CARE-Edit: Condition-Aware Routing of Experts for Contextual Image Editing

Supplementary Material

This appendix provides complete supplementary material to the main manuscript and is organized as follows:

- **Appendix A: Dataset and Implementation Details.** We detail the construction of training dataset in Appendix A.1, including the targeted design of mask-aware image pairs and the generation pipeline. We also provide a systematic comparison with existing public datasets. Appendix A.2 presents implementation specifications, including network architectures, optimization protocols, training schedules, and hyperparameter configurations in our experiments.
- **Appendix B: Extended Qualitative Comparisons.** We report additional qualitative comparisons for instruction-based (Appendix B.1) and subject-driven (Appendix B.2) editing that could not be included in main text due to space limitations, encompassing mainstream tasks such as object removal, replacement, and style transfer. For each task, we include detailed per-category and per-edit-type samples and further discuss the behavior and capabilities of CARE-Edit. We provide a project page, where we host more qualitative examples and interactive visualizations.
- **Appendix C: Additional Empirical Analysis.** We present the empirical analysis and visualizations of latent attention maps in Appendix C to validate the efficacy of specialized experts. Figure 13 demonstrates how each expert in CARE-Edit, including condition-aware routing, reference-guided subject preservation, and mask-aware control, contributes to effective task-specific learning over diffusion timesteps.

A. Dataset and Implementation Details

A.1. Training Dataset

Base corpora. To equip CARE-Edit with diverse editing capabilities, we collect data from several high-quality sources, targeting four key editing tasks: *instruction-based editing*, *object removal/replacement*, and *style transfer*. (i) Instruction-based edits are drawn from MagicBrush [55] and OmniEdit [49], which provide rich natural-language instructions paired with real-world images. (ii) and (iii) Removal and replacement samples are sourced from UNO [48] subset. (iv) Style transfer data is enriched using AnyEdit [54], which contains both the fine-grained appearance- and style-level instructions (*e.g.*, "convert to watercolor painting").

Motivation for Subjects200K. A critical limitation of purely instruction-based datasets is the spatial underspecification of edits. Models are often forced to infer the edit location solely from language, leading to ambiguity. To address this, we require a dataset with precise object masks to explicitly teach the model to align edits with spatial constraints. We thus select Subjects200K [42] as our foundation for two reasons: (i) it offers high-quality foreground masks for a diverse taxonomy of objects and humans; and (ii) reference images are captured on clean white backgrounds, which simplifies downstream composition and in-context editing. From this source, we construct a 20K subset where each sample comprises an image, a fine-grained mask, an instruction, and an optional reference image, enabling CARE-Edit to learn region-specific editing while keeping subject identity.

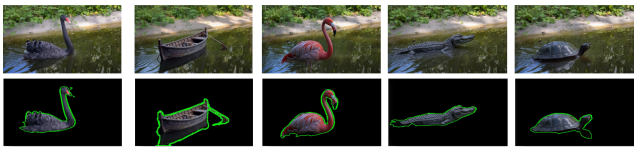

Figure 8. **Pipeline for Mask-aware Image-Pair Generation.** We use a GPT-Image-1 and VLM-based pipeline to create our training data. Starting with a reference subject on a white background, we generate diverse scene descriptions and corresponding images using an image-to-image model. This yields high-quality image pairs with consistent backgrounds but varying foregrounds, annotated with precise segmentation masks ($\mathcal{M}$) and bounding boxes ($\mathcal{B}$).

Mask-Aware Image-Pair Generation. As shown in Figure 8, we construct background-consistent image pairs with varying foregrounds using a GPT-based generation pipeline. Starting with a Subjects200K [42] reference subject, we (i) sample a scene template and a set of subjects, and (ii) query the GPT-image-1 model to synthesize images that share a consistent background but feature diverse foreground objects. (iii) Extract a high-resolution fine mask $\mathcal{M} \in \{0,1\}^{H\times W}$ for foreground using an off-the-shelf segmentation model, followed by manual filtering to ensure quality. In practice, we instantiate these templates with category names (*e.g.*, *swan*, *boat*, *flamingo*) and short descriptions of target background (*e.g.*, *"floating on a calm river near the shore"*). To improve robustness against imperfect user inputs at inference time, we also derive a coarse mask $\mathcal{B}$ in training, defined as the tight axis-aligned bounding box of the fine mask $\mathcal{M}$:

$$\Omega = \{(x,y) \mid \mathcal{M}(x,y) = 1\}, \tag{18}$$

$$x_{\min} = \min_{(x,y)\in\Omega} x, \quad x_{\max} = \max_{(x,y)\in\Omega} x, \tag{19}$$

