This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CAST: Context-Aware Dynamic Latent Space Transformation for Interactive Text-to-Image Retrieval

Xuanzuo Lin[1,2], Min Zhang[3], Daizong Liu[4†], Zhiwen Zuo[1,2]
Xun Yang[5], Changting Lin[6], Xun Wang[1,2], Jianfeng Dong[1,2†]
[1]Zhejiang Gongshang University,
[2]Zhejiang Key Laboratory of Big Data and Future E-Commerce Technology,
[3]East China Normal University, [4]Wuhan University,
[5]University of Science and Technology of China, [6]Zhejiang University

Abstract

Interactive Text-to-Image Retrieval (I-TIR) aims to refine image retrieval results through natural language dialogues, which allows users to progressively supplement or correct their search intention across multiple rounds, enabling a more precise and user-aligned visual search experience. However, existing methods perform cross-modal retrieval within a fixed multimodal feature space, mapping all dialogue text and images onto the same static embedding manifold. Such static formulation easily causes semantic vagueness, making it difficult to capture subtle embedding shifts in the user's updated intention for fine-grained retrieval. To address this limitation, we propose Context-Aware Latent Space Transformation (CAST), a lightweight framework that dynamically transforms the common latent space of both textual and visual representations according to the specific evolving user's search intention, enabling fine-grained and adaptive semantic alignment. The core of CAST is the Context-Aware Space Regulator (CASR), a crucial space transformation module composed of two key components: (1) the Context-Aware Low-Rank Projector (CLP), which learns to predict the projection direction of embedding space based on the intent's context; and (2) the Context-Guided Modulator (CGM), which adaptively determines appropriate projection strength. CASR is highly lightweight, adding negligible parameters and computational overhead, and can be seamlessly integrated into diverse I-TIR frameworks. Extensive experiments demonstrate the effectiveness of our proposed framework, indicating that it can serve as a general, plug-and-play solution for efficient and scalable interactive text-to-image retrieval. Our source code is available at https://github.com/HuiGuanLab/CAST.

† Corresponding authors: Daizong Liu and Jianfeng Dong.

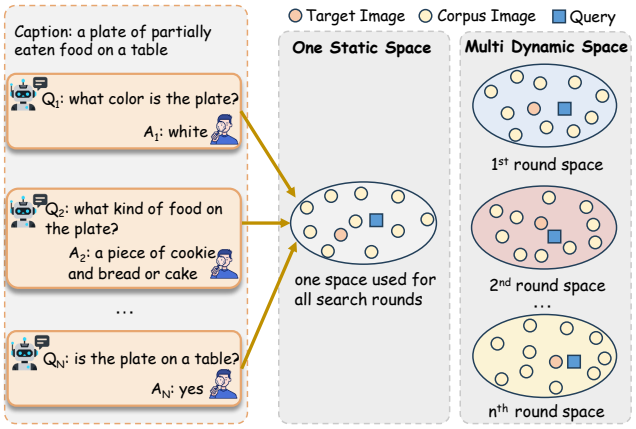


Figure 1. Illustration of our motivation. Different from the previous works that typically perform cross-modal matching in *one* static multimodal feature space throughout all search round, we propose to adaptively utilize *multiple* dynamic feature spaces based on progressively refined user's search intention. Each space is expected to be more discriminative with respect to a specific "topic" that the user is focused on in corresponding search round.

1. Introduction

Traditional text-to-image retrieval [4, 5, 8, 21, 25, 31, 38, 44, 45] typically relies on a single-round text query, where the user must provide a complete and precise description in one attempt to search for the target image from a large candidate pool. However, such one-shot retrieval often demands an overly detailed and comprehensive description, which is difficult for users to formulate in one go. As a result, the retrieval performance is usually suboptimal compared to the users' ultimate goal. To address this limitation, Interactive Text-to-Image Retrieval (I-TIR) [13, 19, 20] has recently been introduced. It enables multi-round interactions between the user and the system, allowing the user to iteratively refine the query, clarify intentions, and progressively guide the retrieval process toward the desired image.

Existing research on this more challenging interactive retrieval task primarily focuses on two key aspects. First, some studies aim to enhance the retrieval model's accurate understanding of dialogue content. These approaches often exploit large language models (LLMs) to reconstruct the dialogue history, transforming redundant or ambiguous expressions into clearer, more focused queries that improve intent comprehension [19, 32, 33]. Second, other methods emphasize the generation of higher-quality dialogue questions. These approaches leverage LLMs to select more discriminative questions, guiding user responses to better distinguish between candidate images [19]. Despite their great progress, these methods remain fundamentally constrained by traditional text-to-image retrieval methods, where cross-modal matching is performed in one static embedding space throughout all search rounds. As illustrated in Figure 1, they implicitly assume that the geometric relationship between textual and visual features remains fixed throughout the dialogue. In practice, each round of user feedback introduces new semantic constraints that dynamically transform the retrieval focus [16, 17] . For instance, if the user emphasizes "the color of the dog" in a particular round, a latent space should become more discriminative with respect to the color. If the feature space is not dynamically transformed according to the evolving semantics, the model merely reformulates queries on a fixed representation manifold. This static paradigm limits the effective utilization of contextual information, preventing the model from accurately adapting to semantic changes in the dialogue, thereby affecting the accuracy of the retrieval results.

