A UNIFIED AND CONTROLLABLE FRAMEWORK FOR LAYERED IMAGE GENERATION WITH VISUAL EFFECTS

Anonymous authors
Paper under double-blind review

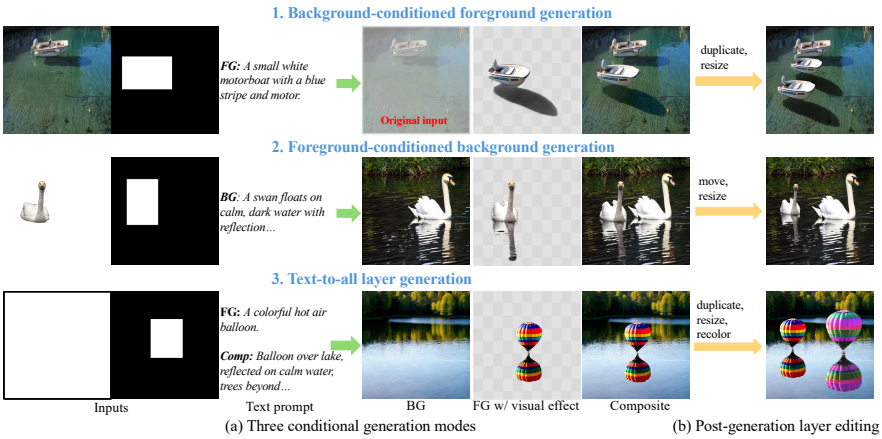

Figure 1: **Layered generation with LASAGNA.** (a) Our framework supports three generation modes: background-conditioned foreground generation, foreground-conditioned background generation, and text-to-all layer generation, which flexibly handle different inputs and jointly synthesize coherent, high-quality composites, backgrounds, and transparent foregrounds with realistic visual effects (*e.g.*, shadows and reflections). (b) Generated layers enable direct post-editing into new, coherent scenes. See the supplementary video demo.

ABSTRACT

Recent image generation models produce impressive composites, but often fail to preserve the identity of user-provided content when editing specific elements: the surrounding scene may shift, and even the edited object's appearance can drift from the original. Layered representations offer a natural remedy—they allow users to independently manipulate individual elements—but existing layered methods typically produce transparent foregrounds without realistic visual effects such as shadows and reflections, forcing the use of a second harmonization model after every edit, which in turn reintroduces drift. To overcome these limitations, we present LASAGNA, which generates a photorealistic background and an RGBA foreground with compelling visual effects in a single forward pass. By treating object-associated visual effects as part of the foreground layer, LASAGNA supports the dominant class of consumer edits (e.g., translation, scaling, recoloring, duplication) via alpha compositing alone, without invoking any model post-edit, thereby eliminating the identity drift inherent to cascade editing pipelines. This single-pass design contrasts with prior layered methods that rely on separate expert models for each task; LASAGNA handles diverse conditioning inputs—text prompts, foreground (FG), background (BG), and location masks—within a unified architecture. We further release two community resources: LASAGNA-48K, the first public dataset of 48K layered image triplets with photorealistic visual effects (curated via a VLM-based filter trained on 30K human-labeled examples), and LASAGNABENCH, the first standardized benchmark for layer-centric generation and editing, comprising 242 expert-annotated samples across six diverse sources. Experiments show that LASAGNA outperforms both general-purpose editors (FLUX, Qwen-Image-Edit, gpt-image-1) and prior layered methods (LayerDiffuse) across three generation modes, and supports a wide range of post-edits without any model re-inference.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Recent advances in text-to-image generation have predominantly leveraged diffusion-based generative models, enabling impressive synthesis quality and semantic accuracy from text prompts (Rombach et al., 2022; Ramesh et al., 2021; Podell et al., 2023; Esser et al., 2024b; Labs et al., 2025; Wu et al., 2025a). Despite their success, these models typically produce images as a single entity, limiting controllability for real-world editing tasks. Consequently, modifications to individual elements within a generated image—such as repositioning, scaling, or adjusting a specific object—often require complex prompt engineering or re-generating the entire image, making it difficult to preserve desired attributes in other regions.

To achieve controllable editing, recent works (Zhong et al., 2025; Zhang & Agrawala, 2024; Dalva et al., 2024; Huang et al., 2025a) explore compositional and layered image generation. This approach, which decomposes generated images into layers, allows for independent manipulation of image components. However, current layered approaches fall short in several critical aspects that prevent their use in real-world scenarios:

I. **Lack of visual effects in foreground layers.** The faithful generation of visual effects like shadows and reflections intrinsically associated with the foreground object is largely overlooked.
II. **Lack of a unified, generative layer framework.** Current approaches often lack a unified framework capable of handling diverse conditional inputs such as foreground (FG), background (BG), masks, and text, thereby limiting their controllability and practical utility.
III. **Absence of public data and standardized evaluation.** As shown in Table 1, most existing methods rely on proprietary or non-public training data. Public datasets like MULAN (Tudosiu et al., 2024) still fall short, as they lack realistic FG visual effects essential for downstream editing. Moreover, the absence of standardized evaluation protocols further hinders meaningful progress comparison across studies.

In this work, we address these limitations and enable controllable, versatile, and realistic layered editing from three complementary perspectives: a unified generation paradigm, publicly available training data, and a standardized benchmark. We present LASAGNA, a novel framework designed to generate images as a composition of FG layers and background layers, explicitly embedding visual effects such as shadows and reflections. Unlike standard editing models that output a single flattened composite, our layered representation guarantees that unedited content is perfectly preserved while remaining freely editable for downstream modifications. This unified architecture simultaneously integrates diverse conditioning inputs and supports three generation modes: background-conditioned foreground layer generation (**FG_Gen**), foreground-conditioned background generation (**BG_Gen**), and text-to-all layer generation (**Text2All**). In **BG_Gen**, where user-provided foreground assets typically lack effects, LASAGNA restores the missing effects while preserving identity, producing an editable RGBA FG for subsequent operations.

To enable LASAGNA training, we introduce LASAGNA-48K, the first publicly available dataset of 48K natural images with faithfully decomposed RGBA FG and BG layers. Critically, these FGs accurately preserve effects like shadows and reflections in relation to the object and its transparency. We will release LASAGNA-48K to facilitate the research and development of models capable of capturing these complex, scene-consistent visual effects.

Furthermore, we introduce LASAGNABENCH to establish a standardized measure for our method and future research. Evaluation in layer editing and generation has been challenging, as prior work relies on bespoke protocols and user studies. LASAGNABENCH provides the first public benchmark for this task, featuring 242 real-world images sourced from 6 diverse datasets, each meticulously decomposed by human experts into high-fidelity, text-paired layers that accurately capture complex visual effects. On LASAGNABENCH, our method achieves superior layer generation while preserving object identity, spatial fidelity, and visual coherence.

In summary, our primary contributions are: **(1)** We present **LASAGNA**, a unified framework supporting three generation modes and flexible conditioning inputs (text, images, masks). It synthesizes realistic composites by jointly or individually generating coherent BGs and RGBA FGs with visual effects, enabling highly controllable, professional photo-editing-tool-style image editing without extra inference. **(2)** A new dataset **LASAGNA-48K**, the first publicly available dataset featuring over 48K natural images with decomposed BG layers as well as FG layers with scene-coherent visual

Table 1: **Layer Dataset Overview.** ✓: available; ×: not available or non-public.

Paper	Layer data (Public)	Eval Bench	Visual Effect
MULAN (Tudosiu et al., 2024)	✓ (✓)	×	×
LayerDiffuse (Zhang & Agrawala, 2024)	✓ (×)	×	×
PSDiffusion (Huang et al., 2025a)	✓ (×)	×	×
LASAGNA (ours)	✓ (✓)	✓	✓

Table 2: **Overview of the three generation modes.**

Mode	Inputs	Targets
FG_Gen	$\{\mathbf{c}_{\text{txt}}, \mathbf{c}_{\text{mask}}, \mathbf{c}_{\text{bg}}\}$	$\{\mathbf{x}_0^{\text{comp}}, \mathbf{x}_0^{\text{fg+ve}}\}$
BG_Gen	$\{\mathbf{c}_{\text{txt}}, \mathbf{c}_{\text{mask}}, \mathbf{c}_{\text{fg}}\}$	$\{\mathbf{x}_0^{\text{comp}}, \mathbf{x}_0^{\text{bg}}, \mathbf{x}_0^{\text{fg+ve}}\}$
Text2All	$\{\mathbf{c}_{\text{txt}}, \mathbf{c}_{\text{mask}}\}$	$\{\mathbf{x}_0^{\text{comp}}, \mathbf{x}_0^{\text{bg}}, \mathbf{x}_0^{\text{fg+ve}}\}$

effects. **(3) LASAGNABENCH**, the first public benchmark for rigorous and standardized evaluation of controllable layer-centric generation and editing. **(4)** Ours achieves high-quality layer generation and editing, particularly for tasks requiring strict identity preservation and harmonious integration.

2 RELATED WORKS

2.1 TEXT-TO-IMAGE AND IMAGE EDITING MODELS

Recent text-to-image diffusion models (Rombach et al., 2022; Ramesh et al., 2021; Podell et al., 2023; Esser et al., 2024b; Labs et al., 2025; Wu et al., 2025a) have made remarkable progress in generating high-fidelity images from text. However, these models are typically confined to single-layer synthesis, lacking an explicit layered representation. Consequently, they cannot produce RGBA outputs or support independent post-generation editing of specific elements without unintended changes to other regions. While specialized image editing models (Chen et al., 2025a; Zhong et al., 2025; Zhang et al., 2023a; Zhao et al., 2024; Liu et al., 2025a) have been developed for common editing tasks, they still struggle with precise object-level control and often introduce non-local artifacts. They are particularly weak in complex spatial edits, such as enlarging or relocating an object while preserving its identity and appearance, as they lack an understanding of the layered composition of the scene. This motivates the development of layer-centric frameworks that inherently support structured, controllable synthesis and editing.

2.2 IMAGE LAYER GENERATION

To enable compositional editing, prior work has explored two main paradigms for layered generation.

Image-layer extraction via post-processing: This common pipeline first uses text-to-image models to generate an RGB composite image. Then, existing segmentation models (Ren et al., 2024; Ravi et al., 2024; Kirillov et al., 2023) can be used to extract an independent FG layer. Finally, inpainting models (*e.g.*, (Zhao et al., 2025; Wei et al., 2025; Winter et al., 2024; Zhuang et al., 2024; Ju et al., 2024)) are typically employed to reconstruct the occluded BG. However, this multi-stage, separately optimized pipeline accumulates errors across stages and often fails to preserve global coherence and cross-layer consistency during post-editing. Operations such as object translation or scaling often produce spatially inconsistent or visually unnatural results.

Direct transparent image layer generation: This paradigm aims to generate layers directly. While initial methods such as those of Fontanella et al. (Fontanella et al., 2024) and LayerDiffuse (Zhang & Agrawala, 2024) target simple scenes, recent works like DreamLayer (Huang et al., 2025b) and PSDiffusion (Huang et al., 2025a) extend this to multi-layer synthesis. However, these models predominantly focus on cartoon or synthetic domains and lack complex physical effects (e.g., shadows and reflections). Furthermore, their published designs do not offer our shared per-frame Type/IO interface for configuring variable input/output sets within one model. This interface reuses learned capabilities for new configurations without changing network structure. In design-centric tasks, ART (Pu et al., 2025) utilizes layout conditioning but remains limited in editing versatility and realistic effect modeling. Other approaches, such as LayerDecomp (Yang et al., 2025a) and Qwen-Image-Layered (Yin et al., 2025), focus on layer decomposition but cannot generate novel content. Crucially, none of these methods provide a unified, controllable framework for generating transparent FGs with high-fidelity, physically-grounded visual effects. Unlike prior works trained on synthetic or design-centric data, LASAGNA leverages real-world images (e.g., COCO) and supports three generation modes with mask-guided spatial constraints, effectively addressing the key barriers to realistic, professional-grade image editing.

