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

7 Overview

This supplementary document provides additional details and extended experimental results for the main paper. Sec. 8 describes additional details of the AlbumFill Benchmark. Sec. 9 presents the prompt templates used in our pipeline. Sec. 10 provides a comparison with the fine-tuning based method. Sec. 11 shows additional qualitative comparisons with reference-based image completion methods.

8 Additional Details of the AlbumFill Benchmark

To better understand the structure of CUFED albums and justify our dominant-identity construction strategy, we analyze the distribution of album sizes as well as statistics of detected faces and identity clusters. Face detection is performed using YOLOv8, and identity embeddings are extracted using InsightFace for identity clustering.

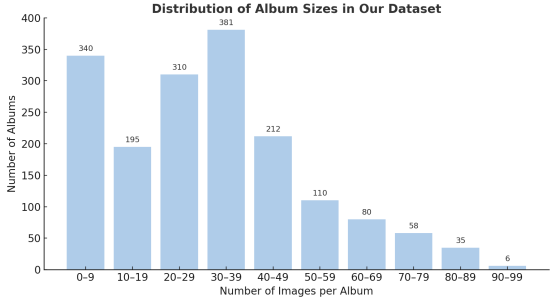

Fig. 5: Distribution of album sizes in the constructed dataset. Each bar represents the number of albums whose sizes fall within the corresponding range of images per album. Most albums contain between 30 and 40 images, with an average of 30.86 images per album, reflecting moderate album-size diversity similar to real-world personal photo collections.

Fig. 5 shows the distribution of album sizes in the dataset. Each bar represents the number of albums whose sizes fall within the corresponding range of images per album. Most albums contain between 30 and 40 images, with an average of 30.86 images per album. This distribution reflects moderate variation in album size, which resembles the structure of real-world personal photo collections.

Table 6: Statistics of detected faces and identity clusters in CUFED albums. Face detection is performed using YOLOv8 and identity embeddings are extracted using InsightFace. Statistics are computed across all 1,727 albums.

Measure	Min	Mean	Max
Total faces per album	0	122.20	1342
Unique identities per album	0	7.47	104

Table 6 summarizes face detection and identity clustering statistics across all 1,727 albums in CUFED. On average, each album contains 122.20 detected faces and 7.47 unique identities, indicating that multiple individuals frequently appear within the same album. This observation motivates the need for identity clustering when constructing personalized albums from CUFED.

After clustering, we select the dominant identity in each album, defined as the individual appearing in the largest number of images. This procedure yields identity-consistent personal albums while preserving variations in pose, clothing, viewpoint, and environment. The dominant identity typically accounts for the majority of faces within an album, which aligns with real-world personal photo collections where a primary subject appears frequently alongside occasional secondary individuals.

Examples from the constructed AlbumFill Benchmark are shown in Fig. 6. These examples illustrate the diversity of scenes, poses, and backgrounds in the dataset. Such real-world variability makes identity-preserving retrieval and completion more challenging than in curated datasets, providing a realistic benchmark for evaluating personalized image completion systems.

9 Prompt Templates

In this section, we provide the prompt templates used in different stages of the AlbumFill pipeline to ensure reproducibility. Specifically, we include the prompt used for the masked visual reasoning stage (Sec. 9.1), the prompt used for the instruction evaluation protocol (Sec. 9.2), and the prompt used for MLLM-based completion baselines (Sec. 9.3). These prompts correspond to the sequential stages of our framework and are presented in the same order as they are applied in the pipeline.

9.1 Stage-1 Reasoning Prompt

In the first stage of our pipeline, a vision-language model performs masked visual reasoning to infer the likely semantics of the occluded region. The goal of this stage is not to generate a full image caption, but to produce a concise manipulation instruction that describes how the missing region should be completed. This instruction serves as a textual cue for composed image retrieval.