To address the above issues, we propose Context-Aware Latent Space Transformation (CAST), a general and efficient framework that dynamically transforms the cross-modal latent feature space according to the evolving dialogue context for fine-grained semantic mining. The core of our framework is a lightweight space transformation module that answers two fundamental questions essential for dynamic feature space transformation: (1) In which user-intent semantic direction should the model refine its search? (2) To what extent should this semantic modulation occur? To address these, we propose a context-aware space regulator that consists of two synergistic components: (1) CLP, which provides context-guided transformation directions through low-rank projections; (2) CGM, which adaptively modulates the appropriate projection magnitude. Together, these components enable the multimodal feature space to dynamically transform according to user intention, thereby directly improving retrieval performance in multi-round interactions. A key advantage of this design lies in its efficiency and plug-and-play compatibility. Thanks to its lightweight design, it can dynamically evolve semantic space with minimal computational overhead, while establishing a general and scalable paradigm for I-TIR.

In summary, our contributions are threefold:

- We identify the limitations of existing frameworks that operate within a static multimodal embedding space, and we are the first to propose the use of more fine-grained dynamic latent space in interactive text-to-image retrieval.
- We propose CAST, which directly modulates the geometry of the multimodal space through two synergistic modules: CLP learns the transformation direction, and CGM controls the magnitude, enabling stable and context-aware transformation of multimodal embeddings.
- CAST can be seamlessly integrated into existing I-TIR architectures with minimal computational overhead. Experiments on the VisDial benchmark demonstrate that CAST consistently outperforms state-of-the-art methods, particularly in later dialogue rounds.

2. Related Work

Text-to-Image Retrieval. Traditional text-to-image retrieval [5, 6, 9, 10, 21, 25, 27–31, 38, 44] follows a single-round matching paradigm, where the goal is to embed visual and textual inputs into a shared semantic space and measure their similarity through contrastive learning. Large-scale vision-language models [15, 23, 38, 48] such as CLIP [38] and BLIP [23] have demonstrated strong alignment capability, enabling efficient cross-modal retrieval by comparing the cosine similarity between image and text embeddings. Different from the above methods relying on a static feature space, QuARI [47] attempts to alleviate this limitation by dynamically transforming both textual and visual features conditioned on the query. Nevertheless, these models operate under a single-round assumption, where each query is treated independently without contextual history. This formulation limits their ability to model evolving user intent or cumulative reasoning across dialogue rounds. In contrast, I-TIR requires the retrieval model to continuously integrate new information from prior exchanges, refining its understanding of visual semantics as the conversation unfolds.

Interactive Text-to-Image Retrieval. Recent advances have extended I-TIR from static single-round matching to iterative dialogue-driven refinement. Early approaches such as ChatIR [20] and PlugIR [19] leverage LLMs to generate discriminative questions and reformulate user interactions into more informative textual queries, effectively guiding the model towards fine-grained intent understanding. This line of work emphasizes dialogue generation and linguistic reformulation. A second line of research focuses on enhancing retrieval representation and reasoning. Methods like Diffusion-Augmented Retrieval (DAR) [32] and ImageScope [33] introduce diffusion priors or LLMs reasoning to enrich semantic diversity and interpretability, yet their underlying feature geometry remains static. In contrast, our framework dynamically reshapes both textual and

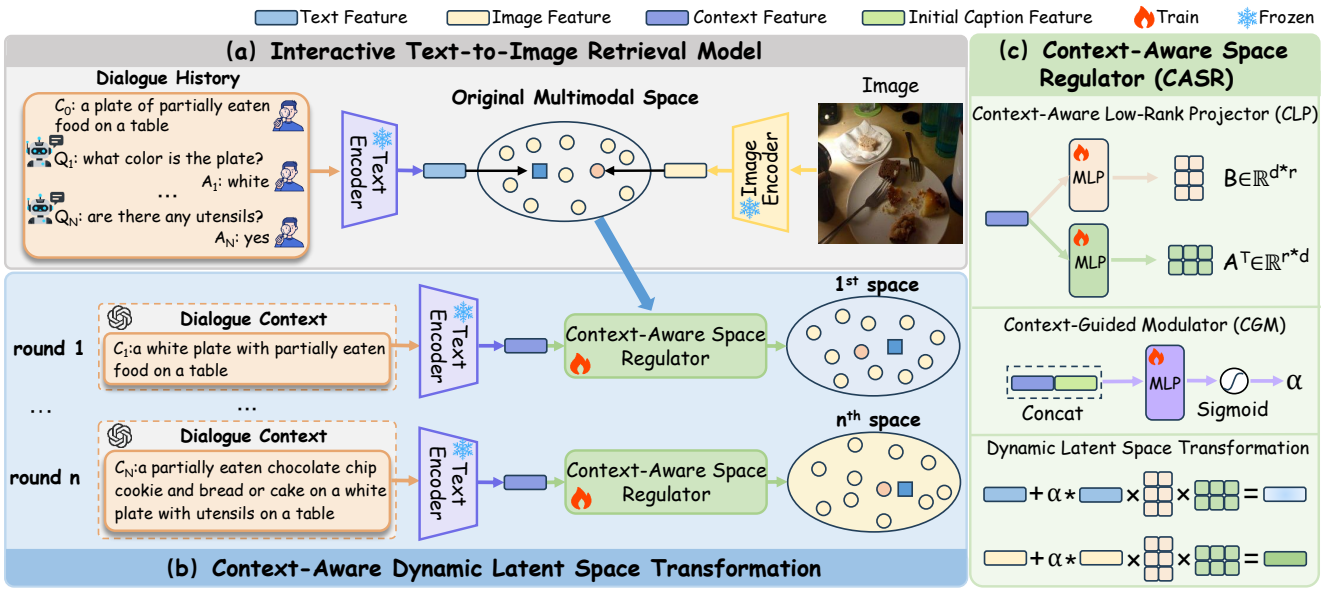


Figure 2. (a) A traditional interactive text-to-image retrieval framework using a static multimodal latent space for cross-modal alignment. (b) Illustration of our proposed context-aware dynamic latent space transformation, which aims to dynamically transform the static multimodal latent space into a new space according to the user's search intention. (c) The CASR consists of two key modules: The CLP module provides the direction for the feature space transformation, while the CGM module controls the strength of the transformation.

visual representations according to dialogue context, enabling continuous adaptation of the retrieval space for fine-grained semantic alignment multi-round interaction.

