

# *AlbumFill*: Album-Guided Reasoning and Retrieval for Personalized Image Completion

Yu-Ju Tsai<sup>1</sup>   Brian Price<sup>2</sup>   Qing Liu<sup>2</sup>   Luis Figueroa<sup>2</sup>  
Daniil Pakhomov<sup>2</sup>   Zhihong Ding<sup>2</sup>   Scott Cohen<sup>2</sup>   Ming-Hsuan Yang<sup>1</sup>

<sup>1</sup> University of California, Merced

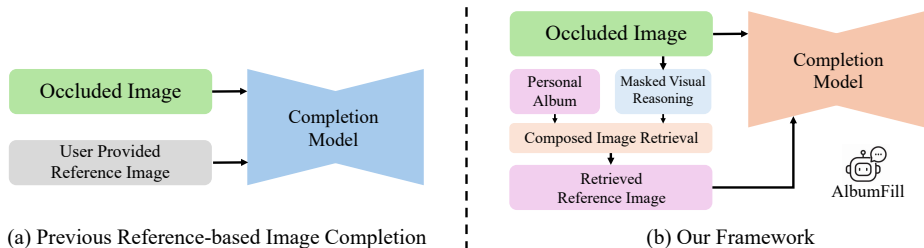
<sup>2</sup> Adobe Research

**Abstract.** Personalized image completion aims to restore occluded regions in personal photos while preserving identity and appearance. Existing methods either rely on generic inpainting models that often fail to maintain identity consistency, or assume that suitable reference images are explicitly provided. In practice, suitable references are often not explicitly provided, requiring the system to search for identity-consistent images within personal photo collections. We present *AlbumFill*, a training-free framework that retrieves identity-consistent references from personal albums for personalized completion. Given an occluded image and a personal album, a vision-language model infers missing semantic cues to guide composed image retrieval, and the retrieved references are used by reference-based completion models. To facilitate this task, we introduce a dataset containing 54K human-centric samples with associated album images. Experiments across multiple baselines demonstrate the difficulty of personalized completion and highlight the importance of identity-consistent reference retrieval.

## 1 Introduction

With the rapid growth of personal photography, individuals now maintain extensive collections of personal images that capture their appearance, clothing, and daily environments. These personal albums [24, 48] inherently contain rich identity-consistent cues, including facial structure, hairstyle, clothing style, and habitual backgrounds. Such collections provide valuable visual priors that can support personalized image editing [19] and completion [50]. In particular, when restoring occluded regions in personal photographs, album images often contain complementary visual evidence that reflects the same identity across different scenes and viewpoints.

Unlike generic web images, personal albums exhibit strong identity consistency across images. A single individual may appear repeatedly across many photographs with similar facial characteristics, clothing preferences, and social contexts. These recurring visual patterns provide valuable cues that can support identity-aware image editing and restoration. However, effectively exploiting such cues remains challenging, as the most relevant reference image may be located



**Fig. 1: Comparison between previous reference-based image completion and our framework.** (a) Previous methods assume that a suitable reference image is provided by the user, limiting their applicability when references are unavailable. (b) Our framework automatically retrieves identity-consistent references from a personal photo album. Given a masked input image, a reasoning module first infers the missing semantic cues, which are used to perform composed image retrieval from the album. The retrieved reference image is then used by a completion model to synthesize the missing region.

anywhere within a large personal album and must be automatically identified from many candidate images.

When editing or completing a personal photo, users typically provide a mask indicating the region to be filled. Conventional inpainting and diffusion-based approaches [25, 32, 42] rely primarily on local context or category-level priors, which may generate visually plausible textures but often fail to preserve identity-specific details. Reference-based completion methods [7, 50, 57, 63] offer a promising alternative by conditioning the generation process on an additional reference image that provides visual evidence of the missing content. However, these approaches typically assume that an appropriate reference image is explicitly provided by the user. In real-world personal photo collections, such references are rarely available beforehand and must instead be discovered from a large album containing numerous candidate images.

Selecting a suitable reference image is itself a challenging problem. The retrieved reference must not only match the semantic content of the missing region but also preserve the identity and appearance of the same individual. In large personal albums containing hundreds or thousands of images, many candidates may share similar visual contexts while differing in identity, pose, or clothing. As a result, naive similarity-based retrieval often fails to identify the most appropriate reference, motivating the need for reasoning-guided retrieval strategies.

Vision-language models (VLMs) [3, 10, 17, 27, 47, 51, 52, 59] provide a natural tool for addressing this challenge. By interpreting the visible context of the occluded image, VLMs can infer the likely semantics of the missing region and generate textual descriptions that capture the expected content. Such descriptions can serve as structured retrieval cues, enabling the system to search for identity-consistent reference images within the album that match both the inferred semantics and the individual’s appearance.

Building upon this insight, we present *AlbumFill*, a framework for personalized image completion that performs reasoning-guided reference retrieval from

personal albums. Given an occluded image and a personal album, our approach first employs a vision-language model to reason about the masked region and generate semantic cues describing the missing content. These cues are then used to perform composed image retrieval (CIR) [18, 22, 28, 36, 41, 44] over the album to identify reference images that are consistent with both the inferred semantics and the individual’s identity. The retrieved reference image is subsequently combined with the masked input and passed to a reference-based completion model to synthesize the final result.

To support this task, we further introduce a new benchmark for personalized album-based completion. Our dataset contains 54,000 human-centric image samples paired with contextual captions and album references, enabling systematic evaluation of both reference retrieval and completion quality. Unlike conventional datasets that focus on generic reconstruction, our benchmark emphasizes identity-aware completion in personal photo collections.