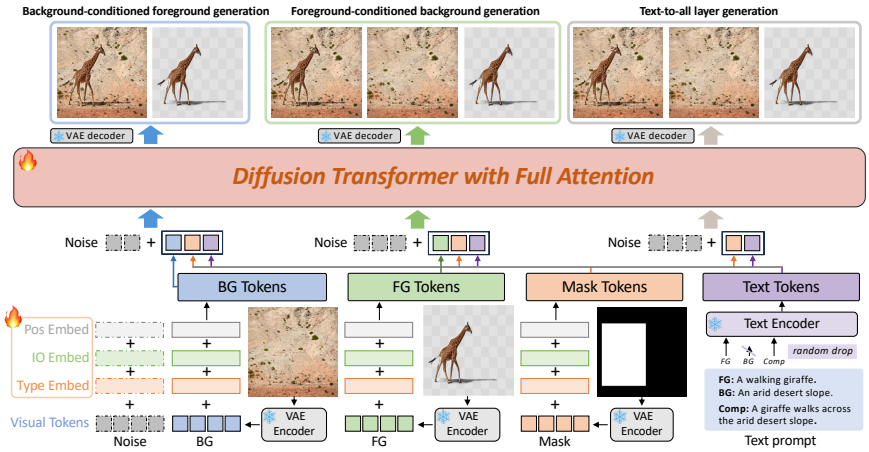

Figure 2: **Framework of LASAGNA.** We formulate the joint generation of composite images, BGs, and FGs as a flexible, layer-conditional denoising task. This unified framework supports multiple workflows, including FG_Gen, BG_Gen, and Text2All. We use a unified input with learnable embeddings to distinguish roles of visual latents (noise, BG, FG, and mask) across tasks, enabling the model to adapt its behavior under various generation settings. This allows a single attention-based model to flexibly process varied combinations of inputs and targets simultaneously.

2.3 LAYER DATASET

Previous studies (Zhang et al., 2023b; Tudosiu et al., 2024; Kang et al., 2025; Huang et al., 2025a) have introduced several layer-related datasets. MULAN, a prominent multi-layer dataset, provides object-level decompositions but does not include explicit visual effects as part of the FG layers. Text2Layer generates a two-layer decomposition, but the dataset is not public and does not incorporate visual effects. PSDiffusion proposes an internal multi-layer dataset, consisting of 30K samples. Except for MULAN, most datasets remain private and none explicitly account for visual effects. To bridge this data gap, we introduce LASAGNA-48K, a new dataset built upon an advanced decomposition model that jointly generates BG and FG layers while faithfully preserving complex visual effects in the FG's alpha channel. In addition, we manually annotate a high-quality layer benchmark. All training and testing data are fully released to encourage transparency and foster further research in this area.

3 APPROACH

3.1 LASAGNA FRAMEWORK

As shown in Fig. 2, LASAGNA models the joint generation of composite images $\mathbf{x}^{\text{comp}}$, BGs $\mathbf{x}^{\text{bg}}$, and FGs with visual effects $\mathbf{x}^{\text{fg+ve}}$ as a flexible, layer-conditional denoising task. Our model learns to denoise a set of target images $\mathbf{X}_t \subseteq \{\mathbf{x}_t^{\text{comp}}, \mathbf{x}_t^{\text{bg}}, \mathbf{x}_t^{\text{fg+ve}}\}$ conditioned on a set of inputs $\mathbf{C} \subseteq \{\mathbf{c}_{\text{txt}}, \mathbf{c}_{\text{mask}}, \mathbf{c}_{\text{bg}}, \mathbf{c}_{\text{fg}}\}$. By varying the composition of $\mathbf{X}_t$ and $\mathbf{C}$, we unify three generation modes (FG_Gen, BG_Gen, and Text2All) in a single model, addressing different real-world editing needs (see Table 2).

We build LASAGNA upon the Diffusion Transformer (DiT) architecture (Peebles & Xie, 2023; Chen et al., 2025a) to support flexible editing tasks, adapted to handle heterogeneous inputs. We employ four embedding types to distinguish between different tasks and image types:

- *Type Embedding*—represents the semantic role of each image, *e.g.*, BG or FG.
- *IO Embedding*—indicates if a frame is used as an input or an output in the current task.
- *Position Embedding*—joint spatial and frame position of image tokens.
- *Timestep Embedding*—the diffusion step.

Text prompts are encoded by T5 (Raffel et al., 2020). All conditional tokens and noisy target tokens are concatenated into a single sequence, allowing the model's self-attention blocks to seamlessly integrate information from any arbitrary set of conditions $\mathbf{C}$ to guide the denoising of targets $\mathbf{X}_t$.

We train our model by optimizing a unified denoising objective across all conditional generation tasks. Each mode uses its specific conditional inputs $\mathbf{C}^{(m)}$ and targets $\mathbf{X}_0^{(m)}$ as defined in Table 2. The model is trained to minimize the joint expectation $\mathcal{L}_{\text{dm}}$ over this multi-task distribution:

$$\mathcal{L}_{\text{dm}} = \mathbb{E}_{m,t,\mathbf{X}_0^{(m)},\boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon}_\theta(\mathbf{X}_t^{(m)}; \mathbf{C}^{(m)}, t) - \boldsymbol{\epsilon}\|_2^2 \right]$$
$$s.t.\ \mathbf{X}_t^{(m)} = \sqrt{\alpha_t}\mathbf{X}_0^{(m)} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon},\ \text{and}\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I).$$

The training loss follows flow matching (Lipman et al., 2022). See Appendix B for technical details. The same interface also supports composite generation conditioned on both a clean FG and a BG (Appendix C).

3.2 LASAGNA-48K DATASET

To enable the training of our controllable, layer-based framework, we introduce LASAGNA-48K, the first publicly available large-scale dataset of over 48K high-quality image triplets with visual effects.

Data sources. We build our dataset from three public sources: MULAN (Tudosiu et al., 2024), COCO 2017 (Lin et al., 2014), and SOBA (Wang et al., 2022). MULAN provides layered data but lacks visual effects. COCO offers diverse scenes with complex layouts for robust editing. SOBA, a shadow dataset, adds rich visual effects for realistic composites. Finally, LASAGNA-48K includes 8K MULAN, 39K COCO, and 1K SOBA samples.

Data construction pipeline. We design a four-stage pipeline to ensure high data fidelity (Fig. 3); examples of generated triplets and captions are shown in Fig. 4.

I. *Non-occluded mask selection.* To isolate unoccluded foremost objects without direct annotations, we generate multiple mask proposals per object utilizing available dataset-specific annotations. Depending on the source, we extract parallel base masks using explicit layer annotations, depth-filtered instance masks (Ke et al., 2025), and a salient segmentation model (Gao et al., 2024). To handle boundary inaccuracies, we augment each base mask with a dilated variant. All proposals are processed independently, and our data curator (Step III) selects the highest-scoring outcome as the final clean layer.

II. *LayerDecomp decomposition.* We process each image and its mask variants with LayerDecomp (Yang et al., 2025a), extracting multiple sets of BGs and FGs with visual effects.

III. *Data filtering.* Due to the consistency loss in LayerDecomp, the quality of extracted BGs and FGs is inherently correlated — enabling BG fidelity alone as a reliable proxy for overall quality. We train a data curator built upon InternVL2.5-8B (Chen et al., 2024b) to assess the BGs fidelity and filter low-quality data. The curator is trained on 30K carefully human-annotated samples (see Appendix K for annotation criteria and examples), and achieves 88.8%/72.3% precision/recall on a 1K human-annotated held-out test set. After filtering, we further use Qwen2.5-VL-32B (Bai et al., 2025) to remove residual artifacts in FG. Finally, a manual inspection of 200 random LASAGNA-48K samples shows that >90% are judged to be good.

IV. *Captioning.* After obtaining high-quality triplets of composite images, BGs, and FGs, we prompt InternVL 2.5-38B to caption these images jointly, considering cross-image relationships for semantically consistent descriptions.

4 EXPERIMENTS

LASAGNA Implementation. We finetune a pre-trained 2B-parameter DiT model. All conditional images and targets are encoded with RGBA-VAE, which is finetuned from DiT VAE using a combination of L1, GAN, and perceptual losses (Zhang et al., 2018). We adopt resolution-specific batching (batch size 6 for 512^2 resolution, 1 for 1024^2) to improve generalization across scales. We use the AdamW optimizer and set the learning rate to 1.2×10^{-5} with linear warm-up for $2K$ steps. The model is trained for 20K iterations. For inference, results are generated using DDIM sampling for 50 steps.

4.1 BENCHMARK

We introduce LASAGNABENCH, the first public benchmark specifically designed for layer-centric image generation. Prior works (Dalva et al., 2024; Zhang & Agrawala, 2024; Huang et al., 2025a)

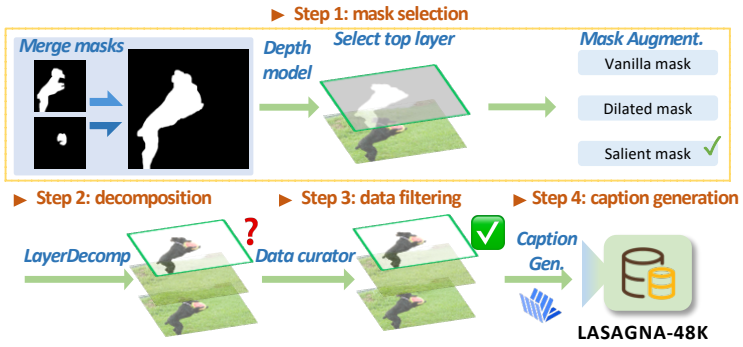

Figure 3: **Data construction pipeline.** Starting with existing datasets, we implement a four-stage data construction pipeline leveraging off-the-shelf models with a custom-trained data curator. This process yields a high-quality dataset as the foundation for subsequent model training.

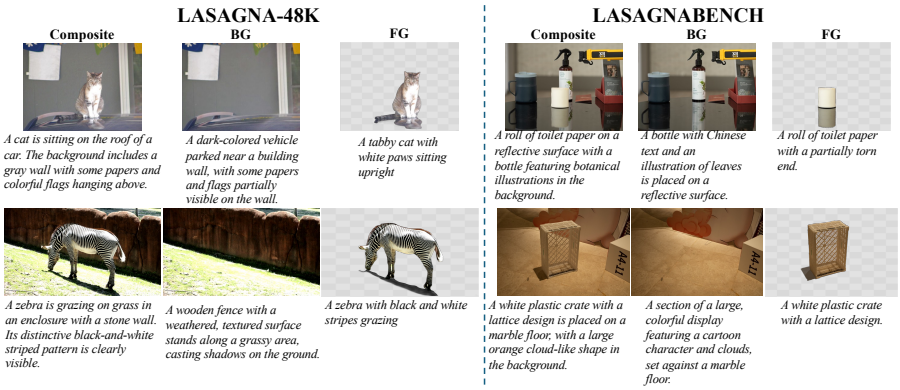

Figure 4: **Samples of LASAGNA-48K and LASAGNABENCH.** Each sample consists of a composite image, a clean BG, and a FG layer with visual effects, along with corresponding captions for all components. Additional examples are shown in Appendix H.

use varied protocols on non-public datasets. LASAGNABENCH contains 242 samples from 6 diverse datasets: 4 public sources (Tudosiu et al., 2024; Lin et al., 2014; Wang et al., 2022; Unsplash) and 2 in-house data sources, all of which will be released (see Fig. 4 for examples and Appendix Table 10 for source statistics). Each sample includes a real photographic composite image, a BG image, an FG with visual effects annotated by professional annotators, and captions. For public datasets, BGs were generated with LayerDecomp, followed by automated curation and manual verification to ensure consistency. For in-house data, we captured controlled real-world photography pairs, following a similar data collection methodology as in prior works (Yang et al., 2025a; Winter et al., 2024). These collection efforts ensure that LASAGNABENCH offers a diverse and high-quality set of layer representations, capturing realistic variations in object appearance and surrounding BG scenes, along with visually authentic visual effects.