Multi-round Composed Image Retrieval. Multi-round Composed Image Retrieval (CIR) [12, 13, 16, 46] is much relevant to our target I-TIR task, where both address retrieval in an iterative and multi-round setting. Their key difference lies in the query formulation: CIR uses both text and image, while I-TIR relies solely on text. A representative CIR work, GeneCIS [43], argues that similarity is not static but can vary under different conditions, which aligns with our motivation. A key difference is that GeneCIS conditions the representation primarily on the query side, while our method performs a coupled transformation that updates both the query and the image.

3. Method

3.1. Problem Formulation

Interactive text-to-image retrieval is formulated as a conversational retrieval task in which a model poses questions [36, 42] and a user provides answers to progressively refine the retrieval target. At round t, the dialogue history can be expressed as:

$$\mathcal{H}_t = C_0, (Q_1, A_1), (Q_2, A_2), \ldots, (Q_t, A_t), \quad (1)$$

where Q_i denotes the model-generated question and A_i represents the user's natural-language response. The most common approach is to concatenate the dialogue context $\mathcal{H}_t$ into a single textual sequence as input, and compute retrieval scores within a static embedding space [11, 18] shared across all rounds:

$$s_t = \mathrm{sim}(f_T(\mathcal{H}_t), f_I(I)), \quad (2)$$

where $f_T(\cdot)$ and $f_I(\cdot)$ denote the text and image encoders, respectively, and I represents a candidate image from the corpus.

3.2. Motivation

Despite recent progress in interactive retrieval, most existing frameworks operate within a static multimodal feature space, assuming that the relationship between text and images remains fixed across all dialogue rounds. However, interactive retrieval is inherently dynamic: each user response provides new semantic constraints that refine and reshape the retrieval intent. When the feature space itself remains unchanged, these evolving semantics are merely projected into the same embedding geometry, causing the query to drift without truly reflecting the updated intent. As a result, subtle contextual cues such as newly mentioned attributes or refined object relations are easily lost in static representations. This limitation motivates a new perspective: each dialogue round should act as a contextual modulation step that dynamically transforms the multimodal representation space according to the current intent context. By allowing both textual and visual embeddings to evolve in response to dialogue context, the retrieval system can continuously refocus on semantically relevant regions, achieving finer-grained alignment and more faithful modeling of user in-

tent. Details of our pipeline are illustrated in Figure 2.

3.3. Context-Aware Space Regulator

The proposed context-aware space regulator module is the core module of our multimodal feature space transformation. It operates in three stages: dialogue context generation, context-aware low-rank projection, and context-guided modulation. In the first stage, we generate the dialogue context u_t from the dialogue history, which serves as the conditioning signal for the subsequent feature space transformation. Then, the direction of the transformation is controlled based on this signal, and finally, the magnitude of the transformation is modulated accordingly. These operations work together to enable the model to adaptively adjust the feature space in alignment with user intent.

At a high level, CASR performs a dynamic transformation of the embedding space, where features are modulated according to the semantic condition u_t. Formally, given a feature vector x_t in the original space, the reshaped one through space transformation at round t can be expressed as:

$$\mathcal{G}(x_t \mid u_t) = \mathrm{LN}\big(x_t + \mathcal{F}(x_t \mid u_t)\big), \tag{3}$$

where $\mathcal{F}$ represents a context-conditioned transformation function. We adopt an additive residual formulation, which preserves the original representation while allowing the transformation term $\mathcal{F}(x_t, u_t)$ to inject context-dependent semantic shifts. This residual design stabilizes training and ensures that contextual modulations act as smooth refinements rather than full replacements for the base feature. Through this formulation, CASR continuously transforms the multimodal feature manifold according to evolving dialogue semantics, achieving context-aware refinement of both modalities. We next detail how each component contributes to realizing this dynamic transformation.

Dialogue Context Generation via LLM. To ensure that the transformation direction accurately aligns with the user's intent, the model requires a semantic condition that comprehensively reflects the entire dialogue context rather than a single utterance. We therefore reference PlugIR [19] and introduce a dialogue context generation mechanism that leverages an LLM. [2, 3, 35] to integrate and refine prior dialogue rounds. By resolving ambiguities, the LLM produces a concise contextual summary serving as the semantic condition u_t for adaptive feature-space transformation:

$$C_t = \mathrm{LLM}(\mathcal{H}_t), \quad u_t = f_T(C_t), \tag{4}$$

where u_t encodes all contextual semantics up to round t, providing a reliable basis for dynamic transformation of the multimodal feature space.