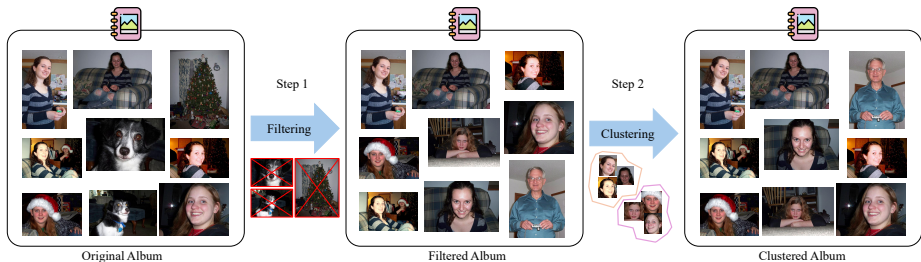
Our work differs from prior prompt-based inpainting methods, which rely on generative priors to hallucinate missing content, as well as recent multimodal editing systems such as BAGEL [14] and Gemini 3 Pro [13] that generate edits directly from textual instructions. It also differs from existing reference-based completion approaches that assume a reference image is provided. Instead, we focus on the realistic setting where identity-consistent reference images must be automatically retrieved from personal albums before performing completion.

Our contributions are summarized as follows:

- We propose *AlbumFill*, a framework that leverages vision-language reasoning to guide reference retrieval and enable identity-consistent image completion.
- We introduce a new benchmark for album-based personalized completion, containing 54K human-centric samples for evaluating retrieval and completion quality.
- We demonstrate through experiments that identity-consistent references are critical for personalized image completion, motivating the need for reference retrieval from personal photo collections.

## 2 Related Work

**Vision Language Model Reasoning.** Recent advances in multimodal large language models (MLLMs) [3, 10, 17, 27, 47, 51, 52, 59] highlight the role of instruction tuning [12] and chain-of-thought (CoT) [55] strategies in enabling robust visual reasoning. LLaVA [29, 30] aligns vision encoders with LLMs using curated multimodal instructions, yielding strong zero-shot generalization across tasks. Complementary CoT approaches extend reasoning supervision to the visual domain: Multimodal-CoT [62] separates rationale generation from answering, while Visual CoT [37] provides large-scale data with region-level and step-by-step annotations to teach interpretable reasoning. Compositional CoT [33] further leverages scene-graph-driven prompting to improve structured compositional inference. MM-REACT [58] couples LLM reasoning with specialized vision tools for complex queries. Kosmos-2 [34] integrates explicit localization tokens to connect



**Fig. 2: Data generation pipeline for constructing our Album Dataset.** As describe in Sec. 3, we start from the original CUFED album [53] (left), we first perform human detection and remove images without sufficiently large or clear human presence (Step 1: Filtering). We then apply identity clustering using InsightFace to group images by subject and select the dominant identity within each album (Step 2: Clustering). The resulting clustered album (right) contains only images of the target person, forming a coherent personal-level collection for identity-consistent retrieval and completion.

language to objects, strengthening referential reasoning. Finally, commercial systems [1, 11, 31, 46] demonstrate scalable, real-time multimodal reasoning capabilities, underscoring the trend toward unified, instruction-following MLLMs.

**Zero-shot Composed Image Retrieval.** Zero-shot Composed Image Retrieval (ZS-CIR) [18, 22, 28, 36, 41, 44] retrieves target images consistent with a reference image and user-provided textual modifications, without requiring extensive labeled triplets. ZS-CIR approaches leverage pre-trained vision-language models such as CLIP [35] to combine visual and textual features within a shared embedding space. Recent textual inversion methods [4, 18] map reference image features into pseudo-word tokens or masked textual representations for query generation. Training-free frameworks [26, 40] further eliminate fine-tuning by using large language models to infer compositional queries directly. OSrCIR [45] introduces a one-stage paradigm that employs multimodal large language models (MLLMs) for joint visual-textual reasoning, preserving full image context and achieving superior efficiency and intent alignment.

**Reference-based Inpainting.** Reference-based image inpainting leverages external visual cues to produce semantically consistent completions. Early approaches focus on geometric alignment between source and target images. For example, TransFill [63] aligns views through depth-aware multi-homography to preserve structural consistency. Recent methods increasingly adopt diffusion-based generation to incorporate reference information. Paint-by-Example [57], ObjectStitch [38], and IMPRINT [39] integrate exemplar features through patch conditioning, content adaptors, or staged identity learning to better preserve appearance and semantics. Other approaches extend reference-guided editing to more flexible settings. AnyDoor [8] enables zero-shot reference-guided synthesis, while LeftRefill [5] combines reference and target views as a unified input. MimicBrush [7] performs localized reference editing using dual diffusion U-Nets trained in a self-supervised manner. CompleteMe [50] further improves identity preservation through a region-focused attention mechanism. More recently, unified multimodal frameworks aim to support broader image editing

tasks. UniReal [9] formulates image editing as discontinuous video generation to capture real-world dynamics. MIGE [49] unifies subject-driven generation and instruction-based editing through a shared multimodal encoder, and EditMaster [60] combines textual and visual controls for multimodal editing. Large multimodal systems such as BAGEL [14] and Gemini 3 Pro [13] further support instruction-following image editing with visual references.

However, these approaches generally assume that appropriate reference images are explicitly provided. In realistic personal photo collections, references must instead be retrieved from large albums, requiring reasoning over incomplete visual inputs to identify identity-consistent references.