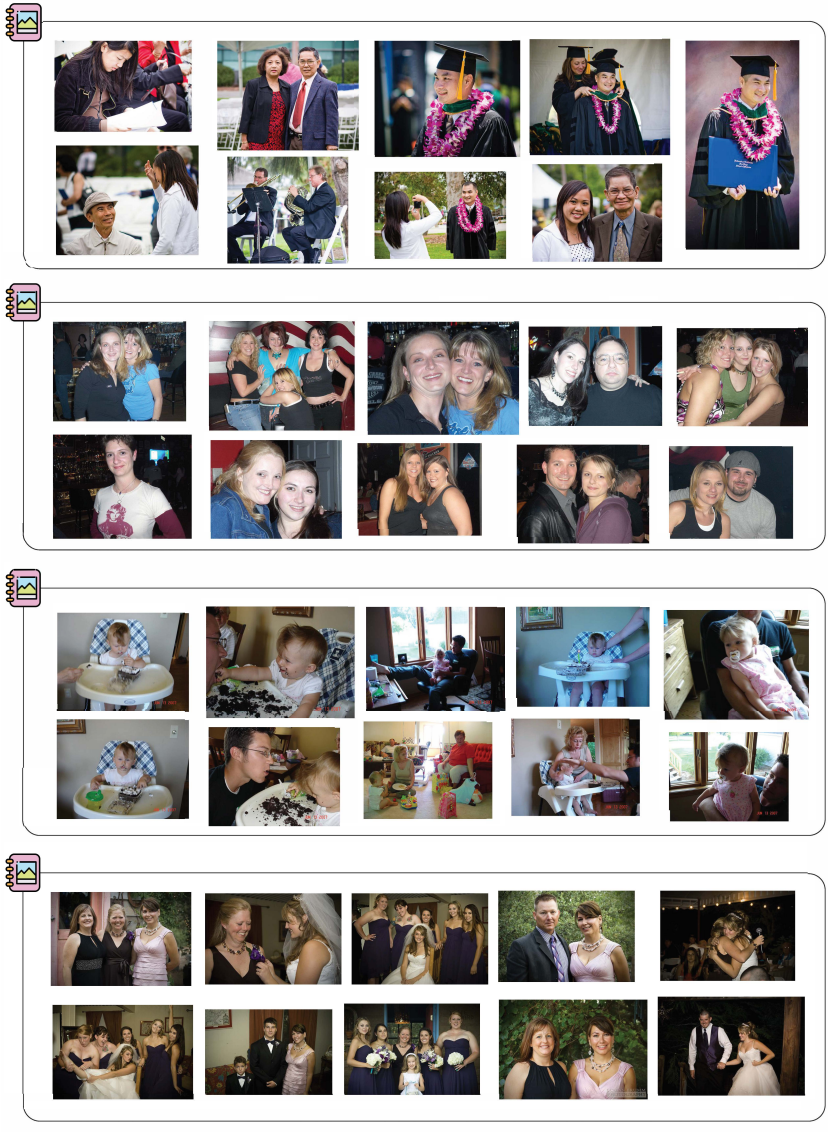

Fig. 6: Examples from AlbumFill Benchmark (Sec. 3 and Sec. 8). Each row shows a sample personal album constructed using the dominant-identity strategy. Images within the same album contain the same primary individual appearing across different poses, viewpoints, and scenes, while occasionally including other people in the background. These examples illustrate the diversity of real-world personal photo collections, including variations in lighting, environment, and human interactions. Such variability makes identity-consistent retrieval and completion challenging and highlights the need for robust personalized completion methods.

You are given an image with randomly masked regions.

Your task is to analyze the visible areas and infer what types of content are likely missing, based only on contextual, structural, and semantic cues.

Do NOT guess specific unseen objects.
Instead, describe the missing regions in terms of:
- **Visual roles** (e.g., a missing body part, missing background area)
- **Attributes** (color, texture, material)
- **Geometric or spatial relations** (left side, upper region, behind a person)
- **Continuity with visible structures** (hair, clothing, floor, environment)

Your output should be a manipulation text suitable for composed image retrieval.
It should be:
- Descriptive enough to guide retrieval **(10–25 words)**
- Focused on the modifications needed to complete the masked image
- Phrased as an edit instruction
- **NOT** a full image caption

Examples:
- Restore the missing left half of the woman's torso and continue the blue dress texture
- Complete the masked background on the right side and extend the wooden wall pattern
- Fill the missing lower body and match the pants texture visible above the mask

Now output **ONLY** the manipulation text.

Fig. 7: Prompt template used for the masked visual reasoning stage. The vision-language model is instructed to analyze the visible regions of a masked image and infer the likely semantics of the missing content using contextual and structural cues. The prompt explicitly discourages speculative guesses and instead encourages grounded descriptions based on observable evidence. The model outputs a short manipulation instruction (10–25 words) describing how the masked region should be completed. This manipulation text is later used as a textual cue for composed image retrieval.

To ensure a fair comparison across different reasoning models, we use the same prompt template for all methods evaluated in our experiments. The prompt instructs the model to analyze only the visible regions of the masked image and infer the missing content based on contextual, structural, and semantic cues. Importantly, the prompt explicitly discourages speculative guessing of unseen objects and instead encourages descriptions grounded in visible evidence.

The output of this stage is a short manipulation text (10–25 words) describing the expected completion of the masked region. This text is then combined with the masked image to form the composed query used in the subsequent image retrieval stage. The full prompt template used for this reasoning stage is shown in Fig. 7.

Masked Region Completion Instruction Evaluation Prompt

You are a strict evaluator for a Masked Image Completion Instruction Generation task.

You are given:
- An image with masked regions
- A manipulation instruction generated by a model

The instruction is NOT a caption and NOT a guess of hidden objects.
It should describe how to complete the missing region using only visible evidence.

Your job is to judge whether the instruction is useful for retrieving a correct reference image for completion.

Be strict. Reward only grounded and retrieval-useful instructions.

Score each dimension from 0 to 20. High scores (18–20) should be rare.

Evaluation Dimensions

1. Evidence Grounding (0–20)
Does every part of the instruction come from visible image evidence?

Penalize:
- guessing hidden objects
- semantic assumptions not supported visually
- adding identity or category labels without visible proof

Reward:
- references to visible edges, textures, boundaries, pose continuation

Score Logic:
High (16–20): fully grounded in visible cues
Mid (8–15): partially grounded but mild assumptions
Low (0–7): speculative

2. Structural Continuity (0–20)
Does the instruction describe how structures should continue across the mask?

Look for:
- continuation of body parts
- surface continuation
- depth ordering
- occlusion relationships

Score Logic:
High (16–20): clear geometric completion constraints
Mid (8–15): vague continuation
Low (0–7): none

Fig. 8: Evaluation prompt used for instruction quality assessment (Part 1). The evaluator receives a masked image and a manipulation instruction generated by a reasoning model. The prompt instructs the evaluator to judge whether the instruction is useful for retrieving a correct reference image for completion, while discouraging speculative reasoning about hidden objects. The first part of the prompt defines the evaluation task and introduces the scoring criteria.