Context-Aware Low-Rank Projector. The most common approach to transform the feature space is to apply a full $d \times d$ linear transformation to map features. However, such an operation is computationally expensive and may distort the geometry of the original embedding space. To achieve efficient and interpretable transformation, inspired by prior studies [1, 14, 40, 47], we instead adopt a low-rank conditional projection. By introducing a low-rank constraint, our model transforms features only along the major directions relevant to the semantic condition u_t, achieving lightweight, stable, and direction-aware semantic transformation. We then generate two feature space projection matrices conditioned on u_t as:

$$A_t = \mathrm{MLP}_A(u_t) \in \mathbb{R}^{d \times r}, \quad B_t = \mathrm{MLP}_B(u_t) \in \mathbb{R}^{d \times r}, \tag{5}$$

where $r \ll d$ is the rank constraint. To maintain numerical stability, we apply column-wise ℓ_2 normalization before computing the conditional projection as:

$$\hat{A}_t = \frac{A_t}{\|A_t\|_2}, \quad \hat{B}_t = \frac{B_t}{\|B_t\|_2}. \tag{6}$$

The conditional projection is then computed as:

$$\mathcal{P}(x_t \mid u_t) = (x_t \hat{B}_t)\hat{A}_t^\top, \quad \mathcal{P}(x_t \mid u_t) \in \mathbb{R}^{1 \times d}. \tag{7}$$

This operation projects x_t into a compact low-rank subspace via $\hat{B}_t$ and reprojects it through $\hat{A}_t^\top$, forming a context-conditioned directional transformation that deforms the semantic manifold along axes defined by u_t.

Context-Guided Modulator. While CLP determines the directional transformation, we further argue that the magnitude of this transformation is also crucial. Excessive modulation may lead to instability, while insufficient adjustment can limit the model's capability. To achieve balanced and adaptive control, we introduce a lightweight context-guided modulator that dynamically scales the magnitude of feature-space modulation based on the semantic divergence between the initial condition u_0 and the current context u_t. It predicts an adaptive coefficient $\alpha_t \in [0, 1]$ as:

$$\alpha_t = \sigma(\mathrm{MLP}([u_0; u_t])), \tag{8}$$

where $\sigma(\cdot)$ denotes the sigmoid activation. We then define the modulation function introduced earlier as:

$$\mathcal{F}(x_t \mid u_t) = \alpha_t \cdot \mathcal{P}(x_t \mid u_t), \tag{9}$$

This formulation allows $\mathcal{F}$ to adaptively modulate its transformation strength according to semantic divergence, ensuring that the transformation of the feature space across dialogue rounds is appropriate.

3.4. Dynamic Latent Space Transformation

After obtaining the context-conditioned projection and modulation operations guided by the dialogue context u_t,

given a textual feature x_t^T and an image feature x_t^I, their corresponding transformed representations are computed as:

$$z_t^T = \mathcal{G}(x_t^T \mid u_t), \quad z_t^I = \mathcal{G}(x_t^I \mid u_t), \tag{10}$$

After modulation, retrieval is performed by computing the cosine similarity between context-guided textual embedding z_t^T and each candidate image representation $z_t^{I_k}$:

$$s_t^{(k)} = \mathrm{sim}(z_t^T, z_t^{I_k}) = \frac{z_t^T \cdot z_t^{I_k}}{\|z_t^T\| \|z_t^{I_k}\|}, \quad k = 1, \ldots, N. \tag{11}$$

All candidate images $\{I_k\}$ are then ranked by their contextual similarity scores $s_t^{(k)}$, yielding the retrieval results for round t. Through this process, CASR dynamically aligns textual and visual representations according to evolving user intent, where the low-rank projection provides direction-aware feature space transformation and the adaptive gating mechanism regulates modulation strength based on contextual divergence. Together, these components ensure stable yet expressive evolution of the multimodal feature space across dialogue rounds.

3.5. Training Objective

In each training batch, dialogue samples are randomly truncated to different lengths and each sample contains distinct dialogue content. Under these sample-specific contexts, both textual and visual features are modulated accordingly. We define a context-guided contrastive loss to align the modulated text embedding $z_t^{T_i}$ with its corresponding context-conditioned image embedding $z_t^{I_i}$:

$$\mathcal{L}_{\text{cgc}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\left(\mathrm{sim}(z_{t_i}^{T_i}, z_{t_i}^{I_i})/\tau\right)}{\sum_{j=1}^{B} \exp\left(\mathrm{sim}(z_{t_i}^{T_i}, z_{t_i}^{I_j})/\tau\right)}, \tag{12}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature parameter.

Unlike conventional contrastive objectives [22, 37, 38] that assume a single static embedding space, our formulation performs alignment within a set of context-adaptive subspaces, each dynamically modulated by its own reconstructed dialogue condition. We introduce a lightweight regularization on α_t to ensure stable modulation and smooth feature evolution across dialogue rounds. Overall, this unified contrastive objective drives the model to learn a coherent, dialogue-aware representation space that evolves smoothly with contextual dynamics.

4. Experiments

4.1. Experimental Setup

Dataset. Following previous works [19, 20, 32, 33], all experiments are conducted on the VisDial [7] dataset, which is built upon COCO [26] images and serves as a standard benchmark for I-TIR. To evaluate the generalization ability of our method, we further adopt four dialogue variants for evaluation: (1) ChatGPT-BLIP2 [24], (2) Human-BLIP2, (3) Flan_Alpaca_XXL-BLIP2 [41] from ChatIR [20] and (4) ChatGPT-BLIP2 from PlugIR [19]. These dialogues are generated through the official construction pipelines of their respective methods, covering diverse linguistic styles and reasoning depths [39, 46]. Note that our model is only trained on VisDial.

Evaluation Metrics. We report Recall@K (R@1, R@5, R@10) as the main metrics, which measure whether the ground-truth image appears in the top-K retrieved results. For experiments evaluating overall performance, we use the 10-round averaged Recall as the evaluation metric. We also provide *Hits* performance in the supplementary material.