### 3 AlbumFill Benchmark

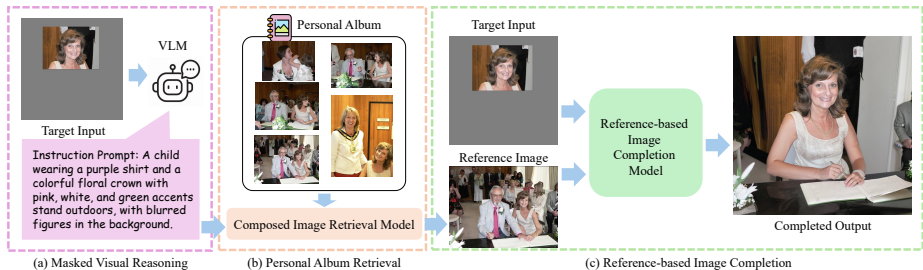
To evaluate our framework in realistic personal-photo scenarios, we construct the **AlbumFill Benchmark**, a dataset designed to reflect the characteristics of real-world personal photo collections. We build our dataset upon CUFED [53], which was originally designed for studying event-specific image importance and is constructed from the YFCC100M dataset [48]. CUFED contains 94,797 images organized into 1,883 folders, with each folder containing approximately 30 to 100 images. This album-like structure naturally reflects groups of photos captured within the same events or contexts, making it well suited for studying album-level visual relationships.

Since personal photo collections are typically human-centric and identity consistency plays a key role in personalized completion, we focus on images containing visible people and ensure that the masked regions correspond to identity-relevant areas.

**Step 1: Human Detection and Filtering.** We therefore detect people using YOLOv8 [23]. Bounding boxes smaller than 0.15 of the image height or width are discarded because they correspond to distant or insignificant figures. Images containing more than 20 people are removed to avoid overly crowded scenes.

**Step 2: Identity Clustering and Album Formation.** Next, we apply InsightFace [20] to cluster identities within each album. For each album, we select the dominant identity, defined as the person who appears in the largest number of images, as the target subject. All images that contain this identity are grouped to form one personal album that spans diverse poses, outfits, and backgrounds. Each image within an album can serve as a potential input image, while the remaining images act as candidate reference images for retrieval. Because multiple visually similar images of the same individual appear in each album under different poses, scenes, and viewpoints, identifying the most appropriate reference image becomes a non-trivial retrieval problem.

**Step 3: Mask Generation.** To construct completion tasks, we generate random occlusion masks on the selected input images. Masks are placed on human-centric regions to simulate realistic occlusions commonly observed in personal photographs, such as missing faces, bodies, or accessories. The masked image serves as the input to the completion model, while the original image is used



**Fig. 3: AlbumFill system overview.** (a) Given a masked target image, a Vision-Language Model (VLM) performs *masked visual reasoning* to generate a textual hypothesis describing the likely content behind the masked region. (b) The reasoning text and visible context are used to compose a multimodal query that retrieves the most semantically aligned and identity-consistent reference image from the user’s personal album. (c) A reference-based image completion model synthesizes the final output by integrating the masked target and the retrieved reference, producing an identity-faithful and contextually coherent restoration.

as the ground-truth reference for evaluation. After filtering and clustering, the final dataset contains **53,294 images grouped into 1,727 personal albums**. Each album corresponds to a single dominant identity with an average of **30.86 photos**. The dataset exhibits substantial diversity in pose, clothing, indoor and outdoor scenes, and background complexity, creating challenging conditions for reference-based completion and retrieval. Overall, the AlbumFill Benchmark provides a realistic testbed for evaluating reasoning-guided retrieval and identity-consistent image completion in personal photo collections. Additional dataset construction details are provided in the supplementary material.

## 4 Method

### 4.1 Overview

We propose *AlbumFill*, a training-free framework that combines vision-language reasoning, album-based retrieval, and reference-guided image completion to enable personalized restoration from personal photo collections. Instead of fine-tuning or training new models, *AlbumFill* connects pre-trained large-scale vision-language and generative components within a modular pipeline. Given an input image  $I_m$  with a missing region and a user’s personal album  $\mathcal{A} = \{I_1, I_2, \dots, I_N\}$ , the system operates in three sequential stages: (1) **Masked visual reasoning**, where a vision-language model analyzes the incomplete image and infers the semantic content of the occluded region based on the visible context; (2) **Album retrieval**, which searches the personal album for reference images that are consistent with the reasoning output and the visible identity cues; and (3) **Reference-based completion**, which synthesizes the final image using both the masked input and the retrieved reference image. This modular design allows flexible substitution of different vision-language models or completion models

while requiring no additional task-specific training. The overall pipeline is illustrated in Fig. 3.

## 4.2 Masked Visual Reasoning

Existing vision-language models (VLMs) typically assume access to complete visual inputs when performing semantic reasoning. However, image editing and inpainting tasks involve intentionally occluded regions, requiring the model to infer plausible content behind the masked area based on the available visual context. To address this setting, we perform **masked visual reasoning**, where a VLM predicts the semantic content of the missing region using only the visible portion of the image together with its learned world knowledge.