4.2 LAYER GENERATION VS GENERAL MODELS

We compare LASAGNA with three leading image generation and editing models: FLUX.1 (Labs et al., 2025; Hugging Face, 2025), Qwen-Image (Wu et al., 2025a), and gpt-image-1[high] (OpenAI, 2025). Since these models do not natively support multi-task generation, we employ their task-specific variants (*e.g.*, inpainting/editing models for conditional tasks, T2I models for Text2All) to ensure fair comparison at their best performance. Because these baselines also lack multi-layer generation capability, we focus evaluation on the generated composite images. We employ FID (Heusel et al., 2017) and CLIP-FID (Parmar et al., 2022) to measure image quality and semantic alignment. Following Complex-Edit (Yang et al., 2025b), we also use a GPT-4o-based score to evaluate Instruction Following and Identity Preservation, averaging them into a final score of 0 to 10. Appendix Table 11 reports separate IF and IP scores.

As shown in Table 3, LASAGNA consistently outperforms prior methods across all modes, achieving higher image quality and better semantic alignment with the input instructions. Qualitative results in

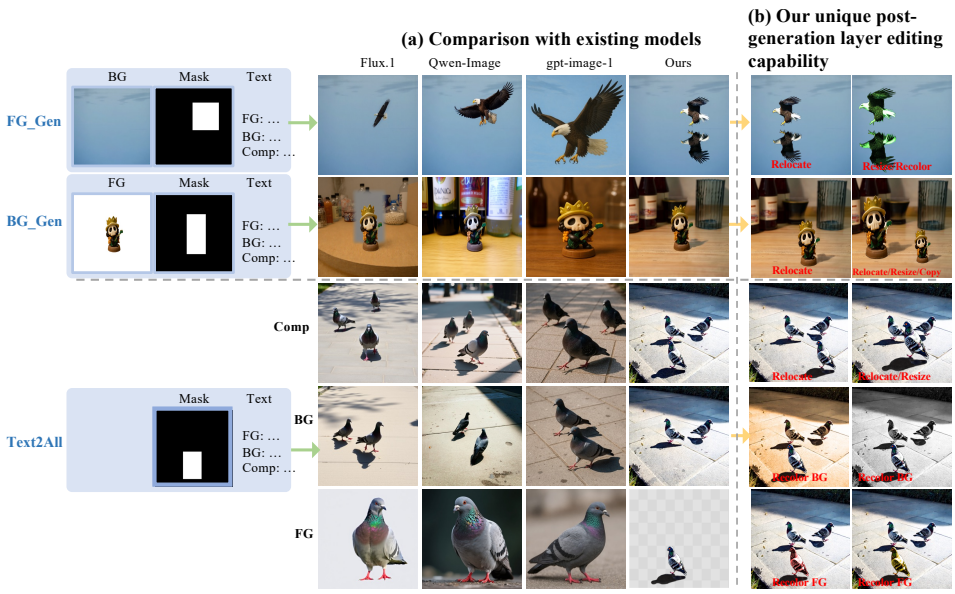

Figure 5: **Layer generation compared with state-of-the-art image generation and editing models.** We compare LASAGNA with Flux.1 (Labs et al., 2025; Hugging Face, 2025), Qwen-Image-Edit (Wu et al., 2025a), and gpt-image-1[High] (OpenAI, 2025). **(a)** Across three distinct generation tasks, LASAGNA consistently achieves superior inter-layer coherence and consistency. In contrast, competing models often fail to maintain these properties. **(b)** Moreover, by generating FGs with faithfully preserved visual effects, LASAGNA enables diverse post-generation editing operations on individual layers directly—a capability not supported by existing models. See Appendix F for additional visualizations and Appendix G for challenging cases.

Table 3: **Comparison with general models for layer generation.** Results for models marked with $*$ are obtained using their respective expert models rather than a single unified model. Specifically, for the FG_Gen and BG_Gen tasks, we use the FLUX.1-Fill-dev, Qwen-Image-Edit-2509, and gpt-image-1[high] editing models, respectively. For the Text2All task, we use the FLUX.1-schnell, Qwen-Image, and gpt-image-1[high] models as text-to-image models, respectively.

Model	# Params	FG_Gen				BG_Gen				Text2All			
		Cond Image	CFID ↓	FID ↓	GPT Score ↑	Cond Image	CFID ↓	FID ↓	GPT Score ↑	Cond Image	CFID ↓	FID ↓	GPT Score ↑
gpt-image-1[high]*	—	BG, Mask	18.1	109.2	8.8	FG, Mask	19.3	108.0	8.9	Mask	26.7	130.2	7.1
FLUX.1*	12B	BG, Mask	10.0	79.6	8.9	FG, Mask	16.1	105.9	8.5	Mask	26.7	133.2	6.0
Qwen-Image-Edit*	20B	BG, Mask	12.1	92.5	9.0	FG, Mask	15.1	101.2	7.9	Mask	20.4	125.7	5.7
Ours	2B	BG, Mask	**9.7**	**72.0**	**9.3**	FG, Mask	**14.1**	**98.6**	**9.0**	Mask	**16.9**	**115.8**	**7.6**

Fig. 5 show superior inter-layer coherence and faithful visual effects. In contrast, competing models generate inconsistent compositions or misalign FGs and BGs. By generating environment-aware visual effects in FG layers, LASAGNA enables direct post-editing and recomposition, which are not supported by existing methods. A blinded user study also favors LASAGNA in average preference on both conditional tasks (Appendix D). Beyond our benchmark, we also report quantitative comparisons on the public benchmarks ImgEdit-Bench (Ye et al., 2025) and GenEval (Ghosh et al., 2023). See Appendix E.1 for benchmark results and Appendix E.2 for ablations on training data, prompt formulation, and joint training.

4.3 LAYER GENERATION VS EXPERT MODEL

We further compare LASAGNA with LayerDiffuse (Zhang & Agrawala, 2024), a prior work specifically designed for layered image generation through multiple expert modes and the only closely related public model. However, LayerDiffuse relies on separate, independently trained models for each task, limiting its controllability and consistency across layers. As shown in Table 4, LASAGNA significantly outperforms LayerDiffuse across all three generation modes in terms of CLIP-FID, confirming the effectiveness of our unified framework in producing coherent and semantically faithful results across diverse generation settings. Qualitative comparisons in Fig. 6 highlight these improvements. FGs generated by LayerDiffuse often appear centered and lack positional diversity due to the absence of explicit spatial constraints. The model also frequently fails to clearly separate FG and BG regions and struggles to represent all described entities when given complex captions. In contrast,

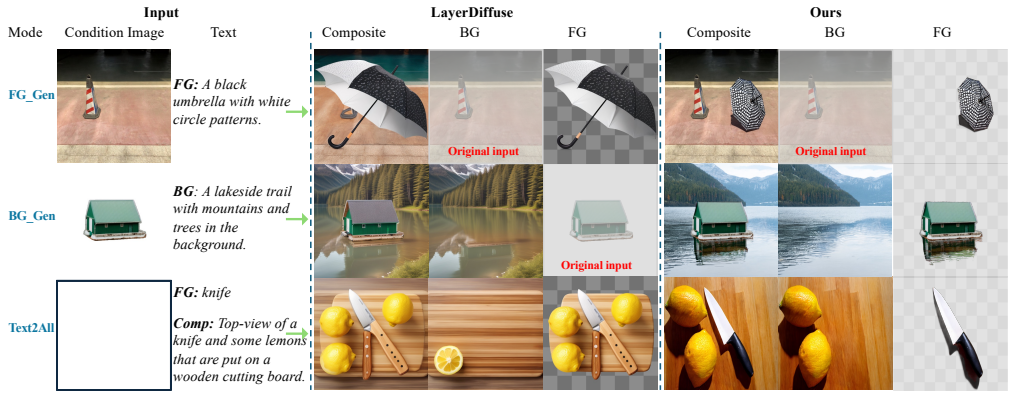

Figure 6: **Layer generation compared with LayerDiffuse (Zhang & Agrawala, 2024).** In FG_Gen, our model produces objects with appropriate size and position, along with realistic shadows consistent with the BG. In BG_Gen and Text2All, our model produces visually consistent results across all layers. Furthermore, it can generate new FGs with corresponding visual effects, enabling flexible and realistic post-editing.

Table 4: **Comparison with LayerDiffuse (Zhang & Agrawala, 2024) for layer generation.** Each cell reports CLIP-FID(↓).

Model	FG_Gen		BG_Gen		Text2All		
	Comp.	FG	Comp.	BG	FG	Comp.	BG
LayerDiffuse	42.0	43.8	43.2	43.1	45.2	46.0	48.2
Ours	**13.4**	**37.3**	**21.0**	**25.6**	**25.5**	**26.2**	**35.8**

Table 5: **Comparison with Qwen (Wu et al., 2025a) for layer editing.** R: recolor; M: movement; C: joint recolor+movement.

Method	**R** CLIP-FID/FID	**M** CLIP-FID/FID	**C** CLIP-FID/FID
Instr. (Qwen)	13.2/102.9	13.4/101.4	15.8/110.8
Casc. Layer (Qwen)	9.5/88.8	8.5/83.6	8.8/86.9
Layer+VE (Ours)	**8.3/71.7**	**6.5/68.4**	**6.4/71.1**

LASAGNA produces spatially controlled, semantically complete, and visually coherent results across all generation modes.

4.4 THE NECESSITY OF LAYERED GENERATION WITH VISUAL EFFECTS

To quantify the value of explicit layer representations with visual effects—and to directly address whether naively cascading existing state-of-the-art models could achieve comparable results—we compare three editing paradigms across three representative editing tasks: recoloring, which alters static appearance attributes; spatial editing, which manipulates object position or scale (*e.g.*, movement, resizing); and complex editing, which combines both to test multi-factor control (see Appendix J for details). The baselines use Qwen-Image (T2I) for composite synthesis and Qwen-Image-Edit-2509 for editing:

I. *Instruct Editing*: Feeds the Qwen-Image composite into Qwen-Image-Edit-2509 and performs direct text-guided modifications.
II. *Cascaded Layer Editing*: A multi-model assembly chaining Qwen-Image → segmentation (Liu et al., 2025b) (FG extraction) → inpainting (BG restoration) → re-composition.
III. *Layer Editing with Visual Effects (Ours)*: Perform the same explicit object-layer editing as above, while additionally incorporating the visual effects generated by LASAGNA.

