46	28.33	72.76	92.72
10. CLIP-ViT-G/14	28.26	49.10	28.50	72.85	92.98
11. CLIP-Ensemble	28.14	48.70	28.72	73.03	93.27

Table 3: Ablation study on FashionIQ and CIRR.

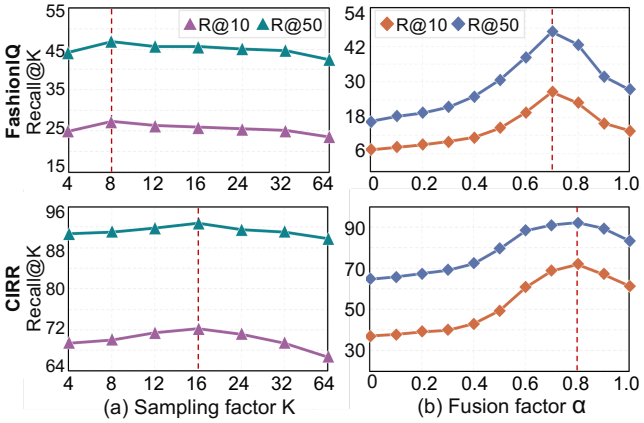

Figure 3: Parameter variation with different K and α.

Conclusion

In this paper, we introduce Test-time CIR (TT-CIR), which efficiently adapts pre-trained VLMs to unlabeled data during testing. Besides, we propose TT-RLDR with two core components: the CMS strategy, which accurately identifies the candidate reward pool for reward estimation, and the DRM module, which generates both stability and adaptability rewards to guide model optimization. Together, CMS and DRM effectively address TT-CIR challenges while achieving precise retrieval with reduced resource and computational consumption. Extensive experiments show that our method outperforms current ZS-CIR and TTA approaches, demonstrating robust adaptability and effectiveness in real-world scenarios. Future work will focus on refining the reward mechanism and applying it to other multimodal tasks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62376196, Grant 62036012, Grant U23A20387, Grant 62106262, Grant 62202331, Grant 62206200, and Grant 62276118, in part by Tianjin Natural Science Foundation under Grant 22JCYBJC00030, and Grant 24JCJQJC00190, and in part by the Beijing Natural Science Foundation under Grant L252032.

References

Ahmadian, A.; Cremer, C.; Gallé, M.; Fadaee, M.; Kreutzer, J.; Pietquin, O.; Üstün, A.; and Hooker, S. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*.

Bai, Y.; Xu, X.; Liu, Y.; Khan, S.; Khan, F.; Zuo, W.; Goh, R. S. M.; and Feng, C.-M. 2024. Sentence-level prompts benefit composed image retrieval. In *ICLR*.

Baldrati, A.; Agnolucci, L.; Bertini, M.; and Del Bimbo, A. 2023. Zero-shot composed image retrieval with textual inversion. In *ICCV*, 15338–15347.

Baldrati, A.; Bertini, M.; Uricchio, T.; and Del Bimbo, A. 2022. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *CVPR*, 4959–4968.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607.

Chen, Y.; Ma, Z.; Zhang, Z.; Qi, Z.; Yuan, C.; Shan, Y.; Li, B.; Hu, W.; Qie, X.; and Wu, J. 2023. Vilem: Visual-language error modeling for image-text retrieval. In *CVPR*, 11018–11027.

Chen, Y.; Zhong, H.; He, X.; Peng, Y.; Zhou, J.; and Cheng, L. 2024. FashionERN: Enhance-and-Refine Network for Composed Fashion Image Retrieval. In *AAAI*, volume 38, 1228–1236.

Cho, J.; Yoon, S.; Kale, A.; Dernoncourt, F.; Bui, T.; and Bansal, M. 2022. Fine-grained Image Captioning with CLIP Reward. In *NAACL*, 517–527.

Döbler, M.; Marsden, R. A.; and Yang, B. 2023. Robust mean teacher for continual and gradual test-time adaptation. In *CVPR*, 7704–7714.

Feng, C.-M.; Bai, Y.; Luo, T.; Li, Z.; Khan, S.; Zuo, W.; Goh, R. S. M.; and Liu, Y. 2025. Vqa4cir: Boosting composed image retrieval with visual question answering. In *AAAI*, volume 39, 2942–2950.

Gu, G.; Chun, S.; Kim, W.; Kang, Y.; and Yun, S. 2024. Language-only training of zero-shot composed image retrieval. In *CVPR*, 13225–13234.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*, 7514–7528.

Hinton, G.; Vinyals, O.; Dean, J.; and et al. 2014. Distilling the knowledge in a neural network. In *NeurIPS*.

Ilharco, G.; Wortsman, M.; Wightman, R.; Gordon, C.; Carlini, N.; Taori, R.; Dave, A.; Shankar, V.; Namkoong, H.; Miller, J.; Hajishirzi, H.; Farhadi, A.; and Schmidt, L. 2021. OpenCLIP.

Jang, Y. K.; Huynh, D.; Shah, A.; Chen, W.-K.; and Lim, S.-N. 2024. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. In *ECCV*, 239–254.