Given a masked image  $I_m$  and its binary mask  $M$ , the visible region is obtained as

$$I_m^{vis} = I_m \odot (1 - M).$$

We then prompt a pre-trained VLM with an instruction such as: “Describe what likely exists in the masked region based on the visible context.” The VLM produces a reasoning hypothesis

$$T_r = \text{VLM}(I_m^{vis}, \text{instruction}),$$

where  $T_r$  is a natural-language description containing semantic cues such as object category, pose, appearance, spatial relations, or activity. The full prompt templates used for reasoning are provided in the supplementary material. This step reframes the pixel-level completion problem as a *conceptual prediction task*, allowing the model to reason beyond directly observable content. Unlike conventional inpainting methods that extrapolate local textures, masked visual reasoning produces global structural priors. For example, the model may infer descriptions such as “the child’s missing hand holding a toy” or “the upper part of the head wearing a floral crown.” These semantic cues guide the subsequent album retrieval stage by providing structured hints about the missing content.

## 4.3 Album Retrieval

Given the reasoning hypothesis  $T_r$ , our goal is to retrieve album images that align with the inferred semantics while preserving the user’s identity. Because each album in our benchmark contains images of the same dominant identity (Sec. 3), the retrieval process focuses on selecting references that best match the inferred content, pose, and scene context.

**Composed Query Embedding.** To integrate visual context with reasoning cues, we construct a multimodal query embedding that combines the visible region of the masked image with the reasoning text:

$$q = f_{\text{CLIP}}(I_m^{vis}, T_r),$$

where  $f_{\text{CLIP}}$  denotes a CLIP-style joint vision-language encoder. This composed embedding captures visual appearance cues from the visible image region (e.g.,

clothing color or hairstyle), contextual information such as indoor or outdoor environments, and semantic predictions produced during the reasoning stage. This formulation follows the standard composed image retrieval paradigm.

**Album-Level Retrieval.** Each album image  $I_i \in \mathcal{A}$  is encoded as

$$v_i = f_{\text{CLIP}}(I_i).$$

We compute the similarity between the query and each candidate image via cosine similarity:

$$s_i = \cos(q, v_i).$$

The top- $k$  most similar images are selected as

$$\mathcal{R} = \text{Top-}k(\{s_i\}_{i=1}^N).$$

In our default pipeline, the highest-ranked image  $I_{\text{ref}}$  is used as the reference for the completion model. However, the full candidate set  $\mathcal{R}$  can also be presented to the user as alternative reference options, allowing manual selection when multiple plausible references exist. This design provides both an automatic retrieval mechanism and an interactive reference selection interface.

By retrieving references from the personal album, this stage provides identity-consistent visual evidence for completion rather than relying solely on generative priors. The reasoning cues further narrow the retrieval space by describing the likely semantics of the occluded region, improving alignment between the retrieved reference and the missing content.

#### 4.4 Reference-based Completion

Given the masked target image  $I_m$  and the retrieved reference image  $I_{\text{ref}}$ , the final stage synthesizes the completed output using a reference-based diffusion inpainting model. We denote the completion module as  $\mathcal{G}$ , which produces the restored image as

$$I_{\text{out}} = \mathcal{G}(I_m, I_{\text{ref}}).$$

**Reference-Guided Conditioning.** The inpainting model incorporates visual features from the reference image to guide the synthesis of the missing region. These reference cues provide fine-grained appearance information such as texture and color, structural information such as pose or silhouette, and scene-level attributes such as lighting and tone. Conditioning on both the masked input and the retrieved reference allows the model to generate content that is consistent with the identity and appearance observed in the album images.

**Consistency Preservation.** To maintain spatial coherence, the completion model leverages both the visible context in  $I_m$  and the appearance cues in  $I_{\text{ref}}$ . This combination encourages structural continuity along the mask boundary while preserving identity-specific characteristics within the synthesized region. As a result, the generated completion aligns with the surrounding scene context and the retrieved reference appearance.

Because the completion module is independent from the reasoning and retrieval stages, different reference-based inpainting models can be used within the same pipeline without additional task-specific training.

## 5 Experiments

### 5.1 Experimental Setups

**Implementation Details.** For masked visual reasoning, we evaluate three state-of-the-art Vision-Language Models (VLMs): Qwen3-VL-8B [56], InternVL3.5-8B [52], and LLaVA-1.6-7B [29]. All models are used in a zero-shot setting without any task-specific fine-tuning. We apply a unified instruction template that asks the model to infer the likely semantic content of the masked region based on the visible context of the image. Each masked input produces a single reasoning hypothesis, which is then used to guide the subsequent retrieval stage. Unless otherwise specified, we use Qwen3-VL-8B as the default reasoning model due to its stronger semantic grounding performance.

For the Composed Image Retrieval (CIR) module, we benchmark three representative zero-shot methods: iSEARLE [2], CIReVL [26], and LinCIR [18]. To ensure consistency across methods, all retrieval models use the same CLIP [35] backbone, ViT-L/14 [15], for encoding both composed queries and album images. Each album image is embedded and ranked using cosine similarity with respect to the composed query embedding. We retrieve the top- $k$  candidates ( $k = 5$  unless otherwise specified) and pass the highest-ranked reference image to the completion module. For reference-based completion, we evaluate three models: MimicBrush [7], CompleteMe [50], and UniReal [9]. Each model is used with the default inference configurations provided by the official implementations.

All experiments are conducted on machines equipped with NVIDIA A100 GPUs. After benchmarking different components, our default pipeline uses Qwen3-VL-8B [56] for masked visual reasoning, LinCIR [18] for composed image retrieval, and UniReal [9] for reference-guided completion. Unless otherwise specified, all reported results follow this configuration.

**Evaluation Metrics.** We evaluate our framework along two key aspects of the AlbumFill pipeline: (1) the retrieval accuracy in the album retrieval stage, and (2) the visual fidelity and perceptual consistency of the completed images.