As shown in Table 5, *Cascaded Layer editing* already outperforms *Instruct editing* across all editing goals, achieving higher fidelity and spatial consistency. *Layer editing with visual effects* achieves the best overall performance with substantial gains in perceptual realism and physical plausibility. Qualitative comparisons in Fig. 7 further highlight these differences. *Instruct editing* often introduces unintended global changes (*e.g.*, altered background tones during recoloring) and struggles to achieve precise spatial edits. In contrast, *Cascaded Layer editing* enables fine-grained spatial manipulation while preserving object identity, which is critical for user-driven image editing applications. Finally, *Layer editing with visual effects* produces the most visually coherent and realistic results, generating shadows and reflections that are physically consistent with the surrounding scene. These results confirm that explicitly modeling layered representations with physically grounded visual effects is crucial for achieving realistic, consistent, and controllable image manipulation.

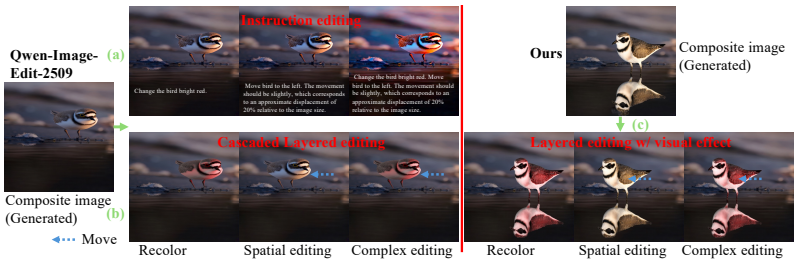

Figure 7: **The Necessity of Layered Generation with Visual Effects.** We demonstrate the benefits of explicit layer representations with visual effects by progressively comparing three paradigms: *Instruct Editing*, *Cascaded Layer Editing*, and *Layer Editing with Visual Effects* (LASAGNA). Across recoloring, spatial, and compositional editing tasks, the lack of explicit layer representations makes *Instruct Editing* prone to unintended changes and less responsive to spatial instructions, while *Layer Editing with Visual Effects* yields more coherent and photorealistic results. *Zoom in for details*.

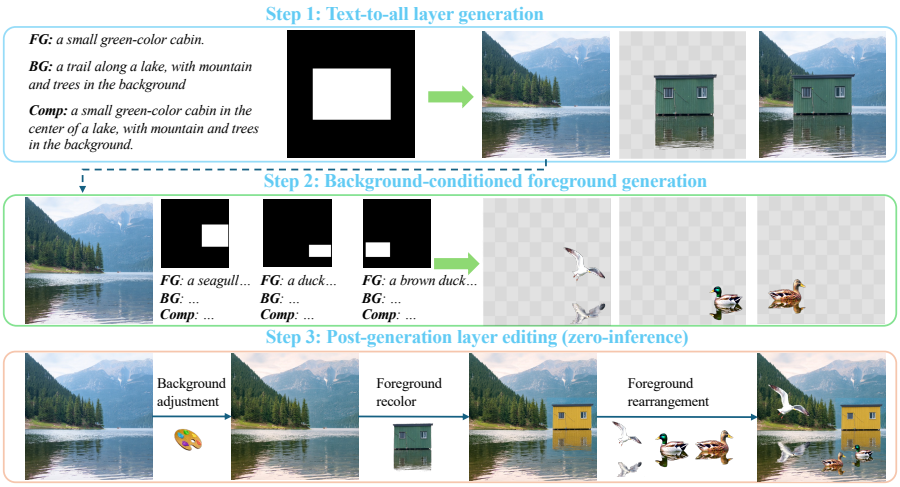

Figure 8: **Diverse multi-object creative applications driven by our model.** We leverage both Text2All and FG_Gen modes to jointly guide the synthesis process, unlocking a broader range of editing possibilities and producing diverse, visually appealing results.

4.5 CREATIVE MULTI-OBJECT EDITING APPLICATIONS

Our framework supports generating multiple layers with realistic visual effects under three different generation modes, enabling natural post-generation layer editing, as shown in Fig. 1 and Fig. 5. In this section, we further demonstrate that, benefiting from the highly consistent and visually coherent results across different modes, our framework allows flexible cross-mode collaborative editing, unlocking a broader range of creative possibilities. As illustrated in Fig. 8, combining the Text2All and FG_Gen modes yields diverse and harmonized editing outcomes.

5 CONCLUSION AND DISCUSSION

In this work, we present LASAGNA, a novel and unified framework for controllable layered image generation and editing. To facilitate research in this direction, we introduce the LASAGNA-48K dataset and the LASAGNABENCH evaluation benchmark. A core advantage of our approach is the explicit binding of high-fidelity, context-aware visual effects to the FG layer. This design is highly practical for real-world applications, addressing the prevalent need for single-object composition and local editing—such as translation and resizing—without requiring model re-inference. Within this common "FG over BG" setting, LASAGNA significantly outperforms existing methods in preserving object identity and maintaining perceptual realism. Extending our framework to dynamically adapt visual effects across complex multi-object interactions and intricate scene geometries remains an exciting direction for future work.

AI USE STATEMENT

We used generative AI tools to polish the manuscript's language and improve clarity. These writing tools were not used to derive scientific conclusions or create schematic figures or plots. The use of AI models for image generation, data curation, captioning, instruction generation, and automated evaluation is documented in the methodology and experimental sections. The authors reviewed the AI-assisted writing and take full responsibility for the paper's content, results, conclusions, and figures.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *OpenAI blog*, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *ECCV*, 2024a.

Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12501–12511, 2025a.

Xiaokang Chen, Chengyue Wu, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Yusuf Dalva, Yijun Li, Qing Liu, Nanxuan Zhao, Jianming Zhang, Zhe Lin, and Pinar Yanardag. Layerfusion: Harmonized multi-layer text-to-image generation with generative priors. *arXiv preprint arXiv:2412.04460*, 2024.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024a.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024b.

Alessandro Fontanella, Petru-Daniel Tudosiu, Yongxin Yang, Shifeng Zhang, and Sarah Parisot. Generating compositional scenes via text-to-image rgba instance generation. *Advances in Neural Information Processing Systems*, 37:43864–43893, 2024.

Shixuan Gao, Pingping Zhang, Tianyu Yan, and Huchuan Lu. Multi-scale and detail-enhanced segment anything model for salient object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9894–9903, 2024.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arxiv:2404.14396*, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Dingbang Huang, Wenbo Li, Yifei Zhao, Xinyu Pan, Yanhong Zeng, and Bo Dai. Psdiffusion: Harmonized multi-layer image generation via layout and appearance alignment. *arXiv preprint arXiv:2505.11468*, 2025a.

Junjia Huang, Pengxiang Yan, Jinhang Cai, Jiyang Liu, Zhao Wang, Yitong Wang, Xinglong Wu, and Guanbin Li. Dreamlayer: Simultaneous multi-layer generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3357–3366, 2025b.

Hugging Face. Flux family models. `https://huggingface.co/docs/diffusers/main/en/api/pipelines/flux`, 2025.

Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pp. 150–168. Springer, 2024.

Kyoungkook Kang, Gyujin Sim, Geonung Kim, Donguk Kim, Seungho Nam, and Sunghyun Cho. Layeringdiff: Layered image synthesis via generation, then disassembly with generative knowledge. *arXiv preprint arXiv:2501.01197*, 2025.

Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *arXiv preprint arXiv:2505.09358*, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Black Forest Labs. Flux, 2024. URL `https://github.com/black-forest-labs/flux`.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

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*, pp. 740–755. Springer, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025a.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025b.

OpenAI. Gpt image 1. `https://platform.openai.com/docs/models/gpt-image-1`, 2025. Accessed: 2025-11-13.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Yifan Pu, Yiming Zhao, Zhicong Tang, Ruihong Yin, Haoxing Ye, Yuhui Yuan, Dong Chen, Jianmin Bao, Sirui Zhang, Yanbin Wang, et al. Art: Anonymous region transformer for variable multi-layer transparent image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7952–7962, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Petru-Daniel Tudosiu, Yongxin Yang, Shifeng Zhang, Fei Chen, Steven McDonagh, Gerasimos Lampouras, Ignacio Iacobacci, and Sarah Parisot. Mulan: A multi layer annotated dataset for controllable text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22413–22422, 2024.

Unsplash. Unsplash: Free high-resolution photos. `https://unsplash.com/`.

Tianyu Wang, Xiaowei Hu, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection with a single-stage detector. *IEEE transactions on pattern analysis and machine intelligence*, 45(3): 3259–3273, 2022.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arxiv:2409.18869*, 2024.

Runpu Wei, Zijin Yin, Shuo Zhang, Lanxiang Zhou, Xueyi Wang, Chao Ban, Tianwei Cao, Hao Sun, Zhongjiang He, Kongming Liang, et al. Omnieraser: Remove objects and their effects in images with paired video-frame data. *arXiv preprint arXiv:2501.07397*, 2025.

Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. In *European Conference on Computer Vision*, pp. 112–129. Springer, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arxiv:2408.12528*, 2024.

Jinrui Yang, Qing Liu, Yijun Li, Soo Ye Kim, Daniil Pakhomov, Mengwei Ren, Jianming Zhang, Zhe Lin, Cihang Xie, and Yuyin Zhou. Generative image layer decomposition with visual effects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7643–7653, 2025a.

Siwei Yang, Mude Hui, Bingchen Zhao, Yuyin Zhou, Nataniel Ruiz, and Cihang Xie. Complexedit: Cot-like instruction generation for complexity-controllable image editing benchmark. *arXiv preprint arXiv:2504.13143*, 2025b.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025.

Shengming Yin, Zekai Zhang, Zecheng Tang, Kaiyuan Gao, Xiao Xu, Kun Yan, Jiahao Li, Yilei Chen, Yuxiang Chen, Heung-Yeung Shum, et al. Qwen-image-layered: Towards inherent editability via layer decomposition. *arXiv preprint arXiv:2512.15603*, 2025.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26125–26135, 2025.

Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *arXiv preprint arXiv:2402.17113*, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023a.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Xinyang Zhang, Wentian Zhao, Xin Lu, and Jeff Chien. Text2layer: Layered image generation using latent diffusion model. *arXiv preprint arXiv:2307.09781*, 2023b.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025.

Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024.

Jixin Zhao, Shangchen Zhou, Zhouxia Wang, Peiqing Yang, and Chen Change Loy. Objectclear: Complete object removal via object-effect attention. *arXiv preprint arXiv:2505.22636*, 2025.

Yichuan Zhong, Yapeng Tian, et al. Tp-blend: Textual-prompt attention pairing for precise object-style blending in diffusion models. *Transactions on Machine Learning Research*, 2025.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arxiv:2408.11039*, 2024.

Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision*, pp. 195–211. Springer, 2024.

Appendix

A VIDEO DEMO

We provide a demonstration video, *Video Demo*, in the Supplementary Material, which offers an intuitive visualization of how our three generation modes operate in practice.

B LASAGNA FRAMEWORK: DETAILED ARCHITECTURE

We provide additional details on how heterogeneous image and text inputs are tokenized, tagged with our task-aware embeddings, and assembled into the single sequence consumed by the DiT backbone. Throughout this section we follow the notation introduced in Table 2: an image $\mathbf{u} \in \{\mathbf{x}_t^{\text{comp}}, \mathbf{x}_t^{\text{bg}}, \mathbf{x}_t^{\text{fg+ve}}, \mathbf{c}_{\text{bg}}, \mathbf{c}_{\text{fg}}, \mathbf{c}_{\text{mask}}\}$ is referred to as a *frame*, and the model processes a variable-length set of F such frames per sample.