3. Retrieval Discriminativeness (0–20)
Would this instruction help distinguish the correct reference image from many candidates?

Penalize generic descriptions:
- complete the missing area
- fill the blank region

Reward constraints:
- location
- material
- part relationship
- interaction

Score Logic:
High (16–20): narrow retrieval space significantly
Mid (8–15): somewhat useful
Low (0–7): too generic

4. Instruction Format Quality (0–20)
Is the output a proper edit instruction rather than a caption?

Penalize:
- full sentence captioning
- unrelated visible descriptions
- narrative wording

Reward:
- concise imperative edit instruction
- focused only on missing region

Score Logic:
High (16–20): clean manipulation instruction
Mid (8–15): partially descriptive
Low (0–7): caption-like

Output Format (strict JSON)

```
{
  "evidence_grounding_score": int,
  "structural_continuity_score": int,
  "retrieval_discriminative_score": int,
  "instruction_format_score": int,
  "final_verdict": "ACCEPT or RETRY",
  "reasoning": "concise strict explanation"
}
```

Fig. 9: Evaluation prompt used for instruction quality assessment (Part 2). The second part of the prompt specifies four evaluation dimensions: evidence grounding, structural continuity, retrieval discriminativeness, and instruction format quality. Each dimension is scored from 0 to 20, and the evaluator returns the scores together with a final decision ($ACCEPT$ or $RETRY$) in a structured JSON format.

Fig. 10: Examples of instruction evaluation results produced by Gemini 2.5 Pro. For each masked input image, we show manipulation instructions generated by different reasoning models and the corresponding evaluation outputs. The evaluator scores each instruction along four dimensions and provides a final decision. Instructions that are grounded in visible evidence and contain discriminative retrieval cues receive higher scores, while speculative or overly generic instructions are penalized.

9.2 Evaluation Protocol Prompt

To assess the quality of manipulation instructions produced during the masked visual reasoning stage, we design an automatic evaluation protocol using a strong vision-language model as an external evaluator. In our experiments, we use Gemini 2.5 Pro [11] as the evaluation model. The evaluator receives two inputs: the masked image and the manipulation instruction generated by a reasoning model.

The objective of this evaluation is not to measure linguistic similarity to a ground-truth description, but rather to determine whether the instruction is useful for retrieving a correct reference image for completion. Therefore, the evaluation protocol focuses on properties that are important for composed image retrieval.

Specifically, the evaluator scores each instruction along four dimensions: *Evidence Grounding*, *Structural Continuity*, *Retrieval Discriminativeness*, and *Instruction Format Quality*. Each dimension is scored on a scale from 0 to 20, where higher scores indicate stronger grounding and higher usefulness for retrieval. The evaluation prompt explicitly encourages strict scoring, rewarding instructions that are grounded in visible image evidence while penalizing speculative guesses about unseen objects.

Figs. 8 and 9 present the full evaluation prompt and scoring rubric used by the evaluator. The evaluator returns scores for all four dimensions together with a final decision (*ACCEPT* or *RETRY*) and a concise explanation.

Fig. 10 shows several examples of evaluation results produced by Gemini 2.5 Pro. For each masked input, we display the instruction generated by different reasoning models and the corresponding evaluation scores. These examples illustrate how the protocol distinguishes well-grounded instructions from speculative or non-discriminative ones.

9.3 Completion Prompts for MLLM-Based Methods

We further investigate the behavior of recent multimodal large language models (MLLMs) when performing reference-based image completion using natural language instructions. To this end, we design three prompt variants with increasing levels of constraint: *Low-level*, *Medium-level*, and *High-level* prompts. These prompts progressively introduce stricter task descriptions and editing constraints in order to guide the model toward faithful reference-based inpainting.

The three prompt templates are illustrated in Figs. 11, 12, and 13. The low-level prompt provides only a minimal task description, instructing the model to fill the missing region in the masked input image using a reference image. The medium-level prompt adds several explicit constraints, such as preserving pixels outside the masked region and maintaining the same image resolution. Finally, the high-level prompt specifies a comprehensive set of editing rules, explicitly prohibiting operations such as resizing, pasting the reference image, modifying non-masked regions, or returning the reference image directly.

To evaluate the effectiveness of these prompts, we apply the same three prompt templates to two representative MLLM-based image editing systems:

You are given two images.

The first image **(IMG1)** is a masked input image containing a missing region (a hole).
The second image **(IMG2)** is a reference image.

Task: Fill the hole in **IMG1** using **IMG2** as reference.

Constraints:
1) Preserve everything in **IMG1** outside the masked region.
2) Use **IMG2** only to guide what should appear in the missing region.
3) Output must keep the same resolution as **IMG1**.

Return only the edited image.

Fig. 11: Low-level completion prompt. A minimal instruction describing the reference-based image completion task. The prompt specifies that the model should fill the missing region in the masked image using a reference image but provides only limited editing constraints.

You are given two images.

The first image **(IMG1)** is the masked input image containing a missing region.
The second image **(IMG2)** is a reference image.

Your task:
Fill only the missing region in **IMG1** using visual information from **IMG2**.

Important constraints:
- The first image **(IMG1)** defines the output canvas.
- **Do NOT** resize, crop, pad, or change the aspect ratio.
- The output resolution must be exactly identical to **IMG1**.
- **Do NOT** return **IMG2**.
- **Do NOT** replace the entire image.
- **Do NOT** paste **IMG2** directly.
- Only modify pixels inside the masked region.
- Preserve every pixel outside the masked region exactly as in **IMG1**.