**Album Retrieval.** For composed image retrieval (CIR), we follow standard evaluation protocols and report Recall@K and mean Average Precision at K (mAP@K) for  $K \in \{1, 5, 10, 25, 50\}$ . Recall@K measures whether relevant reference images appear within the top- $K$  retrieved results, while mAP@K evaluates the ranking quality among all relevant candidates within the album.

**Reference-based Completion.** To assess the quality of generated completions, we measure both pixel-level fidelity and perceptual similarity. PSNR [21] and SSIM [54] evaluate low-level reconstruction accuracy between the generated output and the ground-truth image. To better capture perceptual and semantic consistency, we also report LPIPS [61], DINO similarity [6], CLIP similarity [35], and DreamSim [16].

### 5.2 VLM-based Reasoning Evaluation

To evaluate the quality of manipulation instructions produced during masked visual reasoning, we employ an automatic evaluation protocol using a strong

**Table 1: Evaluation of instruction quality using a VLM-based evaluator.** Instruction quality is evaluated using Gemini 2.5 Pro [11] as an external VLM evaluator to maintain independence from the reasoning models. We measure four aspects: 1) evidence grounding, 2) structural continuity, 3) retrieval discriminativeness, and 4) instruction format quality. Scores range from 0 to 20 for each dimension and are averaged across all evaluation samples. Qwen3-VL [56] achieves the highest scores across all dimensions, indicating stronger grounding and more discriminative instructions, and is therefore selected as the reasoning module in our framework.

Model	Evidence Grounding	Structural Continuity	Retrieval Discriminativeness	Instruction Format Quality
LLaVA-1.6-7B [29]	11.00	11.53	8.90	18.05
InternVL3.5-8B [52]	12.01	12.68	10.56	17.52
Qwen3-VL-8B [56]	<b>15.23</b>	<b>14.88</b>	<b>13.75</b>	<b>18.66</b>

Vision-Language Model (VLM) as an external evaluator. Specifically, we use Gemini 2.5 Pro [11]. The evaluator takes the masked image and the generated instruction as input and assesses whether the instruction provides useful cues for retrieving a correct reference image for completion. Unlike traditional caption evaluation, the goal is not linguistic similarity to ground truth but whether the instruction offers grounded and discriminative information for the retrieval stage. To ensure independent assessment, the evaluation model is different from the reasoning models, avoiding potential bias from using the same model family for both generation and evaluation.

The evaluator scores each instruction along four dimensions: *Evidence Grounding*, *Structural Continuity*, *Retrieval Discriminativeness*, and *Instruction Format Quality*, each ranging from 0 to 20. These metrics measure whether the instruction is supported by visible evidence, specifies structural continuation across the mask, provides discriminative retrieval cues, and follows a concise manipulation-instruction format. The full evaluation prompt and rubric are provided in the supplementary material.

Table 1 reports the results of different reasoning models. Qwen3-VL-8B achieves the highest scores across all dimensions, producing more grounded and discriminative instructions that better support the downstream retrieval stage.

### 5.3 Quantitative Comparison on CIR

We evaluate three composed image retrieval (CIR) methods, iSEARLE, CIReVL, and LinCIR, on a subset of 1,416 albums that contain sufficient valid query-reference pairs, while the full dataset contains 1,727 albums. For each album, a masked query image is paired with a manipulation instruction generated by Qwen3-VL, which describes the expected content behind the masked region. This instruction serves as the language component for all CIR baselines. All methods use CLIP ViT-L/14 encoder for fair comparison.

Table 2 reports Recall@k and mAP@k for  $k \in \{1, 5, 10, 25, 50\}$ . LinCIR achieves the best performance across all evaluation settings. At low retrieval depths ( $k = 1, 5, 10$ ), LinCIR obtains the highest Recall@k and mAP@k values. In particular, Recall@1 increases from 6.44 (iSEARLE) and 6.88 (CIReVL)

**Table 2: Quantitative comparison of composed image retrieval (CIR) performance on our personal album benchmark.** We evaluate three representative CIR models: iSEARLE [2], CIReVL [26], and LinCIR [18], under the masked-query setting, using Recall@K and mAP@K for  $K = \{1, 5, 10, 25, 50\}$ . Overall, LinCIR achieves the highest retrieval accuracy across all metrics, particularly at low  $K$ , indicating stronger fine-grained ranking ability.

Metric	Recall@k				mAP@k			
	k=1	k=5	k=10	k=25	k=5	k=10	k=25	k=50
iSEARLE [2]	6.44	28.42	46.70	80.59	21.61	31.73	48.55	56.19
CIReVL [26]	6.88	29.17	47.70	80.95	22.50	32.69	49.69	57.17
LinCIR [18]	<b>7.00</b>	<b>29.63</b>	<b>48.13</b>	<b>81.21</b>	<b>22.83</b>	<b>33.31</b>	<b>50.21</b>	<b>57.69</b>

**Table 3: Quantitative comparison on the AlbumFill benchmark (Sec. 3).** Methods are grouped into three categories: inpainting methods without reference images (top), MLLM-based methods (middle), and reference-based image completion methods (bottom). Our framework adopts UniReal [9] as the completion backbone.