Karthik, S.; Roth, K.; Mancini, M.; and Akata, Z. 2024. Vision-by-Language for Training-Free Compositional Image Retrieval. In *ICLR*.

Kool, W.; van Hoof, H.; and Welling, M. 2019. Buy 4 REINFORCE Samples, Get a Baseline for Free!

Lee, J.; Jung, D.; Lee, S.; Park, J.; Shin, J.; Hwang, U.; and Yoon, S. 2024. Entropy is not Enough for Test-Time Adaptation: From the Perspective of Disentangled Factors. In *ICLR*.

Lee, J.-H.; and Chang, J.-H. 2024. Continual Momentum Filtering on Parameter Space for Online Test-time Adaptation. In *ICLR*.

Levy, M.; Ben-Ari, R.; Darshan, N.; and Lischinski, D. 2024. Data roaming and quality assessment for composed image retrieval. In *AAAI*, volume 38, 2991–2999.

Li, H.; Hu, P.; Zhang, Q.; Peng, X.; Yang, M.; et al. 2025a. Test-time Adaptation for Cross-modal Retrieval with Query Shift. In *ICLR*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 12888–12900.

Li, Z.; Chen, Z.; Wen, H.; Fu, Z.; Hu, Y.; and Guan, W. 2025b. Encoder: Entity mining and modification relation binding for composed image retrieval. In *AAAI*, 5101–5109.

Liang, J.; He, R.; and Tan, T. 2025. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision (IJCV)*, 133(1): 31–64.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755.

Liu, Z.; Rodriguez-Opazo, C.; Teney, D.; and Gould, S. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, 2125–2134.

Liu, Z.; Sun, H.; Peng, Y.; and Zhou, J. 2024. DART: dual-modal adaptive online prompting and knowledge retention for test-time adaptation. In *AAAI*, volume 38, 14106–14114.

Mirza, M. J.; Micorek, J.; Possegger, H.; and Bischof, H. 2022. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *CVPR*, 14765–14775.

Mnih, A.; and Rezende, D. 2016. Variational inference for monte carlo objectives. In *ICML*, 2188–2196.

Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; and Tan, M. 2023. Towards Stable Test-Time Adaptation in Dynamic Wild World. In *ICLR*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*, volume 35, 27730–27744.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, volume 36, 53728–53741.

Saito, K.; Sohn, K.; Zhang, X.; Li, C.-L.; Lee, C.-Y.; Saenko, K.; and Pfister, T. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *CVPR*, 19305–19314.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shu, M.; Nie, W.; Huang, D.-A.; Yu, Z.; Goldstein, T.; Anandkumar, A.; and Xiao, C. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, volume 35, 14274–14289.

Suo, Y.; Ma, F.; Zhu, L.; and Yang, Y. 2024. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *CVPR*, 26951–26962.

Tang, Y.; Yu, J.; Gai, K.; Zhuang, J.; Xiong, G.; Hu, Y.; and Wu, Q. 2024. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *AAAI*, volume 38, 5180–5188.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *ICLR*.

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022. Continual test-time domain adaptation. In *CVPR*, 7201–7211.

Wang, Y.; Yang, Q.; Zeng, Z.; Ren, L.; Liu, L.; Peng, B.; Cheng, H.; He, X.; Wang, K.; Gao, J.; et al. 2025. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*.

Wen, Z.; Niu, S.; Li, G.; Wu, Q.; Tan, M.; and Wu, Q. 2023. Test-time model adaptation for visual question answering with debiased self-supervisions. *IEEE Transactions on Multimedia (TMM)*, 26: 2137–2147.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8: 229–256.

Wu, H.; Gao, Y.; Guo, X.; Al-Halah, Z.; Rennie, S.; Grauman, K.; and Feris, R. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *CVPR*, 11307–11317.

Zhang, C.; Stepputtis, S.; Sycara, K.; and Xie, Y. 2024a. Dual prototype evolving for test-time generalization of vision-language models. In *NeurIPS*, volume 37, 32111–32136.

Zhang, F.; Qu, S.; Shi, F.; and Xu, C. 2024b. Overcoming the Pitfalls of Vision-Language Model for Image-Text Retrieval. In *ACM MM*, 2350–2359.

Zhang, M.; Levine, S.; and Finn, C. 2022. Memo: Test time robustness via adaptation and augmentation. In *NeurIPS*, volume 35, 38629–38642.

Zhao, S.; Wang, X.; Zhu, L.; and Yang, Y. 2024. Test-time adaptation with clip reward for zero-shot generalization in vision-language models. In *ICLR*.

Zhou, Y.; Fan, Z.; Cheng, D.; Yang, S.; Chen, Z.; Cui, C.; Wang, X.; Li, Y.; Zhang, L.; and Yao, H. 2024. Calibrated self-rewarding vision language models. In *NeurIPS*, volume 37, 51503–51531.

Zuo, Y.; Zhang, K.; Sheng, L.; Qu, S.; Cui, G.; Zhu, X.; Li, H.; Zhang, Y.; Long, X.; Hua, E.; et al. 2025. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*.