Use **IMG2** only as a visual guide for structure, texture, lighting, and semantics.
Return a single edited image.

Fig. 12: Medium-level completion prompt. An extended prompt that introduces additional constraints for reference-based image completion, including preserving pixels outside the masked region and maintaining the same output resolution as the input image.

You are given two images.

Image A (IMG1) is a masked image with a missing region.
Image B (IMG2) is a reference image.

Your task:
Perform reference-based image inpainting on **Image A**, using visual information from reference **Image B**.

Definition of the task:
- **Image A** defines the fixed canvas.
- The output image **MUST** have identical width and height as **Image A**.
- The output image **MUST** have the same aspect ratio as **Image A**.
- **No** resizing, cropping, padding, or scaling is allowed.
- **No** change to global framing is allowed.

Editing constraints:
- Modify **ONLY** the pixels inside the masked region of **Image A**.
- Pixels outside the masked region **MUST** remain identical to **Image A**.
- **Do NOT** replace the entire image.
- **Do NOT** paste or overlay **Image B**.
- **Do NOT** return **Image B**.
- **Do NOT** merge **Image A** and **Image B** directly.
- **Do NOT** change colors, lighting, or structure outside the masked region.

Use **Image B** only as semantic and structural guidance to reconstruct plausible content in the missing region of **Image A**.

The final output must:
- Be a single image
- Have identical dimensions as **Image A**
- Preserve all non-masked pixels exactly

Return only the edited image.

Fig. 13: High-level completion prompt. A highly constrained prompt specifying strict editing rules for reference-based image completion. The prompt explicitly prohibits operations such as resizing the image, modifying pixels outside the masked region, directly returning the reference image, or merging the reference image with the input.

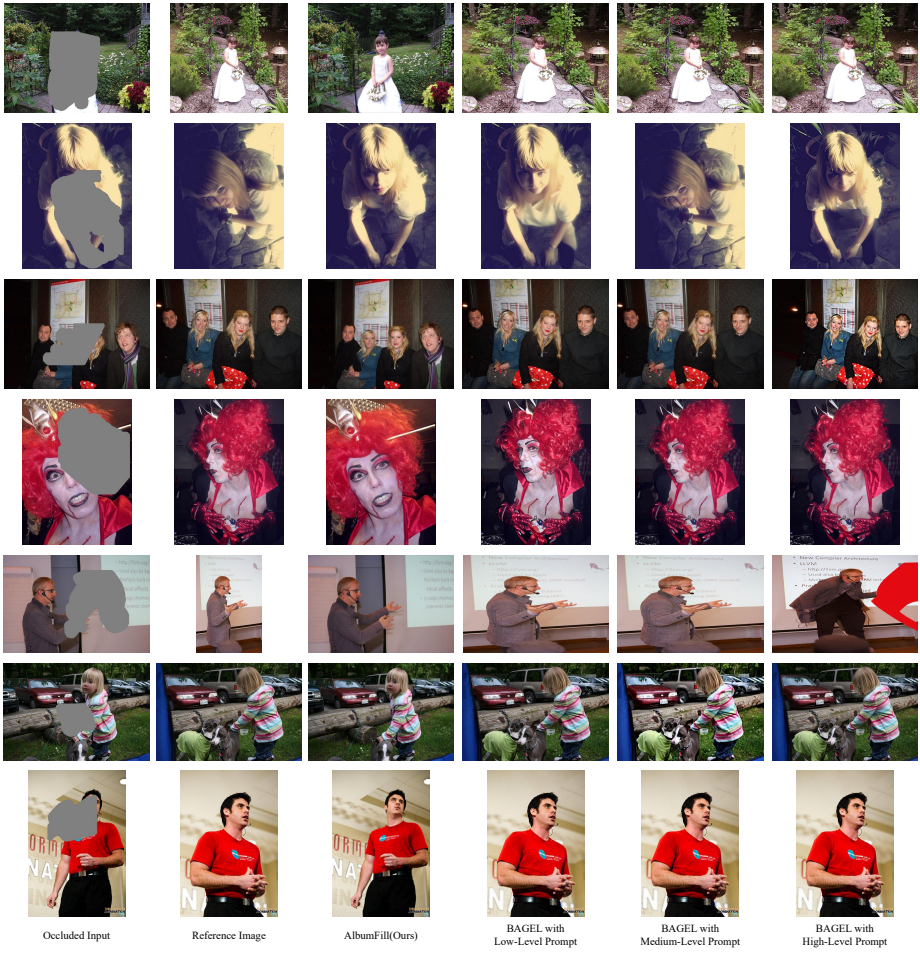

Fig. 14: Visual comparison with BAGEL under different prompt levels. We apply the same low-, medium-, and high-level completion prompts to BAGEL. Despite increasingly strict prompt constraints, BAGEL frequently produces incorrect results, such as copying the reference image directly or generating outputs that violate the masked editing requirement.