B.1 IMAGE TOKENIZATION

All image inputs (both conditional and noisy targets) live in a shared latent space provided by a frozen RGBA VAE encoder $\mathcal{E}_{\text{VAE}}$. Given a raw frame $\mathbf{u}$, we obtain

$$\mathbf{z} \;=\; \mathcal{E}_{\text{VAE}}(\mathbf{u}) \;\in\; \mathbb{R}^{C\times H\times W}, \tag{1}$$

including the mask condition $\mathbf{c}_{\text{mask}}$, which we encode with the same VAE so that all conditioning lives in the same latent manifold. Each frame latent is then patchified and linearly projected to the model dimension D by an embedder $\mathbf{W}_x$, yielding N_{img} tokens per frame. Stacking the F frames gives the image token block $\mathbf{H}_{\text{img}} \in \mathbb{R}^{FN_{\text{img}}\times D}$.

B.2 FOUR EMBEDDINGS: TYPE, IO, POSITION, TIMESTEP

Because all frames—inputs and noisy targets, BG/FG/composite/mask—are flattened into a single token stream, the network must be told (i) what each frame *is*, (ii) whether it is being conditioned on or denoised, and (iii) where each token sits in space and in the diffusion process. We therefore augment every image token belonging to frame $f \in \{1,\dots,F\}$ with four embeddings.

Type Embedding $\mathbf{e}^{\text{type}}$. This is the central mechanism that lets a single network reason jointly over the BG, FG, composite, and mask layers. We assign each frame a discrete type $\tau_f \in \mathcal{T} = \{\text{ASSET}, \text{CANVAS}_{\text{bg}}, \text{CANVAS}_{\text{comp}}, \text{CANVAS}_{\text{fg+ve}}, \text{CONTROL}\}$, where ASSET marks a clean FG reference, the three CANVAS variants enumerate the BG, composite, and FG-with-visual-effect canvases, and CONTROL marks the mask frame. A learned embedding $\phi_{\text{type}}\colon \mathcal{T} \to \mathbb{R}^D$ is then broadcast over all tokens of frame f:

$$\mathbf{e}^{\text{type}}_{f,i} \;=\; \phi_{\text{type}}(\tau_f). \tag{2}$$

Without this embedding, the transformer cannot tell apart frames of different semantic roles—*e.g.*, the target composite $\mathbf{x}_t^{\text{comp}}$ vs. the conditional background $\mathbf{c}_{\text{bg}}$.

IO Embedding $\mathbf{e}^{\text{io}}$. This embedding tells the model whether a frame is an *input* ($\mathbf{c} \in \mathbf{C}$) to be preserved or an *output* ($\mathbf{x} \in \mathbf{X}_t$) to be denoised. We assign each frame an indicator $\rho_f \in \{\text{INPUT}, \text{OUTPUT}\}$, and a learned ϕ_{io} produces

$$\mathbf{e}^{\text{io}}_{f,i} \;=\; \phi_{\text{io}}(\rho_f). \tag{3}$$

Together with $\mathbf{e}^{\text{type}}$, this single scalar flag is the only difference between the three modes in Table 2: under FG_GEN, the BG frame carries ρ=INPUT; under BG_GEN, the FG frame does; and under TEXT2ALL, all image frames carry ρ=OUTPUT. By varying the assignment $\{(\tau_f, \rho_f)\}_{f=1}^{F}$, the same set of weights learns to support every editing mode without any architectural change.

Spatial–Temporal Position Embedding $\mathbf{e}^{\text{pos}}$. A key technical ingredient of LASAGNA is a unified *3D rotary position embedding* (RoPE) (Su et al., 2024) over the joint $(\text{height}, \text{width}, \text{frame})$ axes. Concretely, for a token at spatial location (h, w) within frame index f, we apply RoPE on the Q, K projections of every attention head:

$$\mathbf{e}^{\text{pos}}_{f,h,w} \;=\; \text{RoPE}_{\text{3D}}\big(h,\, w,\, \eta_f\big), \tag{4}$$

where η_f is a per-frame index. Adding the frame index as an explicit positional axis lets the model preserve the 2D spatial structure within each frame while still knowing which layer (BG, FG, composite, mask) each token belongs to, so cross-frame correspondences can be consistently modeled in attention.

Timestep Embedding $\mathbf{e}^{\text{time}}$. The diffusion step t is encoded by a sinusoidal positional encoding followed by an MLP, $\mathbf{e}^{\text{time}}(t) \in \mathbb{R}^D$, and broadcast over all visual tokens.

The Type, IO, and Timestep embeddings are aggregated into a per-token modulation

$$\mathbf{m}_{f,i} \;=\; \text{Norm}\big(\mathbf{e}^{\text{time}}(t) + \mathbf{e}^{\text{type}}_{f,i} + \mathbf{e}^{\text{io}}_{f,i}\big), \tag{5}$$

which is fed to every DiT block to modulate its computation, while $\mathbf{e}^{\text{pos}}$ is injected inside attention via 3D RoPE.

B.3 Sequence Assembly and DiT Forward Pass

Let $\mathbf{H}_{\text{cond}} \in \mathbb{R}^{L_{\text{cond}} \times D}$ denote the condition tokens encoded from the prompt $\mathbf{c}_{\text{txt}}$. We concatenate them with the frame-wise image tokens into a single sequence

$$\mathbf{S} \;=\; \big[\, \mathbf{H}_{\text{cond}};\; \mathbf{H}^{(1)}_{\text{img}};\; \ldots;\; \mathbf{H}^{(F)}_{\text{img}} \,\big], \tag{6}$$

which is processed by a stack of DiT blocks with full bidirectional self-attention—so that every output frame jointly attends to all conditional frames and tokens. The output tokens for frames flagged ρ_f=OUTPUT are then mapped to the noise prediction $\boldsymbol{\epsilon}_\theta$ used in the flow-matching loss.

Why Type and IO Embeddings Matter. Our set-to-set framework draws from a shared frame pool—mask, clean FG asset, FG+VE, BG, and composite—with input/output roles varying across generation modes. Because clean conditions and noisy targets are processed within a single attention sequence, the model must distinguish both what each frame represents and whether it is given or generated. We found that positional embeddings alone were insufficient to jointly train the three modes, motivating explicit Type and IO embeddings.

These embeddings encode complementary distinctions, and neither substitutes for the other. In TEXT2ALL, the BG, FG+VE, and composite frames all carry IO = OUTPUT; Type distinguishes their semantic roles. Conversely, a BG frame retains the same Type whether it is provided in FG_GEN or generated in BG_GEN and TEXT2ALL; IO distinguishes conditioning from generation. The IO embedding therefore serves all three modes, including identifying the clean FG asset as an input in BG_GEN. Together, Type and IO embeddings are essential to our unified training setup and enable task switching through frame-role assignments, without changing model weights or architecture.

C Extensibility to Diverse Tasks

Building on the unified and extensible framework introduced in Section 3.1 and Appendix B, our approach can be readily extended to tasks beyond the three considered in the main paper. For example, as shown in Fig. 9, the same framework supports composite image generation conditioned on a FG–BG pair by simply modifying the IO and Type IDs. This example demonstrates how our unified interface naturally accommodates different input/output set sizes while enabling capability reuse across tasks.

D Human Evaluation

Setup and criteria. Eight independent researchers evaluated FG_Gen and BG_Gen with model identities anonymized. For each task, we randomly sampled ten images from each of the six

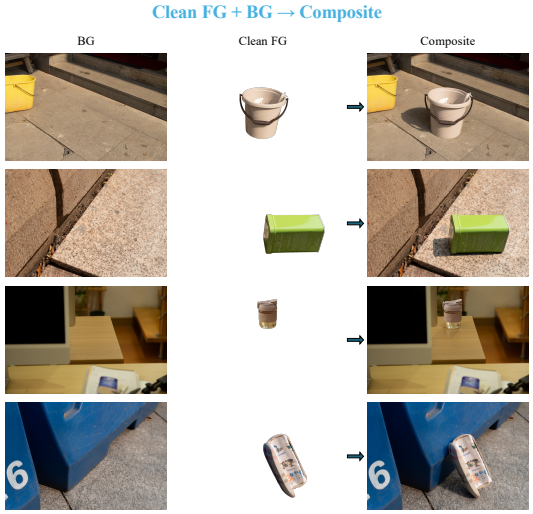

Figure 9: **Clean-FG-plus-BG composition with the configurable frame interface.** A supplied BG and clean FG asset are combined into a generated composite, with scene-dependent integration and visual effects. This additional mode reuses the same model by changing frame-role assignments.

LASAGNABENCH sources, yielding 60 images per task and 120 task–image instances. After excluding incomplete decisions, we retained 1,878 criterion-level votes: 987 for FG_Gen and 891 for BG_Gen.

Annotators independently assessed **Instruction Following (IF)**: compliance with text and task-specific spatial constraints; **Identity Preservation (IP)**: preservation of the original BG in FG_Gen or FG object in BG_Gen; and **Harmonization and Quality (HQ)**: natural integration of boundaries, lighting, perspective, and shadows, together with perceptual quality and artifacts.

Protocol and score. For each image and criterion, annotators selected the best and worst outputs among four models. Multiple selections were allowed for indistinguishable outputs. For decision r, let B_r and W_r denote the best and worst sets. Ties share one unit of weight:

$$s(m,r) = \frac{\mathbf{1}[m \in B_r]}{|B_r|} - \frac{\mathbf{1}[m \in W_r]}{|W_r|}, \qquad \mathrm{BWS}(m) = \frac{100}{N} \sum_{r=1}^{N} s(m,r), \tag{7}$$

where N is the number of valid decisions for a task and criterion. Best–Worst Score (BWS) ranges from -100 to 100; higher is better. Task Avg is the equally weighted mean of IF, IP, and HQ. The FG_Gen counts for IF/IP/HQ are 341/327/319, and the BG_Gen counts are 289/301/301.

Table 6: **Overall blinded user study on all six benchmark sources.** BWS scores (↑); bold indicates the highest value within each task.

Task	Model	IF	IP	HQ	Task Avg
FG_Gen	LASAGNA	**+61.3**	**+39.6**	**+17.5**	**+39.5**
	FLUX	−7.3	+36.5	−17.5	+3.9
	Qwen	−13.7	+16.5	+0.9	+1.2
	GPT	−40.3	−92.6	−0.9	−44.6
BG_Gen	LASAGNA	**+38.6**	**+50.2**	+23.8	**+37.5**
	Qwen	−25.4	+18.9	+4.5	−0.7
	GPT	+28.7	−87.8	**+32.3**	−8.9
	FLUX	−41.9	+18.7	−60.5	−27.9

Results. LASAGNA obtains the highest Task Avg in both tasks, exceeding the best baseline by 35.6 points in FG_Gen and 38.2 in BG_Gen. It leads all three FG_Gen criteria and the IF/IP criteria of BG_Gen. GPT has the highest BG_Gen HQ, but its low IP score indicates a trade-off between perceptual integration and preservation of the supplied object. These human judgments complement the automated metrics by directly assessing the practical balance of control, identity, and visual quality.

E MORE EXPERIMENTS

E.1 COMPARISON ON PUBLIC BENCHMARKS

While primarily designed for the novel task of layer generation, we further investigate its versatility by adapting it to standard public benchmarks: ImgEdit-Bench (Ye et al., 2025) and GenEval (Ghosh et al., 2023), aligning our generation modes (FG_Gen, BG_Gen, and Text2All) with their corresponding benchmark tasks ("Addition", "Background", and standard text-to-image generation).

As shown in Table 7 and Table 8, our model demonstrates strong performance on the composite images. It is important to emphasize that current SOTA methods benefit from significantly larger training corpora and model sizes, and are aligned with these standard benchmark protocols. In contrast, LASAGNA approaches these tasks without specific optimization, yet still maintains highly competitive results. Moreover, a key advantage of our approach is its ability to perform layer generation with visual effects, a capability not supported by existing models. The effectiveness of layer representations—and their clear benefits for subsequent editing quality—has been thoroughly validated in the above, highlighting a unique strength of our method.