$$y_{\min} = \min_{(x,y)\in\Omega} y, \quad y_{\max} = \max_{(x,y)\in\Omega} y, \tag{20}$$

$$\mathcal{B}(x,y) = \begin{cases} 1, & x_{\min} \le x \le x_{\max} \wedge y_{\min} \le y \le y_{\max}, \\ 0, & \text{otherwise}, \end{cases} \tag{21}$$

where $x_{\min}, x_{\max}, y_{\min}, y_{\max}$ are the extrema of coordinates in Ω. We utilize $\mathcal{M}$ for the pixel-accurate supervision (*e.g.*, boundary consistency loss) and $\mathcal{B}$ as a coarse spatial prior for the condition-aware expert routing in CARE-Edit.

Prompt Taxonomy. To systematically construct diverse yet controllable training triplets for our CARE-Edit, we organize the generation pipeline along two axes:

(1) Category and Operation. We factorize the synthetic space into (i) *Categories* (people, animals, everyday objects, stylized assets) and (ii) *Operation Types* (instruction-based, removal, replacement, style transfer, multi-subject cases).

(2) Scene-level Templates. Conditioned on a subject category, we query the LLM to generate multiple scene descriptions. Crucially, the scene prompt is constrained to a single line and constrains the generation to modify only the environment, lighting, camera view, or overall atmosphere, while strictly preserving the core subject identity or attributes. For each *Category*, we sample five such scene descriptions. This separation of subject and scene allows us to reuse identical subjects across heterogeneous contexts, and in turn to build foreground-consistent but background-varying pairs for mask-aware training. Please refer to Figure 9 for details.

(3) Task-level Templates. Given two subject descriptions a and b, a scene description s, and a style phrase p (*e.g.*, "in a retro vintage style"), we instantiate three mask-friendly templates: (i) *Replacement* (T_{rep}) keeps the background unchanged while swapping the main subject from a to b; (ii) *Addition* (T_{add}) forms a diptych where one panel contains only a and the other contains both a and b under the same scene s; (iii) *style / attribute change* (T_{sty} / T_{attr}) preserves the subject identity and scene but alters only high-level appearance attributes p; All templates explicitly ask the generator to keep background layout, lighting, and camera viewpoint as similar as possible across paired images, so that the resulting pairs differ primarily in well-localized foreground regions. This naturally aligns with our fine masks $\mathcal{M}$ and coarse boxes $\mathcal{B}$, and provides clean supervision for subject-centric, mask-aware editing. Please view Figure 10 for details.

Data Efficiency vs Previous Training Setups. Despite being trained on a significantly smaller data corpus ($\sim$ 120K triplets in total) compared to recent state-of-the-art editing and personalization baselines, CARE-Edit achieves superior performance. Table 4 reports multiple-object results on DreamBench++ [31], together with the approximate scale of the training data used by each method. The results show that CARE-Edit outperforms strong baselines such as OmniControl [43], UNO [48], and OmniGen2 [50] on all multiple-object metrics, despite relying on substantially fewer training samples. This suggests that our mask-aware, subject-centric curriculum applied in CARE-Edit and the curated multi-paired construction are substantially more data-efficient than simply scaling up instruction-only datasets, particularly for applications with complicated multi-object compositions.

Table 4. Quantitative results on the *multiple-object* subset of DreamBench++ [31]. We report three metrics along with the approximate number of training examples used by each method. The best and second-best results are marked in **bold** and underlined, respectively.

Method	#Train data	DINO-I ↑	CLIP-I ↑	CLIP-T ↑
OmniControl [43]	~$1M$	0.501	0.641	0.316
UNO [48]	~$1M$	0.508	0.649	0.303
OmniGen2 [50]	$\geq 533K$	<u>0.560</u>	<u>0.713</u>	<u>0.319</u>
CARE-Edit (Ours)	$120K$	**0.568**	**0.720**	**0.327**

A.2. Implementation Details

Backbone and Training Setup. CARE-Edit is built upon the FLUX.1 [18] variant of the Rectified Flow Transformer family. Unless otherwise specified, we select `FLUX.1-dev` as the backbone for all experiments, as it offers a good balance between visual quality and training stability in the editing setup. Following the design of OmniControl [43], we employ condition-aware modules via LoRA [13] on top of the base model, and keep the original backbone weights frozen during fine-tuning. We adopt a standard LoRA rank [13] of $r = 4$ for all attention modules, and only enable the LoRA branches [13] when processing condition-related tokens. For regular text-only tokens, the LoRA scale is set to zero so that the backbone behaves identically to the original FLUX.1 [18] model. All models are trained on 8×NVIDIA L20 GPUs, which corresponds to roughly 800 GPU hours in total. We use a per-GPU batch size of 1 with gradient accumulation over 8 steps (effective batch size of 8). For most experiments, we follow a two-stage training schedule: the model is first trained for 40K iterations on basic, single-subject samples, and then switches to complex multi-subject data for the remaining 60K iterations. We also apply EMA to the LoRA parameters [13] with a decay rate of 0.999.