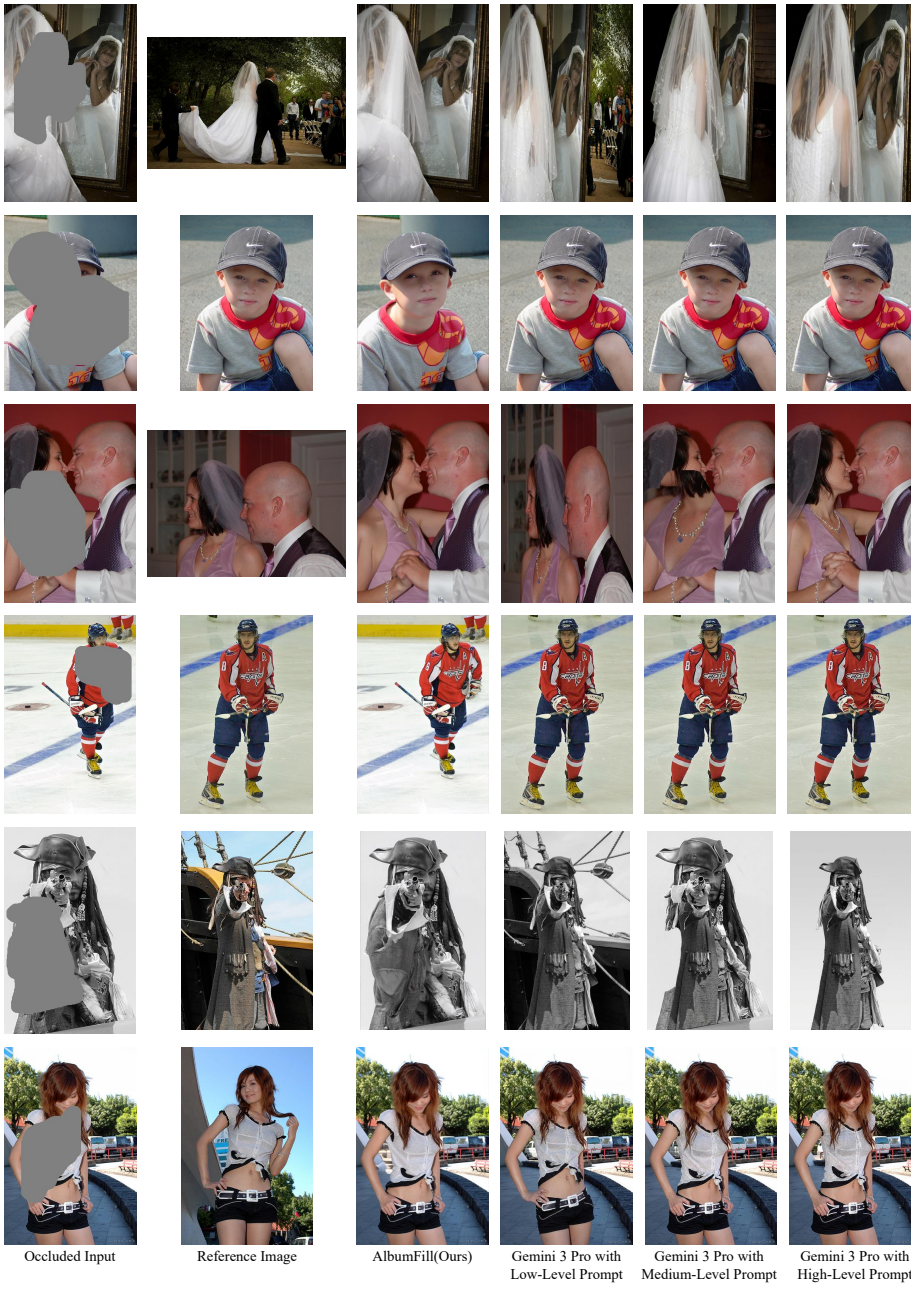

Fig. 15: Visual comparison with Gemini 3 Pro under different prompt levels. Even with explicit prompt constraints, Gemini 3 Pro occasionally returns the reference image, produces outputs with incorrect aspect ratios, or directly combines content from the reference and target images. These behaviors indicate that prompt-based control alone is insufficient to reliably enforce reference-based editing constraints.