Baselines. We compare our approach with the retrieval strategies adopted in several representative frameworks: BLIP (zero-shot) [23], ChatIR [20], PlugIR [19] and ImageScope [33]. For fair comparison, all models adopt the BLIP backbone for feature extraction and text–image alignment. Unlike ChatIR, which optimizes BLIP using a recall-based objective [21, 37], our approach fine-tunes BLIP under a standard contrastive learning paradigm [38].

4.2. Comparison to the state-of-the-art

Quantitative Results on VisDial. Table 1 summarizes the performance on the VisDial dataset, our proposed CAST consistently outperforms the existing methods across all metrics and dialogue rounds with clear margins. At the first dialogue round, our method achieves an R@1 of 43.60%, surpassing ChatIR (42.78%) and BLIP (40.31%). As the dialogue progresses, this advantage steadily expands by the tenth round, CAST reaches 57.56% in R@1, outperforming ChatIR (53.15%) and PlugIR (46.61%). Unlike previous methods that used a static feature space cross-modal similarity computation, our proposed CAST continuously reshapes textual and visual embeddings in response to evolving dialogue semantics, leading to progressively refined alignment between textual intent and visual content. Overall, CAST achieves best performance and exhibits an increasing performance margin with dialogue depth, confirming the effectiveness of dynamic feature transformation in interactive retrieval.

Robustness across Various Dialogue Styles. To further examine the robustness and generalization ability of CAST, we evaluate its performance on multiple dialogue variants generated from the ChatIR [20] and PlugIR [19] methods. These dialogues differ in prompting strategies, linguistic diversity, and reasoning depth, covering a broad range of

Table 1. Performance comparison with the state-of-the-art methods on VisDial. Note that the compared methods all utilize a *static* multi-modal latent space for retrieval, while our proposed method uses *dynamic* latent spaces according to the dialogue context. The results demonstrate the effectiveness of our proposed context-aware dynamic latent space for interactive text-to-image retrieval.

Metric	Method	1	2	3	4	5	6	7	8	9	10	Avg.
Recall@1	BLIP	40.31	38.86	36.68	34.59	33.28	32.22	32.56	30.91	30.23	30.09	33.97
	ChatIR	42.78	43.85	45.69	47.14	48.50	49.22	50.39	51.26	52.37	53.15	48.44 (+14.47)
	PlugIR	40.21	42.10	43.02	43.85	45.78	46.95	46.90	47.53	46.22	46.61	44.92 (+10.95)
	ImageScope	32.17	33.58	33.87	32.46	31.69	34.35	31.20	33.24	29.26	30.04	32.19 (-1.78)
	Ours	**43.60**	**45.59**	**48.35**	**49.76**	**51.89**	**53.39**	**55.43**	**56.10**	**56.78**	**57.56**	**51.85 (+17.88)**
Recall@5	BLIP	61.82	59.54	57.95	57.03	55.43	53.59	52.81	52.33	51.41	50.39	55.23
	ChatIR	65.41	66.86	67.64	69.53	71.27	72.53	72.92	73.93	74.13	74.56	70.88 (+15.65)
	PlugIR	63.08	64.83	66.09	67.15	67.78	68.51	68.07	68.60	67.88	68.12	67.01 (+11.78)
	ImageScope	57.46	58.24	57.03	58.14	55.52	57.21	56.59	56.30	55.28	56.35	56.81 (+1.58)
	Ours	**66.47**	**68.99**	**71.03**	**73.06**	**74.61**	**76.02**	**76.79**	**78.15**	**78.10**	**79.07**	**74.23 (+19.00)**
Recall@10	BLIP	70.64	68.75	66.62	65.12	64.44	63.23	61.58	61.09	60.47	59.40	64.13
	ChatIR	74.81	76.79	78.00	79.60	80.77	81.15	81.59	82.07	81.98	82.22	79.90 (+15.77)
	PlugIR	72.24	73.74	74.52	75.92	75.34	76.45	76.41	75.48	76.02	76.50	75.26 (+11.13)
	ImageScope	68.36	67.97	67.30	69.09	67.64	67.78	67.93	68.51	66.09	67.88	67.86 (+3.73)
	Ours	**75.19**	**77.66**	**79.51**	**81.93**	**82.56**	**83.43**	**84.35**	**84.98**	**84.84**	**86.00**	**82.05 (+17.92)**

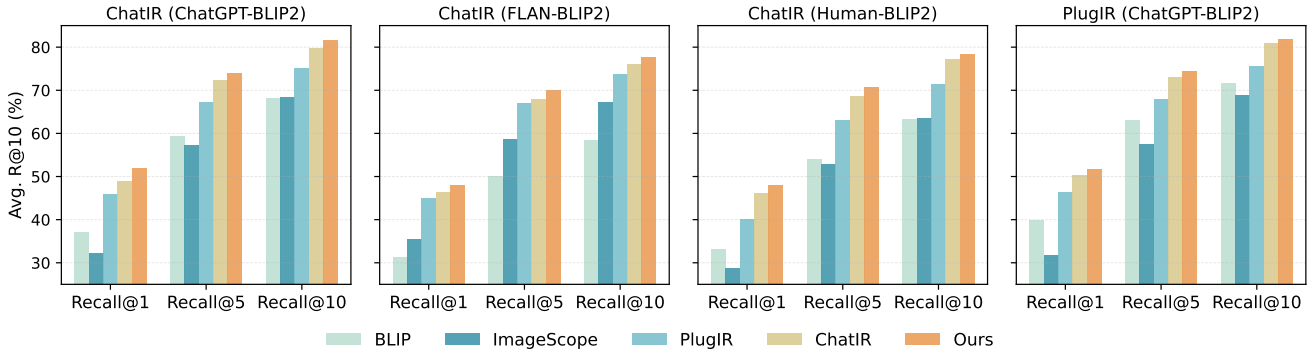


Figure 3. Performance comparison using different dialogue sources. Our proposed method consistently outperforms previous works, demonstrating its strong adaptability to various dialogue styles.