Table 7: **Evaluation of image editing ability on ImgEdit-Bench (Ye et al., 2025).** "Addition" corresponds to FG_Gen, and "Background" corresponds to BG_Gen.

Model	Addition	Background	Model	Addition	Background
Instruct-P2P (Brooks et al., 2023)	2.45	1.44	UniWorld-V1 (Lin et al., 2025)	3.82	2.99
AnyEdit (Yu et al., 2025)	3.18	2.24	BAGEL (Deng et al., 2025)	3.81	3.39
UltraEdit (Zhao et al., 2024)	3.44	2.83	OmniGen2 (Wu et al., 2025b)	3.57	3.57
ICEdit (Zhang et al., 2025)	3.58	3.08	Kontext-dev (Labs et al., 2025)	3.83	3.98
Step1X-Edit (Liu et al., 2025a)	3.88	3.16	LASAGNA	3.86	3.32

Table 8: **Evaluation of text-to-image generation ability on GenEval (Ghosh et al., 2023) benchmark.** † refers to methods using an LLM rewriter.

Model	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
PixArt-α (Chen et al., 2024a)	0.98	0.50	0.44	0.80	0.08	0.07	0.48
Emu3-Gen (Wang et al., 2024)	0.98	0.71	0.34	0.81	0.17	0.21	0.54
DALL-E 3 (Betker et al., 2023)	0.96	0.87	0.47	0.83	0.43	0.45	0.67
SD3-Medium (Esser et al., 2024a)	0.99	0.94	0.72	0.89	0.33	0.60	0.74
FLUX.1-dev† (Labs, 2024)	0.98	0.93	0.75	0.93	0.68	0.65	*0.82*
SEED-X (Ge et al., 2024)	0.97	0.58	0.26	0.80	0.19	0.14	0.49
Transfusion (Zhou et al., 2024)	-	-	-	-	-	-	0.63
Show-o (Xie et al., 2024)	0.98	0.80	0.66	0.84	0.31	0.50	0.68
Janus-Pro-7B (Chen et al., 2025b)	0.99	0.89	0.59	0.90	0.79	0.66	0.80
MetaQuery-XL† (Pan et al., 2025)	-	-	-	-	-	-	0.80
BAGEL† (Deng et al., 2025)	0.98	0.95	0.84	0.95	0.78	0.77	0.88
LASAGNA†	0.99	0.97	0.78	0.83	0.74	0.65	0.83

E.2 ABLATIONS

We ablate our design choices in Table 9, ensuring fair and interpretable comparisons and highlighting the effectiveness of our complete LASAGNA framework. Comparing the model variant trained solely on internal data with LASAGNA, which is trained on the same internal data augmented with our public LASAGNA-48K, we observe consistent performance gains across all generation modes. This demonstrates the effectiveness and value of the newly proposed LASAGNA-48K. To examine the impact of language formulation, we convert the captions into instruction-based prompts. The resulting variant achieves comparable or slightly lower performance than LASAGNA, indicating that our framework is robust to different language input formats. We further train three independent models, each dedicated to a single generation mode. Unified LASAGNA still achieves superior performance, suggesting that joint training enables beneficial knowledge sharing and synergy across the different generation tasks.

Table 9: **Ablation study.** Each cell reports **CLIP-FID / FID** (↓). Training with LASAGNA-48K improves over the internal-only variant, confirming its benefit. The instruction-based setting yields comparable results, showing robustness to language variation. The unified model outperforms single-task models, indicating synergy across generation modes.

Ablation	FG_Gen		BG_Gen			Text2All		
	Composite	FG	Composite	BG	FG	Composite	BG	FG
Internal data	11.3/79.7	28.9/155.9	15.9/102.9	17.8/134.0	22.5/145.2	19.6/120.4	18.1/136.2	28.8/158.0
Instruction	10.6/77.1	27.4/153.1	15.2/107.0	17.2/134.8	**22.6**/140.9	18.0/119.5	**16.8**/134.3	**26.6**/152.5
Separate task	11.0/79.7	28.4/156.7	15.6/104.1	17.8/133.6	37.2/142.0	19.2/119.6	18.4/133.4	28.6/153.3
LASAGNA	**10.3/75.9**	**27.3/151.8**	**14.6/102.9**	**16.8/132.5**	22.7/**139.3**	**17.8/119.4**	17.1/**129.3**	28.1/**150.3**

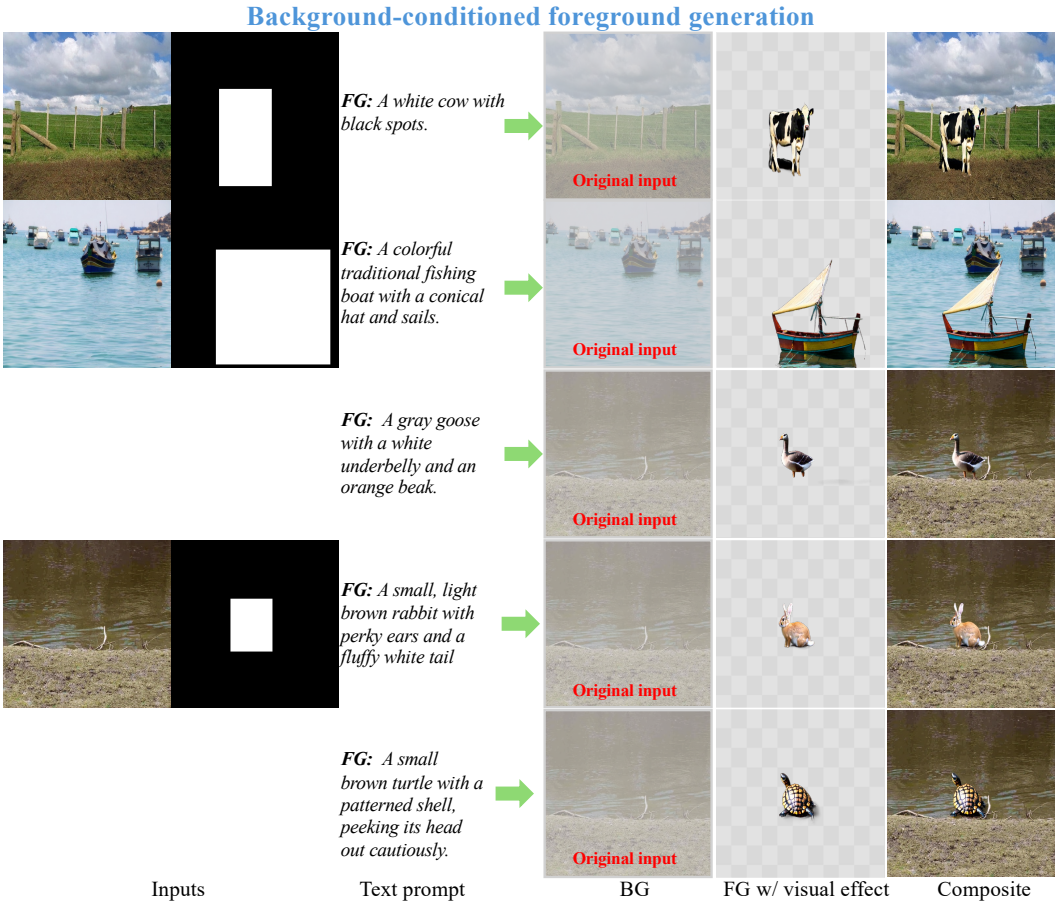

Figure 10: **More results from LASAGNA under the background-conditioned foreground generation (FG_Gen) mode.**

F MORE VISUALIZATION RESULTS FROM LASAGNA

F.1 MORE QUALITATIVE RESULTS FROM LASAGNA UNDER THREE GENERATION MODES

As shown in Fig. 10, Fig. 11, and Fig. 12, we provide more results of our model under three different modes (FG_Gen, BG_Gen, and Text2All).

Specifically, in Fig. 10 and Fig. 11, for the FG_Gen and BG_Gen modes, we further demonstrate more flexible applications. We can fix the background and generate different foregrounds, or fix the foreground and generate different backgrounds.

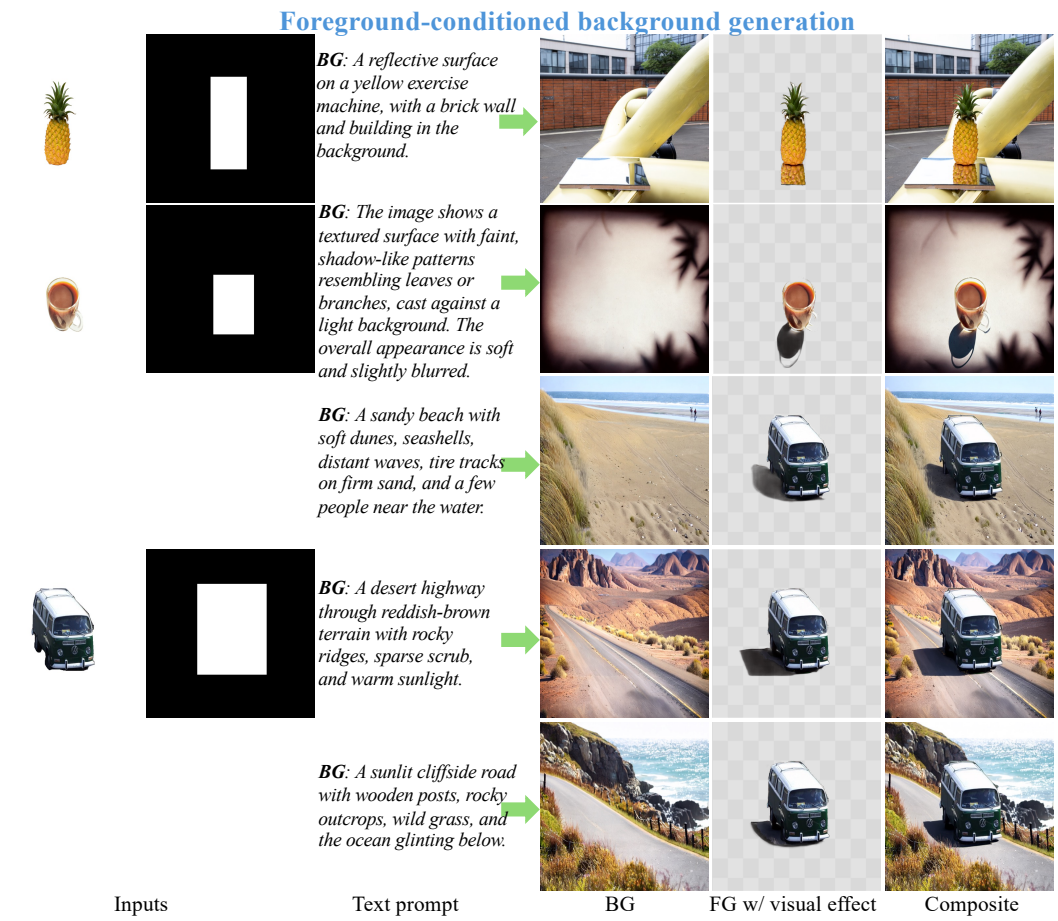

Figure 11: **More results from LASAGNA under the foreground-conditioned background generation (BG_Gen) mode.**

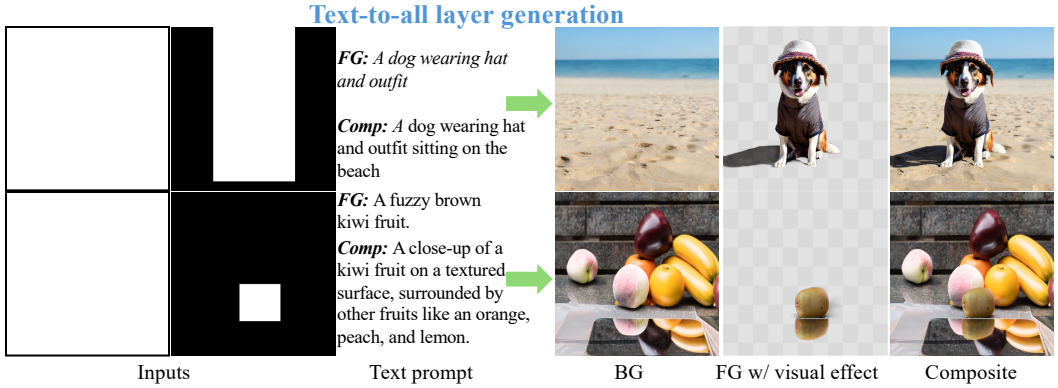

Figure 12: **More results from LASAGNA under the text-to-all layer generation (Text2All) mode.**

F.2 FLEXIBILITY OF THE FOREGROUND-CONDITIONED BACKGROUND GENERATION (BG_GEN) MODE

Our primary use case assumes that user-provided foregrounds (FGs) are clean, i.e., without visual effects (VEs). This assumption aligns with most real-world editing workflows, where extracted assets typically lack physically consistent visual effects such as shadows, reflections, or lighting interactions. In this setting, the key advantage of our model is its ability to synthesize the missing effects and generate backgrounds that are physically and visually consistent with the foreground.