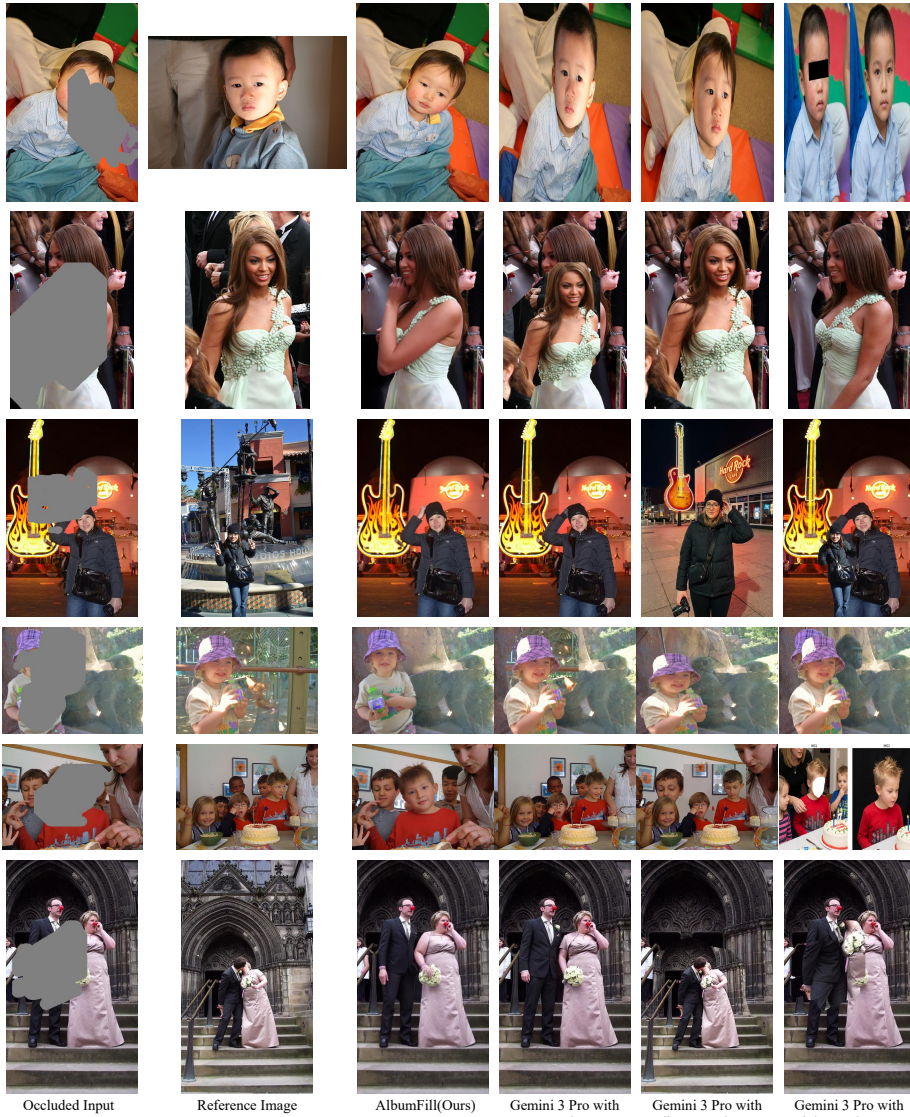

Fig. 16: Additional qualitative comparisons with Gemini 3 Pro. These examples further illustrate common failure modes of MLLM-based editing systems, including replacing the entire input image, blending reference content directly into the target image, or generating outputs that do not preserve the original canvas.

Table 7: Effect of prompt complexity on MLLM-based reference image completion. We evaluate three prompt templates with increasing levels of editing constraints (Low, Medium, High) for BAGEL [14] and Gemini 3 Pro [13]. The same prompts are used for both models. Results indicate that increasing prompt strictness does not reliably improve editing performance. BAGEL produces nearly identical results across prompt levels, while Gemini 3 Pro shows unstable performance variations. Despite explicitly specifying editing constraints in the prompt, both models remain significantly worse than AlbumFill, highlighting the limitations of prompt-based control for reliable reference-guided image completion.

Method	Prompt Level	CLIP ↑	DINO ↑	DreamSim ↓	LPIPS ↓	PSNR ↑	SSIM ↑
BAGEL [14]	Low	77.61	66.09	0.363	0.654	9.43	0.173
BAGEL [14]	Medium	77.52	66.11	0.363	0.654	9.43	0.172
BAGEL [14]	High	77.52	66.29	0.364	0.652	9.47	0.174
Gemini 3 Pro [13]	Low	88.19	85.76	0.167	0.343	15.36	0.499
Gemini 3 Pro [13]	Medium	85.43	82.65	0.208	0.393	14.11	0.448
Gemini 3 Pro [13]	High	89.35	88.26	0.147	0.322	15.83	0.511
AlbumFill(Ours)		**94.70**	**95.71**	**0.055**	**0.142**	**22.60**	**0.790**

BAGEL [14] and Gemini 3 Pro [13]. This controlled setup allows us to analyze how prompt complexity influences editing performance for these models.

Table 7 reports quantitative results across multiple similarity and reconstruction metrics. The results show that increasing prompt strictness does not consistently improve performance. In particular, BAGEL exhibits almost identical performance across all prompt levels, indicating that stronger textual constraints have limited impact on the model's editing behavior. Gemini 3 Pro shows slightly larger variations across prompts, but the overall performance remains substantially lower than our method.

More importantly, qualitative comparisons reveal several common failure patterns for MLLM-based approaches. As shown in Figs. 14, 15, and 16, these models frequently violate the editing constraints specified in the prompt. Typical failure cases include:

- Returning the reference image directly instead of completing the masked input.
- Generating outputs with incorrect aspect ratios or image resolutions.
- Directly copying or blending large portions of the reference image into the target image.
- Replacing the entire input image rather than modifying only the masked region.

These results indicate that, even with carefully designed prompts that explicitly forbid such behaviors, current MLLM-based editing systems still exhibit hallucination-like behaviors and constraint violations. In contrast, our method performs structured reference-based completion and consistently produces faithful reconstructions that respect the masked input image.

Table 8: Comparison with the fine-tuning based method RealFill. We evaluate both methods on 16 albums. For RealFill, four images containing the primary subject are manually selected from each album to construct the fine-tuning set, following the default training configuration from the original paper. Each RealFill model requires approximately one hour of fine-tuning per album. While RealFill achieves slightly better LPIPS and SSIM values, AlbumFill obtains higher CLIP, DINO, DreamSim, and PSNR scores, indicating stronger semantic alignment with the reference images and improved reconstruction quality without any per-subject fine-tuning.

Method	CLIP ↑	DINO ↑	DreamSim ↓	LPIPS ↓	PSNR ↑	SSIM ↑
RealFill [43]	95.26	94.92	0.060	**0.124**	22.02	**0.840**
AlbumFill(Ours)	**95.97**	**97.21**	**0.035**	0.148	**23.05**	0.782

10 Comparison with Finetuned Method

We further compare our approach with a fine-tuning based reference completion method, RealFill [43]. RealFill performs subject-specific fine-tuning to adapt a diffusion model to a particular identity before performing image completion.