Method	CLIP $\uparrow$	DINO $\uparrow$	DreamSim $\downarrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$
BrushNet [25] w/o Prompt	87.29	87.82	0.162	0.233	18.35	0.654
BrushNet [25] with Prompt	88.38	88.19	0.138	0.229	18.81	0.655
BAGEL [14]	77.52	66.29	0.364	0.652	9.47	0.174
Gemini 3 Pro [13]	89.35	88.26	0.147	0.322	15.83	0.511
MimicBrush [7]	91.24	90.55	0.115	0.115	20.11	<b>0.821</b>
CompleteMe [50]	90.70	91.70	0.102	0.212	20.81	0.768
UniReal [9]	<b>94.70</b>	<b>95.71</b>	<b>0.055</b>	<b>0.142</b>	<b>22.60</b>	0.790

to **7.00**. This result suggests that LinCIR forms more discriminative composed representations for masked-query retrieval.

At larger retrieval depths, the three methods show similar recall trends, reaching approximately 80 percent Recall@25. However, LinCIR consistently achieves the highest ranking quality, obtaining the best mAP scores across all evaluated depths, including **50.21** at  $k = 25$  and **57.69** at  $k = 50$ . These results indicate that LinCIR maintains more stable ranking performance when a larger number of candidate references are considered. iSEARLE and CIReVL exhibit comparable performance across most metrics, although both methods remain slightly below LinCIR in terms of top- $k$  retrieval accuracy and ranking quality. Overall, LinCIR provides the most reliable retrieval performance for masked-query CIR on the AlbumFill benchmark.

## 5.4 Quantitative and Qualitative Comparison on Image Completion

We evaluate representative image completion approaches on the AlbumFill benchmark, including reference-based inpainting models and prompt-driven generative systems. Table 3 reports CLIP similarity, DINO similarity, DreamSim distance, LPIPS, PSNR, and SSIM. All methods use the same masked inputs and reference images for fair comparison, as shown in Fig. 4.

Among the evaluated methods, UniReal [9] achieves the best overall performance, obtaining the highest CLIP (94.70) and DINO (95.71) scores, the lowest



**Fig. 4: Visual comparison with different categories of completion methods on our Album Benchmark (Sec. 3).** We compare our method with three types of baselines: prompt-based inpainting (BrushNet [25] with and without prompts), MLLM-based image editing models (BAGEL [14] and Gemini 3 Pro [13]), and reference-based completion methods (MimicBrush [7] and CompleteMe [50]). Prompt-based methods often produce plausible textures but fail to preserve identity. MLLM-based approaches struggle with identity consistency or structural continuity under large occlusions. Reference-based methods rely heavily on the quality of the provided reference image. In contrast, our method retrieves identity-consistent references from personal albums and produces more coherent completions. For more qualitative comparisons, please refer to our supplementary material.

**Table 4: Effect of reasoning prompts on composed image retrieval under different mask ratios.** We evaluate LinCIR with and without reasoning-generated prompts produced by Qwen3-VL. Mask ratios are categorized as small (<20%), medium (20–50%), and large (>50%). Incorporating reasoning prompts consistently improves Recall@k and mAP@k across all mask regimes, with larger gains observed when the visible region becomes more limited.

Metric	Reasoning Prompt	Mask Ratio	Recall@k			mAP@k		
			k=5	k=10	k=25	k=5	k=10	k=25
Method								
LinCIR [18]		small	30.70	48.90	81.46	<b>24.18</b>	34.31	50.77
LinCIR [18]	✓	small	<b>31.08</b>	<b>49.74</b>	<b>82.28</b>	24.02	<b>34.66</b>	<b>51.41</b>
LinCIR [18]		medium	25.91	43.18	77.28	19.82	29.32	46.31
LinCIR [18]	✓	medium	<b>26.80</b>	<b>45.10</b>	<b>78.65</b>	<b>20.62</b>	<b>30.93</b>	<b>48.29</b>
LinCIR [18]		large	20.66	36.99	79.00	14.85	22.00	38.36
LinCIR [18]	✓	large	<b>22.47</b>	<b>39.47</b>	<b>79.63</b>	<b>16.22</b>	<b>24.56</b>	<b>40.96</b>

DreamSim distance (0.055), and the best PSNR (22.60). MimicBrush [7] produces reasonable structural alignment but lower perceptual consistency, while CompleteMe [50] shows comparable reconstruction fidelity.

We also evaluate prompt-driven baselines to examine whether large multi-modal generative models can directly solve masked completion without specialized reference-guided architectures. For BrushNet [25], we test both a default setting and a prompt generated by the masked reasoning stage using Qwen3-VL. Although prompt guidance slightly improves results, the gap with reference-based inpainting models remains large.

We further evaluate BAGEL [14] and Gemini 3 Pro [13] with the same reference images and detailed instructions. Despite the prompts, these models often produce unstable outputs, such as incorrect aspect ratios, returning the reference image, or directly combining the masked and reference images (see Fig. 4), leading to significantly worse scores. The full prompts used for these models are provided in the supplementary material.

Overall, reference-based inpainting models remain substantially more reliable for personalized image completion. Among them, UniReal provides the most consistent performance across perceptual and reconstruction metrics, and is therefore adopted as the completion module in our framework.

## 5.5 Ablation Studies

**Effectiveness of the Reasoning Stage.** We conduct an ablation study to evaluate the impact of the masked visual reasoning stage on composed image retrieval. To isolate its effect, we fix the retrieval model to LinCIR and use Qwen3-VL as the reasoning model, comparing retrieval performance with and without the reasoning-generated prompt. To analyze the effect of occlusion severity, we divide the benchmark by mask ratio: *small* (<20%), *medium* (20–50%), and *large* (>50%).

Table 4 reports Recall@k and mAP@k under these settings. Across all mask ratios, incorporating the reasoning prompt consistently improves retrieval performance. For example, Recall@5 increases from 30.70 to **31.08** for small masks, from 25.91 to **26.80** for medium masks, and from 20.66 to **22.47** for large masks.