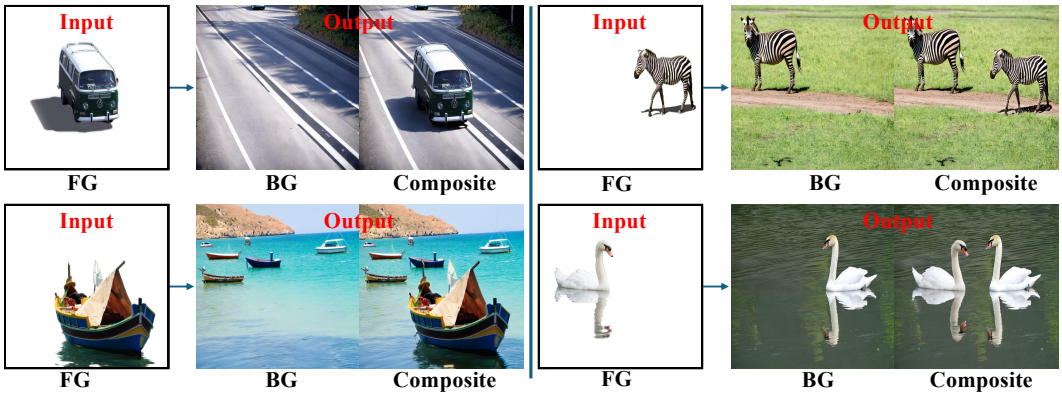

Figure 13: **Foregrounds with visual effects**. When the input foreground already contains visual effects (e.g., shadows or reflections), our model preserves these effects and generates backgrounds that remain physically and visually consistent with the provided cues.



Figure 14: **More results from LASAGNA under the Text-to-All layer generation (Text2All) mode across diverse domains.**

Nevertheless, our framework is not limited to this clean-foreground setting. When the input foreground already contains visual effects (e.g., shadows or reflections), our model is able to preserve and respect these existing effects during background generation. As illustrated in Fig. 13, the generated background adapts to the visual cues present in the input foreground, producing lighting and geometric relationships that remain consistent with the provided effects.

This flexibility arises from the fact that our model is trained on real triplets of foreground, visual effects, and background, allowing it to learn the joint distribution between background structures and visual effects. As a result, the model can both synthesize missing effects when they are absent and maintain existing effects when they are present, enabling robust behavior across diverse real-world editing scenarios.

F.3 MORE QUALITATIVE RESULTS FROM DIVERSE DOMAINS

As shown in Fig. 14, we further present results on more diverse domains, including *sketch*, *watercolor*, *cartoon*, and *cyberpunk* styles. Although our LASAGNA-48K dataset is composed of real-world imagery, our method performs well even in these stylistically distinct domains. We attribute this generalization ability to the physical knowledge embedded in our real-world training data, which enables the model to effectively transfer the knowledge learned by the pretrained model to other domains.

G HARD CASES

To explore the limits of our model's capabilities, we further evaluate FG_Gen on a diverse set of challenging cases. As shown in Fig. 15. We examines FG_Gen near object/support boundaries, on uneven surfaces such as wavy water, and across strong lighting changes or existing shadows. These results demonstrate that our model can produce plausible generations even under such challenging conditions.

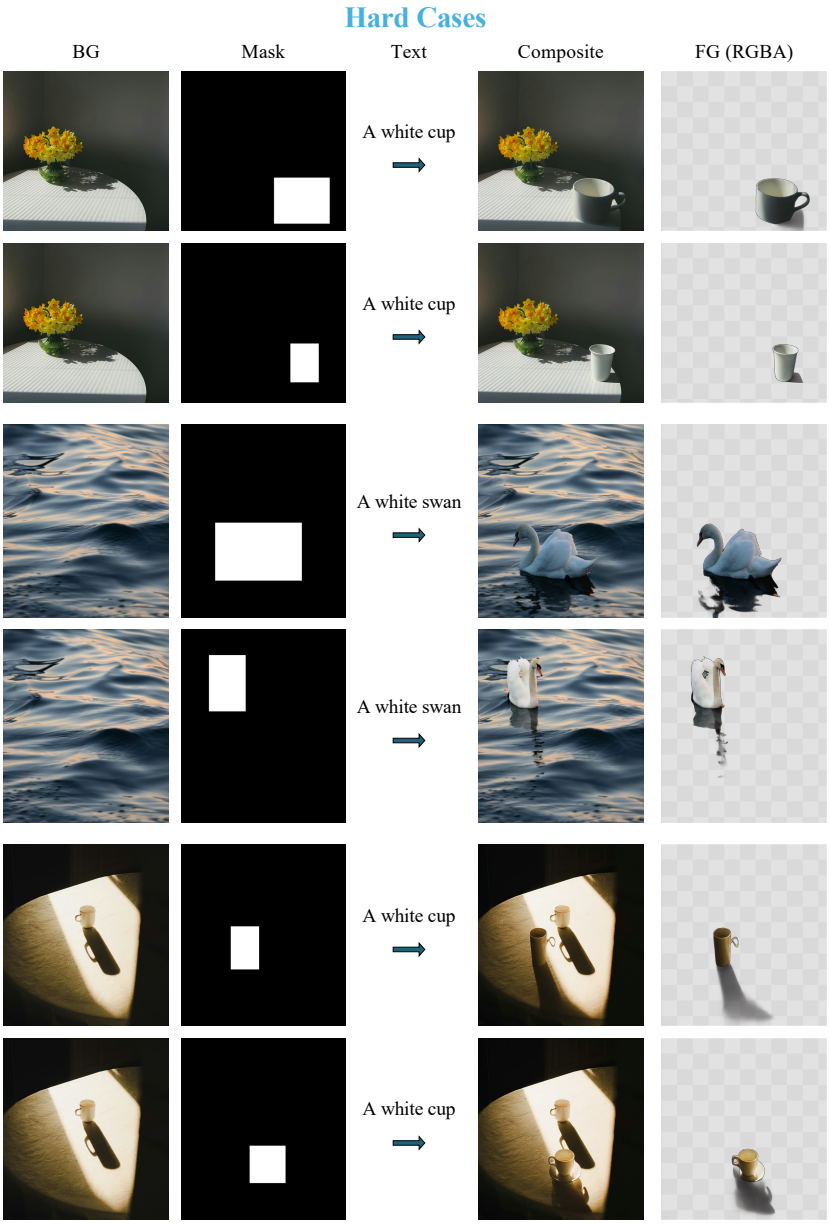

Figure 15: **Qualitative hard cases for FG_Gen.** Top: object placement near the edge of a supporting surface. Middle: foregrounds and reflections on uneven water. Bottom: placement under strong lighting changes and existing cast shadows.

H MORE SAMPLES FROM LASAGNA-48K AND LASAGNABENCH

LASAGNABENCH comes from 6 different data sources, with the specific quantity distribution shown in Table 10.

Table 10: **Benchmark statistics.** LASAGNABENCH is built from 6 distinct sources—four public datasets (Tudosiu et al., 2024; Lin et al., 2014; Wang et al., 2022; Unsplash) and two in-house data—to ensure diversity and representativeness.

Data Source	MULAN	COCO 2017	SOBA	Unsplash	Camera-Indoor	Camera-Outdoor
Num. Images	45	40	50	27	40	40

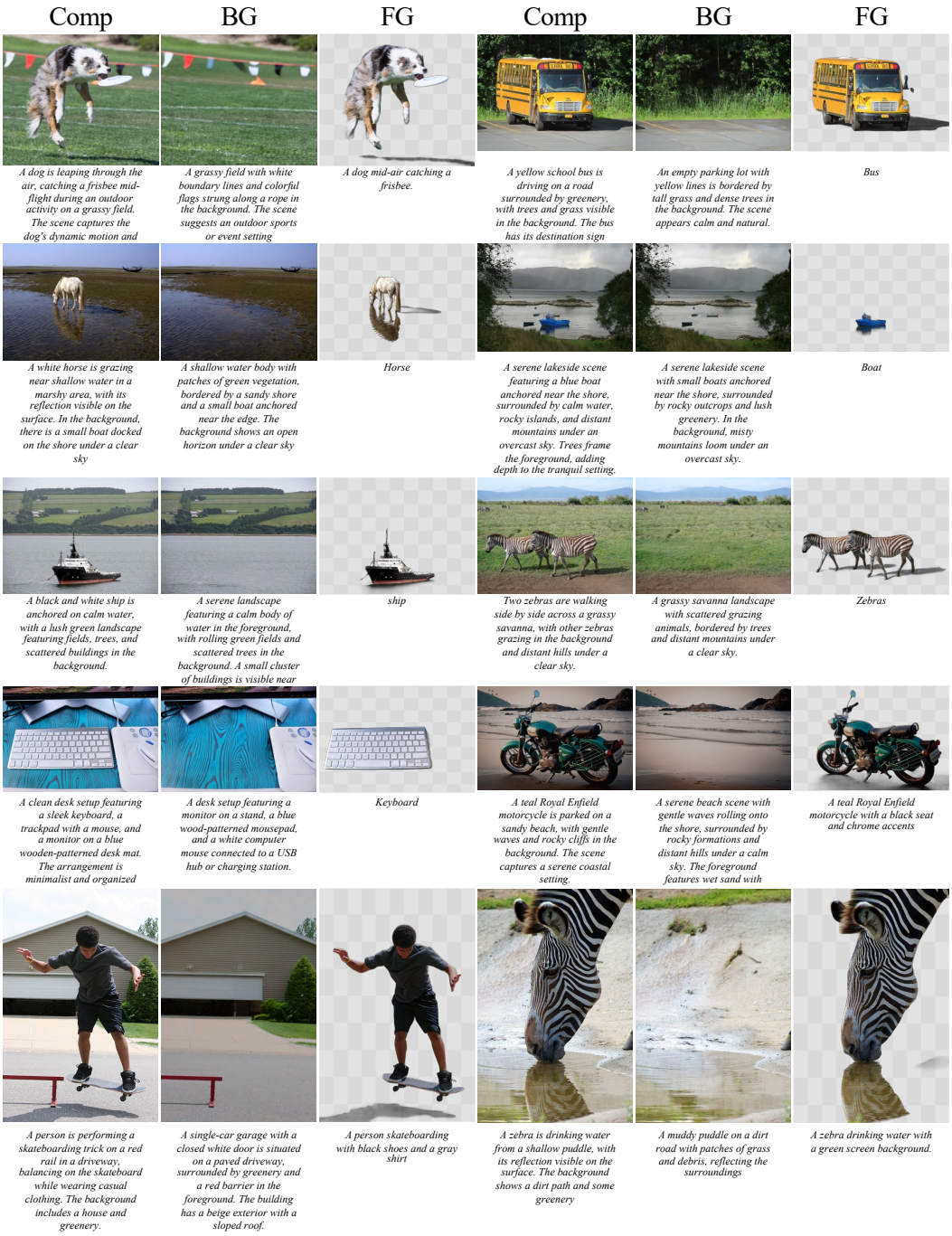

Figure 16: **More samples of LASAGNA-48K.** Each sample consists of a composite image, a clean background, and a foreground layer with visual effects, along with corresponding captions for all components.