For this experiment, we randomly select 16 albums from the dataset. For each album, we manually select four images that contain the primary subject to construct the fine-tuning set required by RealFill. Following the original RealFill implementation, we use the default training configuration provided by the authors.

In practice, RealFill requires performing a separate fine-tuning process for each album. Each fine-tuning run takes approximately **one hour** on a GPU. Moreover, the model must be retrained whenever the subject's appearance changes significantly (e.g., clothing changes or different contexts). This makes the method computationally expensive and difficult to scale for large personal photo collections.

In contrast, our method performs reference-based completion without any per-identity fine-tuning. The same model can generalize to new albums and subjects without additional training.

Quantitative results averaged over the 16 albums are reported in Table 8. While RealFill achieves slightly better LPIPS and SSIM values, our method obtains higher CLIP, DINO, DreamSim, and PSNR scores, indicating better semantic consistency with the reference image and improved reconstruction fidelity.

Qualitative comparisons are shown in Figs. 17 and 18. The examples demonstrate that both methods are capable of reconstructing plausible content in the masked region. However, RealFill occasionally produces less consistent facial structures or mismatched appearances relative to the reference image. Overall, our method achieves competitive or better visual quality without requiring any expensive fine-tuning.

Fig. 17: Qualitative comparison with RealFill. We compare the completion results produced by AlbumFill and RealFill. For each example, we show the masked input image, the reference image used for guidance, and the reconstructed outputs from both methods. AlbumFill produces results that better preserve semantic consistency with the reference image while maintaining the global structure of the original input image.

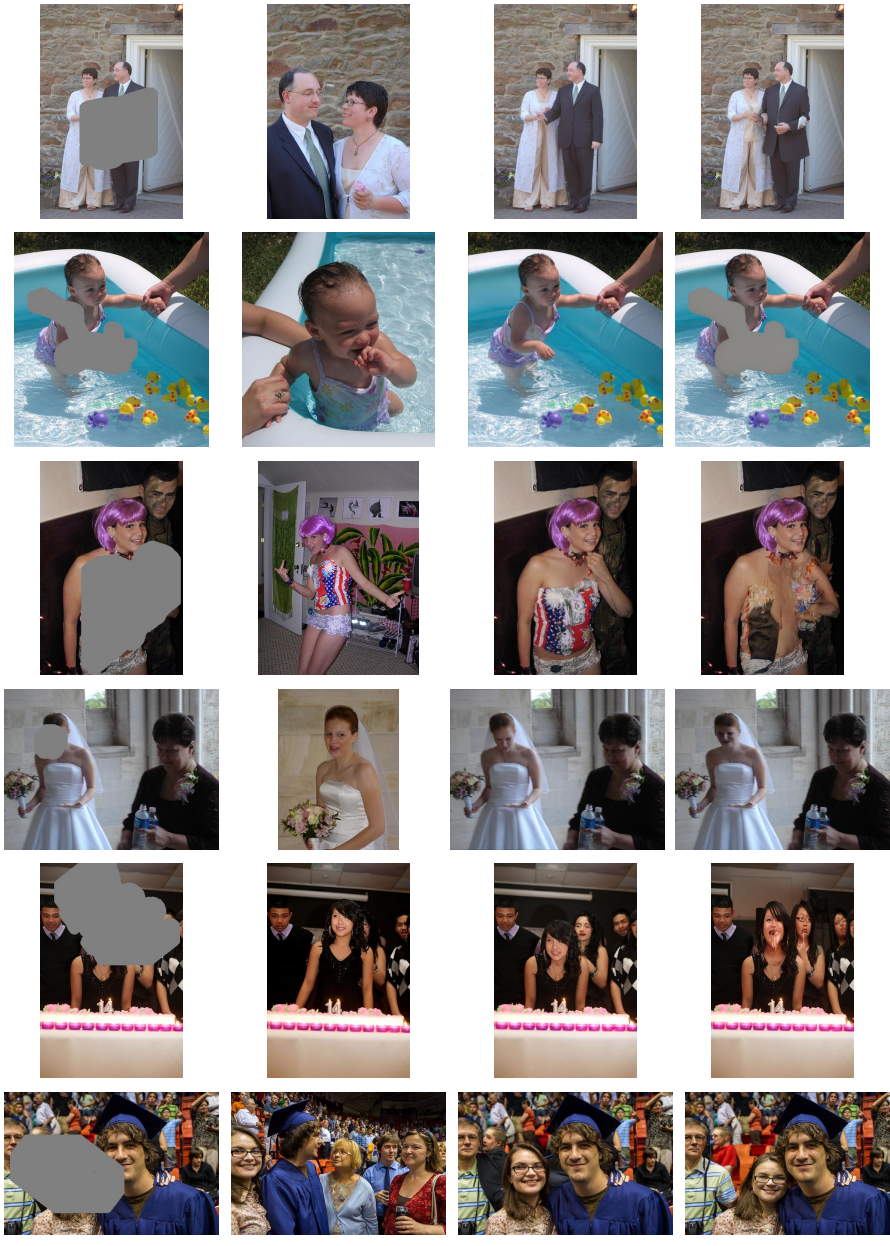

Fig. 18: Additional qualitative comparison with RealFill. These examples further illustrate the differences between AlbumFill and RealFill. While RealFill requires subject-specific fine-tuning for each album, our method performs reference-based completion without additional training and produces comparable or improved reconstruction results.