**Table 5: Wrong-reference ablation under different mask ratios.** The retrieval stage remains unchanged (Qwen3-VL + LinCIR retrieves from the correct album), while the completion reference is replaced with a random image from a different album (wrong identity). Mask ratios are categorized as small (<20%), medium (20–50%), and large (>50%). Using a wrong-identity reference consistently degrades completion quality, with the largest performance drop observed under heavy occlusion.

Metric	Wrong Reference	Mask Ratio	CLIP ↑	DINO ↑	DreamSim ↓	LPIPS ↓	PSNR ↑	SSIM ↑
UniReal [9]	✓	small	96.01	97.30	0.037	0.114	24.00	0.827
UniReal [9]		small	<b>96.65</b>	<b>97.72</b>	<b>0.030</b>	<b>0.110</b>	<b>24.32</b>	<b>0.829</b>
UniReal [9]	✓	medium	89.02	90.84	0.121	0.202	18.65	0.718
UniReal [9]		medium	<b>91.75</b>	<b>93.09</b>	<b>0.092</b>	<b>0.193</b>	<b>19.24</b>	<b>0.724</b>
UniReal [9]	✓	large	69.89	66.99	0.348	0.389	14.10	0.508
UniReal [9]		large	<b>78.35</b>	<b>75.59</b>	<b>0.263</b>	<b>0.378</b>	<b>14.48</b>	<b>0.513</b>

These results indicate that reasoning-derived instructions provide additional semantic cues that help the CIR model better align masked inputs with relevant album images, with larger gains observed as occlusion increases.

### Ablation on Identity-Consistent References.

We perform a wrong-reference ablation to evaluate the importance of identity-consistent references in the completion stage. To isolate this effect, the retrieval stage remains unchanged: references are still retrieved from the correct album using the Qwen3-VL + LinCIR pipeline, but during completion the retrieved reference is replaced with a random image from a different album. To analyze the impact of occlusion severity, we group samples by mask ratio: *small* (<20%), *medium* (20–50%), and *large* (>50%).

Table 5 reports the results using UniReal as the completion model. Across all mask ratios, replacing the reference with a wrong-identity image consistently degrades completion quality. For example, CLIP similarity drops from **96.65** to 96.01 for small masks, from **91.75** to 89.02 for medium masks, and from **78.35** to 69.89 for large masks. These results show that identity-consistent references become increasingly important as the mask ratio grows, highlighting the role of personalized reference retrieval for reliable album-based image completion.

## 6 Conclusion

We present *AlbumFill*, a reasoning-guided framework for personalized image completion that integrates masked-region reasoning, album-based retrieval, and reference-based completion. The framework decomposes the task into reasoning, retrieval, and completion, enabling the system to infer missing semantics, retrieve identity-consistent references from personal albums, and synthesize coherent results. To support this task, we introduce the *AlbumFill* benchmark, which reflects the diversity and identity consistency of real-world personal photo collections. Experiments show that reasoning-derived instructions improve composed image retrieval under occlusion, while ablation studies demonstrate that identity-consistent references are crucial for successful completion, especially when large regions are missing. Overall, our results highlight the importance of combining semantic reasoning with album-based retrieval for personalized image completion.

## References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) [4](#)
2. Agnolucci, L., Baldrati, A., Del Bimbo, A., Bertini, M.: isearle: Improving textual inversion for zero-shot composed image retrieval. TPAMI (2025) [9](#), [11](#)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) [2](#), [3](#)
4. Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A.: Zero-shot composed image retrieval with textual inversion. In: ICCV (2023) [4](#)
5. Cao, C., Cai, Y., Dong, Q., Wang, Y., Fu, Y.: Leftrefill: Filling right canvas based on left reference through generalized text-to-image diffusion model. In: CVPR (2024) [4](#)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021) [9](#)
7. Chen, X., Feng, Y., Chen, M., Wang, Y., Zhang, S., Liu, Y., Shen, Y., Zhao, H.: Zero-shot image editing with reference imitation. In: NeurIPS (2024) [2](#), [4](#), [9](#), [11](#), [12](#), [13](#), [18](#)
8. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. In: CVPR (2024) [4](#)
9. Chen, X., Zhang, Z., Zhang, H., Zhou, Y., Kim, S.Y., Liu, Q., Li, Y., Zhang, J., Zhao, N., Wang, Y., et al.: Unireal: Universal image generation and editing via learning real-world dynamics. In: CVPR (2025) [5](#), [9](#), [11](#), [14](#), [18](#)
10. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR (2024) [2](#), [3](#)
11. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025) [4](#), [10](#), [8](#)
12. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. In: NeurIPS (2023) [3](#)
13. DeepMind, G.: Gemini 3 pro image model card (November 2025), <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Image-Model-Card.pdf> [3](#), [5](#), [11](#), [12](#), [13](#), [14](#)
14. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025) [3](#), [5](#), [11](#), [12](#), [13](#), [14](#)
15. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) [9](#)
16. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In: NeurIPS (2024) [9](#)
17. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024) [2](#), [3](#)