dialogue generation styles from natural conversational reformulations to instruction-driven interactions. As summarized in Figure 3, we report the average performance across all dialogue rounds. CAST consistently achieves the best performance on three different dialogue sources. This demonstrates that our framework generalizes effectively to diverse dialogue formulations and robustly adapts its feature modulation mechanism to different contextual distributions, further validating the semantic adaptability of CAST in interactive retrieval.

4.3. Integration with Existing Frameworks

Recall that our proposed context-aware space regulator can be used as a plug-and-play module for existing vision-language models. In this experiment, we integrate it into the classical vision-language model BLIP [23] and two representative interactive retrieval models PlugIR [19] and ChatIR [20]. For PlugIR, the retrieved images at each round directly influence the subsequent question generation, so we regenerate the questions accordingly. As shown in Figure 4, the performance of BLIP is the worst, as it is not specifically designed for I-TIR. Fortunately, with the help of our context-aware space regulator, it achieves comparable performance to the I-TIR model PlugIR. The results demonstrate the potential of our proposed module for adapting the traditional vision-language model to interactive retrieval. Additionally, replacing the static latent spaces with the dynamic spaces produced by our context-aware space regulator, both PlugIR and ChatIR achieve further performance gain. This not only demonstrates the effectiveness of context-aware dynamic spaces for I-TIR, but also highlights the plug-and-play capability of our module.

4.4. Ablation Studies

Overall contribution of CLP and CGM. Table 2 summarizes the overall impact of the proposed components.

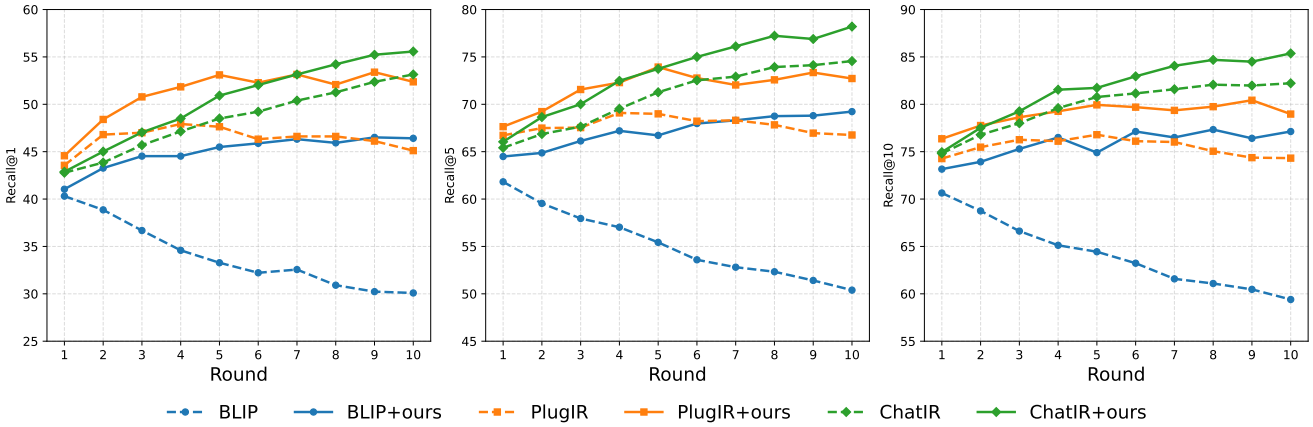


Figure 4. Performance of different methods without or with our proposed context-aware space regulator on the VisDial dataset. Using our proposed module consistently improves performance, showing it can be served as a plug-and-play module for enhancing I-TIR.

Table 2. Ablation study on the VisDial dataset. All results report the 10-round averaged Recall performance.

CLP	CGM	R@1	R@5	R@10
×	×	48.44	71.71	79.89
✓	×	51.04	73.45	80.90
✓	✓	**51.85**	**74.23**	**82.05**

Table 3. Performance comparison between our proposed context-aware space regulator and the alternative projection structures.

Projection Design	Avg. R@10	Δ vs. Ours
Ours	**82.05**	—
Single context-aware W	80.57	-1.48
Single context-agnostic W	79.18	-2.87
Multiple context-agnostic W_t	75.52	-6.53

Introducing the CLP significantly improves retrieval accuracy across all metrics. CLP introduces context-conditioned space transformation that transforms the embedding manifold along meaningful semantic directions, enabling finer adaptation to evolving dialogue semantics. Adding the CGM brings further improvements by adaptively controlling the transformation strength according to contextual divergence between dialogue rounds. This adaptive mechanism ensures smooth and stable feature evolution while maintaining sensitivity to major semantic shifts. Together, CLP and CGM facilitate a progressive and coherent evolution of the feature space, validating the effectiveness of lightweight, context-aware space transformation for multi-round interactive retrieval.