11 More Visual Comparison

We provide additional qualitative comparisons on our AlbumFill Benchmark with several reference-based image completion methods, including MimicBrush [7], CompleteMe [50], and UniReal(Ours) [9]. The results are shown in Fig. 19 and Fig. 20.

In all examples, the reference images used by our method are obtained entirely through our agentic pipeline, without any manual intervention. Given an occluded input image, the agent first performs masked visual reasoning using a vision-language model to infer the missing content. Based on the inferred description, the system constructs a composed retrieval query and searches within the user's album to identify the most relevant reference image.

Importantly, the reference images shown in these figures are not manually selected or curated. Instead, they are automatically retrieved from the album by the reasoning–retrieval pipeline. This automatic retrieval process enables the system to identify images that contain consistent identity cues, clothing patterns, and contextual information.

As shown in Fig. 19 and Fig. 20, the retrieved references provide strong identity and appearance cues that support realistic completion of the missing regions. Compared with existing reference-based editing approaches, our method produces results that better preserve facial identity, clothing details, and overall appearance consistency while maintaining the original image structure.

These examples further demonstrate that the proposed reasoning, retrieval, and completion pipeline enables reliable and scalable reference-based image completion for real-world personal photo collections.

Fig. 19: Qualitative comparison with reference-based completion methods on our AlbumFill Benchmark (Sec. 3 and Sec. 8). For each occluded input image, the reference image used by our method is automatically retrieved through the proposed reasoning–retrieval pipeline without manual selection. Compared with MimicBrush and CompleteMe, our method produces more identity-consistent and context-aware completions, better preserving facial structure, clothing details, and overall appearance coherence.

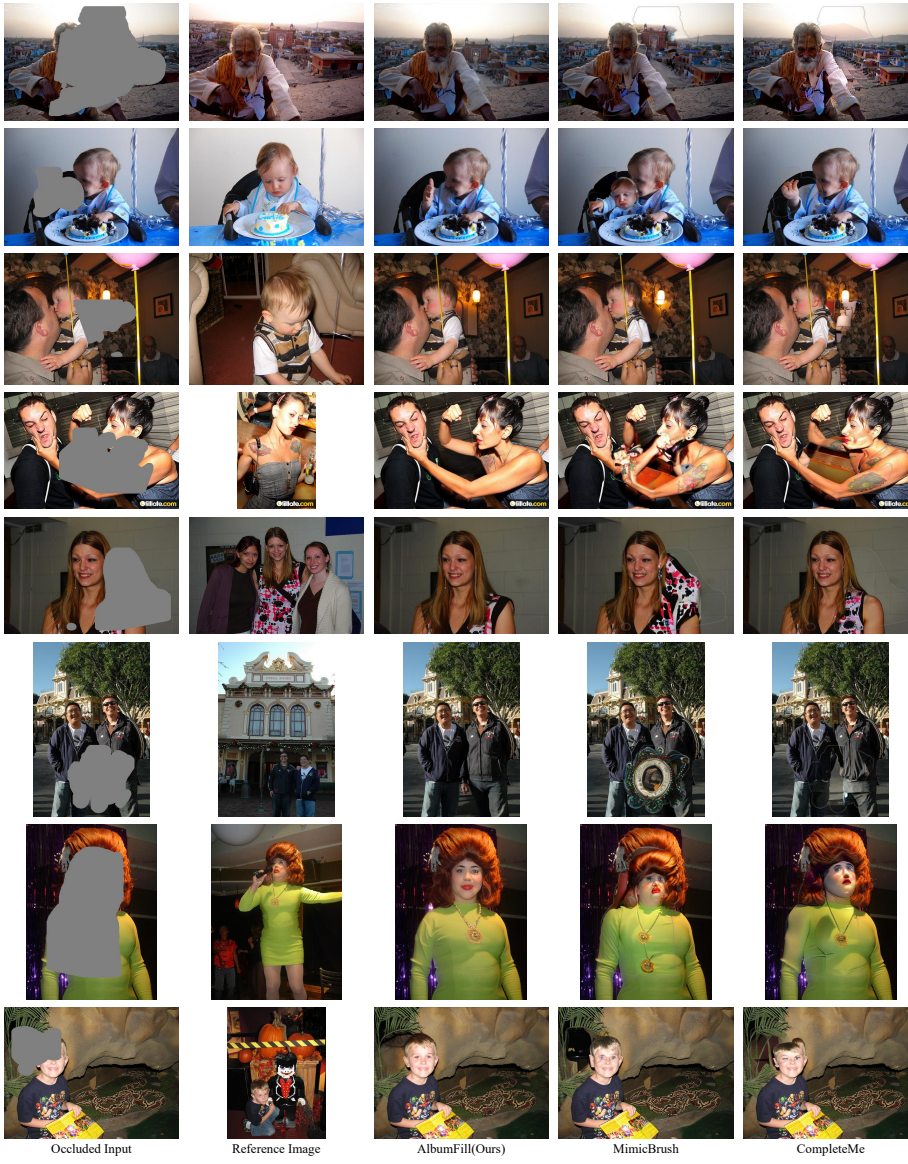

Fig. 20: Additional qualitative comparisons on the AlbumFill Benchmark. The reference images used by AlbumFill are automatically retrieved from the album through masked visual reasoning and composed image retrieval. These automatically retrieved references provide strong identity cues that guide accurate reconstruction of the missing regions, enabling consistent and realistic completion results.