18. Gu, G., Chun, S., Kim, W., Kang, Y., Yun, S.: Language-only training of zero-shot composed image retrieval. In: CVPR (2024) [3](#), [4](#), [9](#), [11](#), [13](#)
19. Gu, J., Zhao, N., Xiong, W., Liu, Q., Zhang, Z., Zhang, H., Zhang, J., Jung, H., Wang, Y., Wang, X.E.: Swapanything: Enabling arbitrary object swapping in personalized image editing. In: ECCV (2024) [1](#)
20. Guo, J., Deng, J., et al.: Insightface: State-of-the-art 2d and 3d face analysis project. <https://github.com/deepinsight/insightface> (2025) [5](#)
21. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: ICPR (2010) [9](#)
22. Jang, Y.K., Huynh, D., Shah, A., Chen, W.K., Lim, S.N.: Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. In: ECCV (2024) [3](#), [4](#)
23. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics yolov8 (2023) [5](#)
24. Joon Oh, S., Benenson, R., Fritz, M., Schiele, B.: Person recognition in personal photo collections. In: ICCV (2015) [1](#)
25. Ju, X., Liu, X., Wang, X., Bian, Y., Shan, Y., Xu, Q.: Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In: ECCV (2024) [2](#), [11](#), [12](#), [13](#)
26. Karthik, S., Roth, K., Mancini, M., Akata, Z.: Vision-by-language for training-free compositional image retrieval. In: ICLR (2024) [4](#), [9](#), [11](#)
27. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023) [2](#), [3](#)
28. Lin, H., Wen, H., Song, X., Liu, M., Hu, Y., Nie, L.: Fine-grained textual inversion network for zero-shot composed image retrieval. In: ACM SIGIR (2024) [3](#), [4](#)
29. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR (2024) [3](#), [9](#), [10](#)
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023) [3](#)
31. Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al.: Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525 (2024) [4](#)
32. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: CVPR (2022) [2](#)
33. Mitra, C., Huang, B., Darrell, T., Herzig, R.: Compositional chain-of-thought prompting for large multimodal models. In: CVPR (2024) [3](#)
34. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023) [3](#)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) [4](#), [9](#)
36. Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T.: Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In: CVPR (2023) [3](#), [4](#)
37. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In: NeurIPS (2024) [3](#)
38. Song, Y., Zhang, Z., Lin, Z., Cohen, S., Price, B., Zhang, J., Kim, S.Y., Aliaga, D.: Objectstitch: Object compositing with diffusion model. In: CVPR (2023) [4](#)

39. Song, Y., Zhang, Z., Lin, Z., Cohen, S., Price, B., Zhang, J., Kim, S.Y., Zhang, H., Xiong, W., Aliaga, D.: Imprint: Generative object compositing by learning identity-preserving representation. In: CVPR (2024) 4
40. Sun, S., Ye, F., Gong, S.: Training-free zero-shot composed image retrieval with local concept reranking. arXiv preprint arXiv:2312.08924 (2023) 4
41. Suo, Y., Ma, F., Zhu, L., Yang, Y.: Knowledge-enhanced dual-stream zero-shot composed image retrieval. In: CVPR (2024) 3, 4
42. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: WACV (2022) 2
43. Tang, L., Ruiz, N., Chu, Q., Li, Y., Holynski, A., Jacobs, D.E., Hariharan, B., Pritch, Y., Wadhwa, N., Aberman, K., et al.: Realfill: Reference-driven generation for authentic image completion. ACM Transactions on Graphics (ToG) (2024) 15
44. Tang, Y., Yu, J., Gai, K., Zhuang, J., Xiong, G., Hu, Y., Wu, Q.: Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In: AAAI (2024) 3, 4
45. Tang, Y., Zhang, J., Qin, X., Yu, J., Gou, G., Xiong, G., Lin, Q., Rajmohan, S., Zhang, D., Wu, Q.: Reason-before-retrieve: One-stage reflective chain-of-thoughts for training-free zero-shot composed image retrieval. In: CVPR (2025) 4
46. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024) 4
47. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025) 2, 3
48. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM (2016) 1, 5
49. Tian, X., Li, W., Xu, B., Yuan, Y., Wang, Y., Shen, H.: Mige: Mutually enhanced multimodal instruction-based image generation and editing. In: ACM MM (2025) 5
50. Tsai, Y.J., Price, B., Liu, Q., Figueroa, L., Pakhomov, D., Ding, Z., Cohen, S., Yang, M.H.: Completeme: Reference-based human image completion. In: ICCV (2025) 1, 2, 4, 9, 11, 12, 13, 18
51. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) 2, 3
52. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025) 2, 3, 9, 10
53. Wang, Y., Lin, Z., Shen, X., Mech, R., Miller, G., Cottrell, G.W.: Event-specific image importance. In: CVPR (2016) 4, 5
54. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP (2004) 9
55. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: NeurIPS (2022) 3
56. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 9, 10

57. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: CVPR (2023) [2](#), [4](#)
58. Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., Wang, L.: Mm-react: Prompting chatgpt for multimodal reasoning and action. arXiv preprint arXiv:2303.11381 (2023) [3](#)
59. Zeng, A., Lv, X., Zheng, Q., Hou, Z., Chen, B., Xie, C., Wang, C., Yin, D., Zeng, H., Zhang, J., et al.: Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. arXiv preprint arXiv:2508.06471 (2025) [2](#), [3](#)
60. Zhang, J., Li, M., Tang, J., Deng, J., Tian, S., Liu, X., Zhang, M., Ye, G., Jiang, Y.G.: Editmaster: Bridging text instruction and visual example for multimodal guided image editing. In: ACM MM (2025) [5](#)
61. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) [9](#)
62. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. TMLR (2024) [3](#)
63. Zhou, Y., Barnes, C., Shechtman, E., Amirghodsi, S.: Transfill: Reference-guided image inpainting by merging multiple color and spatial transformations. In: CVPR (2021) [2](#), [4](#)