Effect of projection structure in CLP. In order to further verify the effectiveness of our designed context-aware low-rank projector, we compare three space transformation methods: (1) a context-aware projector using a matrix W without matrix decomposition; (2) a context-agnostic projector W shared for each dialogue round; and (3) multiple context-agnostic projectors W_t separately for each dialogue round. Illustrated in Table 3, replacing the context-aware low-rank projection with a non-decomposed context-aware matrix W leads to a significant performance drop, confirming the advantages of matrix decomposition. Additionally, using a shared context-agnostic projector or multiple independent projectors for each dialogue round results in further performance degradation, highlighting the importance of context awareness in dynamic feature transformation. These results demonstrate that the low-rank, context-dependent design of CLP offers a more effective and efficient way to transform the feature space.

Table 4. Performance comparison between our context-guided modulator and the context-agnostic counterpart using a learnable scalar for space transformation magnitude computation.

Gating Mechanism	Avg. R@10	Δ vs. Ours
Context-Guided Modulator	**82.05**	—
Learnable scalar	81.51	-0.54

Effectiveness of the Context-Aware Modulator (CGM). As shown in Table 4, our context-guided modulator is better than the context-agnostic counterpart using a learnable scalar for space transformation magnitude computation. It is worth noting that the context-guided modulator adaptively learns the transformation magnitude based on the input dialogue context, while the context-agnostic one uses a fixed scalar regardless of input context. This result highlights the importance of considering the input context for transformation magnitude computation.

5. Analysis

Visualization Analysis. Figure 5 illustrates how the model dynamically transforms the latent feature space un-

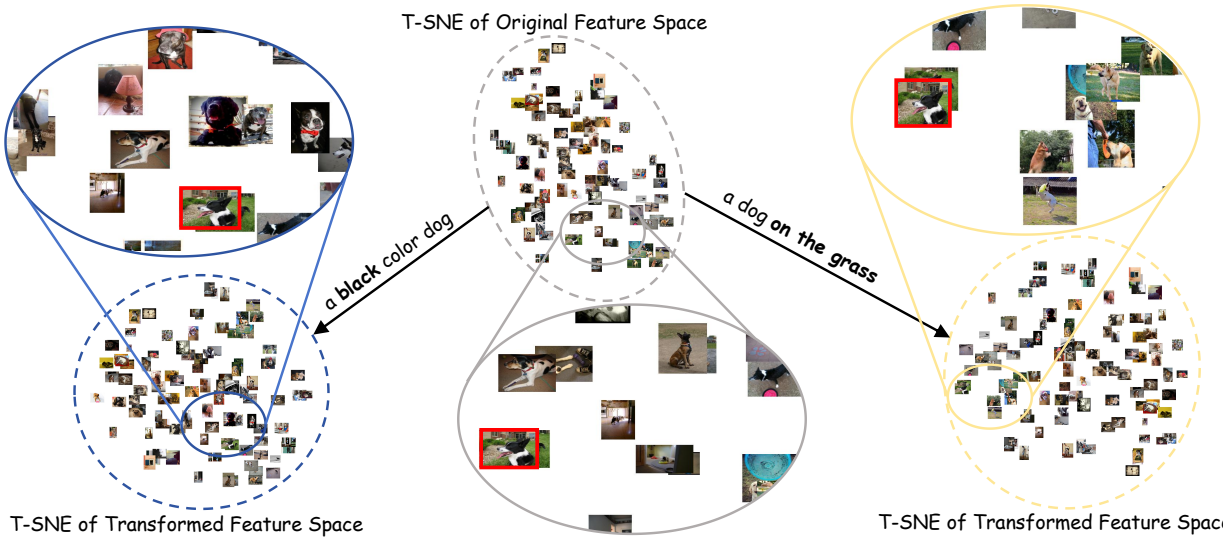


Figure 5. T-SNE visualization shows the 100 images about the topic of *dog*. The red bounding box marks a sample image, with its local region zoomed-in to illustrate space transformation. When the input dialogue context is "a black dog", the feature space seems to be transformed into a new space of emphasizing color distinctions. As a result, images with the "black" attribute are brought closer together in the transformed space. Similarly, when the dialogue context is "a dog on the grass", the space adjusts to prioritize scene, bringing images in grassy environments closer together. This demonstrates how the input context dynamically transforms the representation of the images, highlighting the model's ability to adapt its feature space according to context.

der different contextual conditions. We first select the top 100 images about "dog" and visualize their embeddings using T-SNE [34]. Then, we take the sample highlighted by the red bounding box as an example and zoom into its local neighborhood to observe the semantic transitions more clearly. When the input context condition is "a black dog", the model emphasizes the color attribute in the new transformed feature space, making black dogs more semantically related and clustered together. As the condition further evolves to "a dog on the grass", the model highlights the scene attribute "on the grass" causing dogs with grassy backgrounds to become more concentrated in the embedding space. This demonstrates how the input context dynamically transforms the representation of the images, highlighting the model's ability to adapt its feature space according to context.

Efficiency Analysis. Despite incorporating dynamic space transformation, CAST remains extremely efficient. Its main computational cost lies in the feature transformation step. When using a rank of r=8, CAST can transform the features of 50,000 images in approximately 0.002 seconds without any acceleration, demonstrating its lightweight and scalable design. This efficiency stems from the low-rank formulation, which decomposes the transformation into two low-rank matrix operations that can be executed in parallel across samples. Such speed makes CAST not only suitable for interactive settings but also practical for large-scale retrieval systems.

6. Conclusion

In this paper, we have introduced CAST, a lightweight framework for dynamic latent-space transformation in I-TIR. Unlike prior methods that operate in a static multimodal space, CAST enables the feature manifold to transform adaptively with dialogue context, achieving finer semantic alignment between text and image representations. At its core lies the CASR, which combines the CLP for context-conditioned projection and the CGM for adaptive control of modulation strength. Together, these modules balance semantic flexibility and feature stability throughout multi-round interactions. Extensive experiments show that CAST consistently outperforms state-of-the-art methods across diverse dialogue sources. With minimal computational overhead and plug-and-play compatibility, CAST offers a scalable solution for interactive retrieval.

7. Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2023YFC3304902), Pioneer and Leading Goose R&D Program of Zhejiang (No. 2024C01110), National Natural Science Foundation of China (No. 62472385, 62306278, 62402439), Zhejiang Provincial Natural Science Foundation of China (No. LQN25F020012), Fundamental Research Funds for the Provincial Universities of Zhejiang (No. FR2402ZD) and Tongxiang Institute of Artificial General Intelligence (No. TAGI2-B-2024-0004).

References

[1] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020. 4

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 4

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 4

[4] Rui Cai, Jianfeng Dong, Tianxiang Liang, Yonghui Liang, Yabing Wang, Xun Yang, Xun Wang, and Meng Wang. Cross-lingual cross-modal retrieval with noise-robust fine-tuning. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):5860–5873, 2024. 1

[5] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*, 2022. 1, 2

[6] Xianke Chen, Daizong Liu, Xun Yang, Xirong Li, Jianfeng Dong, Meng Wang, and Xun Wang. Prvr: Partially relevant video retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[7] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017. 5

[8] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 20(12):3377–3388, 2018. 1

[9] Jianfeng Dong, Xirong Li, and Duanqing Xu. Cross-media similarity evaluation for web image retrieval in the wild. *IEEE Transactions on Multimedia*, 20(9):2371–2384, 2018. 2

[10] Jianfeng Dong, Zhe Ma, Xiaofeng Mao, Xun Yang, Yuan He, Richang Hong, and Shouling Ji. Fine-grained fashion similarity prediction by attribute-specific embedding learning. *IEEE Transactions on Image Processing*, 30:8410–8425, 2021. 2

[11] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 3

[12] Zhixiao Fu, Xinyuan Chen, Jianfeng Dong, and Shouling Ji. Multi-order adversarial representation learning for composed query image retrieval. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1685–1689. IEEE, 2021. 3

[13] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. *Advances in neural information processing systems*, 31, 2018. 1, 3

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yun-Cheng Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021. 2

[16] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual compositional learning in interactive image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1771–1779, 2021. 2, 3

[17] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Interactive image search with relative attribute feedback. *International Journal of Computer Vision*, 115(2):185–210, 2015. 2

[18] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018. 3

[19] Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. Interactive textto-image retrieval with large language models: A plugand-play approach. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 11–16, 2024. 1, 2, 4, 5, 6

[20] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Chatting makes perfect: Chat-based image retrieval. *Advances in Neural Information Processing Systems*, 36: 61437–61449, 2023. 1, 2, 5, 6

[21] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and quality assessment for composed image retrieval. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2991–2999, 2024. 1, 2, 5

[22] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 5

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2, 5, 6

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 5

[25] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European conference on computer vision*, pages 121–137. Springer, 2020. 1, 2

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In

European conference on computer vision, pages 740–755. Springer, 2014. 5

[27] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078, 2020. 2

[28] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. Adaptive proposal generation network for temporal sentence localization in videos. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9292–9301, 2021.

[29] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11235–11244, 2021.

[30] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. Memory-guided semantic learning network for temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1665–1673, 2022.

[31] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2125–2134, 2021. 1, 2

[32] Zijun Long, Kangheng Liang, Gerardo Aragon Camarasa, Richard Mccreadie, and Paul Henderson. Diffusion augmented retrieval: A training-free approach to interactive text-to-image retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 823–832, 2025. 2, 5

[33] Pengfei Luo, Jingbo Zhou, Tong Xu, Yuan Xia, Linli Xu, and Enhong Chen. Imagescope: Unifying language-guided image retrieval via large multimodal model collective reasoning. In *Proceedings of the ACM on Web Conference 2025*, pages 1666–1682, 2025. 2, 5

[34] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008. 8

[35] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5):1, 2023. 4

[36] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 3

[37] Yash Patel, Giorgos Tolias, and Jiří Matas. Recall@ k surrogate loss with large batches and similarity mixup. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7502–7511, 2022. 5

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 2, 5

[39] Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. Image chat: Engaging grounded conversations. *arXiv preprint arXiv:1811.00945*, 2018. 5

[40] Nick Stracke, Stefan Andreas Baumann, Joshua Susskind, Miguel Angel Bautista, and Björn Ommer. Ctrloralter: conditional loradapter for efficient 0-shot control and altering of t2i models. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024. 4

[41] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023. 5

[42] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022. 3

[43] Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6862–6872, 2023. 3

[44] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448, 2019. 1, 2

[45] Yabing Wang, Shuhui Wang, Hao Luo, Jianfeng Dong, Fan Wang, Meng Han, Xun Wang, and Meng Wang. Dual-view curricular optimal transport for cross-lingual cross-modal retrieval. *IEEE Transactions on Image Processing*, 33:1522–1533, 2024. 1

[46] 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 *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021. 3, 5

[47] Eric Xing, Abby Stylianou, Robert Pless, and Nathan Jacobs. Quari: Query adaptive retrieval improvement. *arXiv preprint arXiv:2505.21647*, 2025. 2, 4

[48] Zhiyi Yao, Xiang Zhang, Lei Hou, Xu Sun, Yue Huang, Wei Lin, and Lei Zhang. Filip: Fine-grained interactive language–image pre-training. In *International Conference on Learning Representations*, 2022. 2