As shown in Fig. 16, we provide more samples from LASAGNA-48K.

As shown in Fig. 17, we provide more samples from LASAGNABENCH.

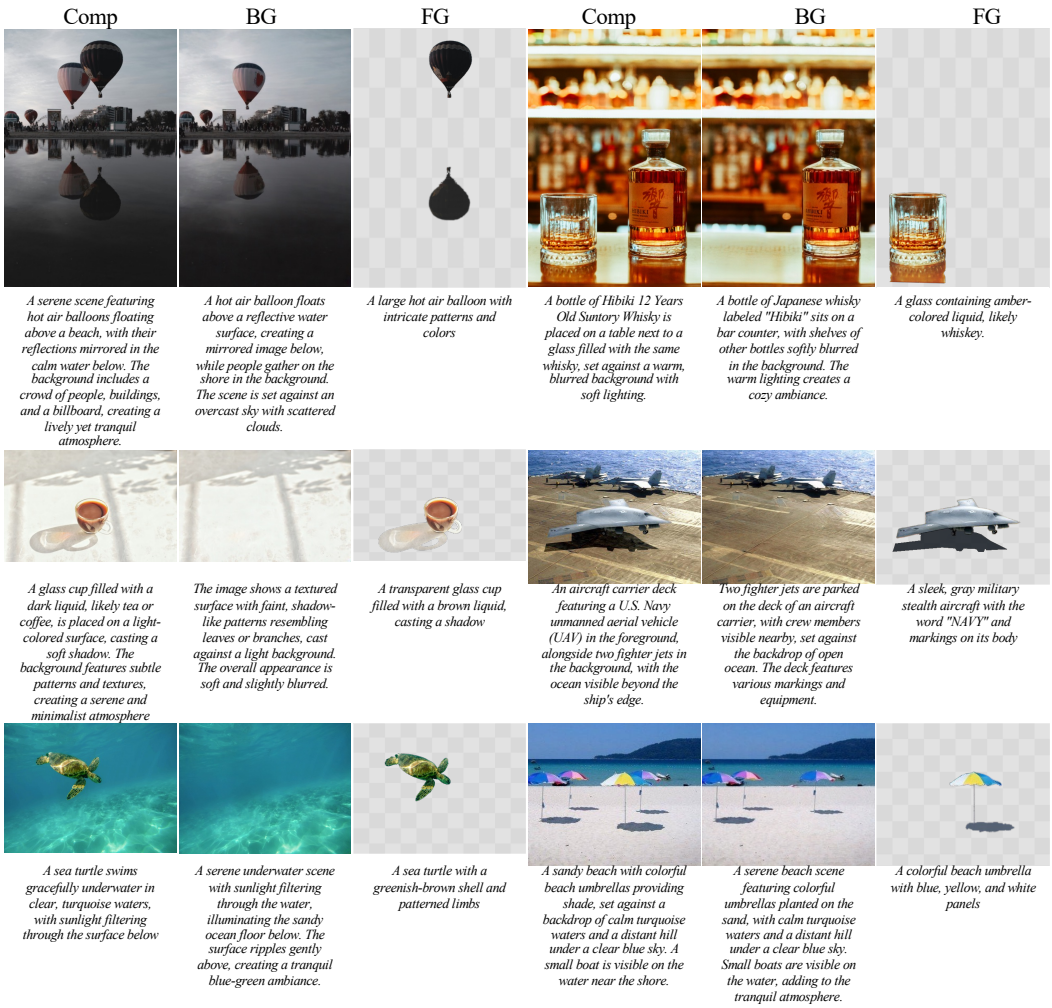

Figure 17: **More samples of LASAGNABENCH.** Each sample consists of a composite image, a foreground layer with visual effects, and a clean background along with corresponding captions for all components. **Foreground with visual effects annotated by professional annotators. Background are decomposed by expert models or captured by camera.**

I DETAILED SCORES OF THE GPT SCORE

As shown in Table 3 in the main manuscript, we provide the "GPT Score", which is the average score of the instruction-following and identity-preserving metrics proposed by Complex-Edit (Yang et al., 2025b). Here, we additionally provide the original score in Table 11. We run the same prompt three times and take the average as the original score for instruction-following and identity-preserving.

J DETAILS OF THE METRIC IN SECTION 4.4 (THE NECESSITY OF LAYERED GENERATION WITH VISUAL EFFECTS)

To illustrate the necessity of our proposed generation paradigm (Layer Editing with Visual Effects), we conduct quantitative experiments in Section 4.4 of the main manuscript. The results show the superiority of our generation paradigm.

We compare three editing modes: *Instruct Editing*, *Cascaded Layer Editing*, and *Layer Editing with Visual Effects*. For both *Cascaded Layer Editing* and *Layer Editing with Visual Effects*, the foreground is represented in RGBA format, which allows us to perform programmatic, pixel-accurate modifications. This also ensures that the editing parameters remain fully consistent across the two modes.

Table 11: **GPT scores.** "FG_Gen" denotes background-conditioned foreground layer generation, "BG_Gen" denotes foreground-conditioned background generation, and "Text2All" denotes text-to-all layer generation. "IF" stands for Instruction Following and "IP" for Identity Preservation. Results for models marked with $*$ are obtained using their respective expert models rather than a single unified model. Specifically, for the FG_Gen and BG_Gen tasks, we use the FLUX.1-Fill-dev, Qwen-Image-Edit-2509, and gpt-image-1[high] editing models, respectively. For the All_Gen task, we use the FLUX.1-schnell, Qwen-Image, and gpt-image-1[high] models as text-to-image models, respectively.

Model	FG_Gen			BG_Gen			Text2All		
	IF ↑	IP ↑	Avg	IF ↑	IP ↑	Avg	IF ↑	IP ↑	Avg
gpt-image-1[high]*	9.77	7.88	8.8	9.84	7.95	8.9	6.95	7.26	7.1
FLUX.1*	8.32	9.46	8.9	8.43	8.48	8.5	6.46	5.61	6.0
Qwen-Image-Edit*	8.66	9.34	9.0	7.25	8.46	7.9	6.41	4.98	5.7
Ours	9.27	9.29	9.3	9.34	8.67	9.0	8.31	6.91	7.6

Template of Instruction Editing

Recolor

1.Red: Change the {Object} bright red.
2.Orange: Change the {Object} to vibrant orange.
3.Yellow: Change the {Object} sunny yellow.
4.Green: Change the {Object} intense green.
5.Cyan: Change the {Object} to vivid cyan.
6.Blue: Change the {Object} deep blue.
7.Purple: Change the {Object} rich purple.

Spatial editing

Magnitude: {20,30,50}
Intensity: {slightly, moderately, significantly}
Direction: {to the left, to the right, upwards, downwards}

↓

Move {Object} {Direction}. The movement should be {Intensity}, which corresponds to an approximate displacement of {Magnitude}% relative to the image size

Figure 18: When performing instruct editing, the content within {} will be replaced by specific parameters. In spatial editing, the values of the Magnitude and Intensity parameters correspond one-to-one from left to right.

We benchmark these three approaches on recoloring, spatial editing, and complex compositional editing tasks. For recoloring, we define seven random color transformation operations. For spatial editing, we randomly select a movement direction (up, down, left, or right) and apply one of three displacement magnitudes (20%, 30%, or 50%). For compositional editing, we randomly combine recoloring and spatial editing to test multi-factor control. All evaluations are conducted automatically to ensure objective and reproducible comparison across methods.

For *Instruct Editing* mode, when performing recolor and spatial editing, we input the editing parameters from the previous Layer Editing into GPT-5. Based on the context of our question and the value of the parameter, GPT-5 generates an appropriate natural language description, as shown in the template in Fig. 18. In the subsequent Complex Editing task, we combine the two types of instructions accordingly. This helps ensure that all three types of editing perform the same actions as much as possible, thereby ensuring comparability.

K Training Samples of Data Curator

In Section 3.2, LASAGNA-48K Dataset in the main manuscript, we demonstrate the complete data construction pipeline. The data curator is a key component in the data construction pipeline. To obtain high-quality data filtering results, we carefully annotated about 30K samples by humans, as shown in Fig. 19, to train the data curator. Each sample is a triplet data: a composite image, mask and a background without the foreground, which are input into the data curator. The curator outputs a confidence score (between 0 and 1) indicating the probability that the background is good.

In manual annotation, we provide a binary label for each data sample. Specifically, we determine the final binary label based on the following three aspects:

- **Hole filling**: Whether the object has been successfully removed.
- **Background consistency**: Whether the removed area is consistent with the surrounding environment.

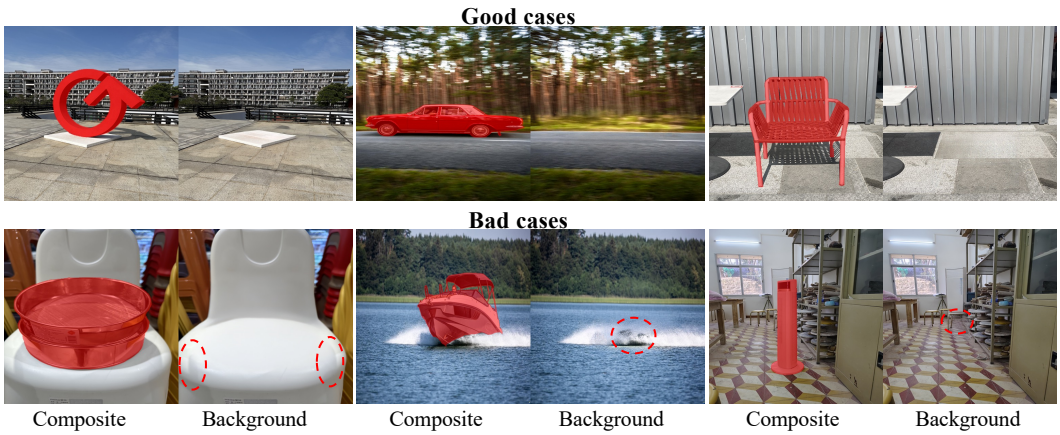

Figure 19: **Training Samples of Data Curator.** The red dashed box indicates the main problematic area in the bad cases.

- **Visual effect**: Whether the visual effects caused by the object (*e.g.*, shadow and reflection) have been removed.

If **any one** of the above criteria is not met, the sample is considered a bad case. Only when all conditions are satisfied is it labeled as a good case. We highlight these unnatural regions with dashed boxes in Fig. 19. These hard negative samples ensure that, after training the data curator, the filtered data it produces consists exclusively of high-quality background images.