Loss Functions and Hyperparameters. The total training objective $\mathcal{L}_{\text{CARE}}$ combines the standard diffusion reconstruction loss $\mathcal{L}_{\text{diff}}$ (Sec. 3.1) with three regularization terms as:

$$\mathcal{L}_{\text{CARE}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{load}}\,\mathcal{L}_{\text{load}} + \lambda_{\text{mask}}\,\mathcal{L}_{\text{mask}} + \lambda_{\text{mix}}\,\mathcal{L}_{\text{mix}} \quad (22)$$

where $\mathcal{L}_{\text{load}}$ ensures balanced expert utilization (Sec. 3.2), $\mathcal{L}_{\text{mask}}$ enforces the boundary consistency (Sec. 3.3), and $\mathcal{L}_{\text{mix}}$ encourages spatial smoothness in the mixture map (Sec. 3.4). To prioritize the reconstruction term while maintaining regularization, we empirically set small weights to regularizers:

$$(\lambda_{\text{load}},\, \lambda_{\text{mask}},\, \lambda_{\text{mix}}) = (0.01,\, 0.1,\, 0.05). \quad (23)$$

where the hyperparameters were identified according to prior works [41] and fixed for all experiments w/o extra tuning.

Role:
You are an expert scene designer for image-to-image generation.
Goal:
Given a brief subject prompt [asset], generate 5 different Scene Descriptions that place the same subject in diverse environments.
Rules:
1. Each Scene Description must reuse the identical subject [asset] but change the background, environment, camera view, lighting, or atmosphere.
2. Focus on describing ONLY the background and global conditions; do NOT alter the core identity of the subject.
3. Each Scene Description must be one line and should not exceed ~65 tokens.
4. Be highly creative and cover a wide range of locations, times, and moods.
Format:
[asset]: Koala
*[SceneDescription1]: [A koala] **sitting** on a wooden railing of a city rooftop garden at sunset, with skyscrapers in the distance.*
*[SceneDescription2]: [A koala] **sitting** on a picnic blanket in a sunny park, surrounded by scattered fruits and leaves.*
*[SceneDescription3]: [A koala] **sitting** on a tree trunk above a small creek in a lush tropical valley after rain.*
*[SceneDescription4]: [A koala] **clinging** to a tall eucalyptus tree in a misty forest at dawn.*
*[SceneDescription5]: [A koala] **resting** on a branch inside a glass-walled rainforest dome in a modern zoo.*

Figure 9. **GPT Prompting for Scene Diversity.** We explicitly instruct the LLM to generate varied scene descriptions (*e.g.*, "sunny park") while keeping the subject (*e.g.*, "Koala") constant. As such, by synthesizing multiple environments for the identical subject entity, we create training data that compels the image editing model to learn robust subject grounding independent of the background correlations.

Evaluation. We evaluate CARE-Edit on four representative image editing tasks across three diverse benchmarks that probe different aspects of contextual image editing:

- **EMU-Edit** [39]: this benchmark tests both the object-level and attribute-level modifications on real-world photos using fine-grained text prompt descriptions.
- **MagicBrush** [55]: this dataset involves complex, region-based editing tasks guided by free-form natural language instructions and the user-provided masks.
- **DreamBench++** [31]: this benchmark evaluates personalized subject-driven image editing and composition, covering single-object and complex multiple-object scenarios.

We follow the official data splits and evaluation subsets for each benchmark whenever available. Following the data processing pipelines in OmniControl [43] and UNO [48], we resize images to 512×512 while preserving aspect ratio.

B. Extended Qualitative Comparisons

Appendix B reports additional experimental results that could not be included in the main paper due to space limitations. In particular, we organize these results by task to demonstrate the model's robustness in handling diverse semantic demands, from subtle attribute changes to complex scene re-contextualization. For each task, we include representative per-category and per-edit-type samples and briefly discuss the behavior of CARE-Edit across different settings.

B.1. Instruction-based Image Editing

In this subsection, we introduce and discuss qualitative results on instruction-based image editing on EMU-Edit [39] and MagicBrush [55]. Figure 15 shows a large-scale visual comparison between our method and several strong baselines, including OmniGen2 [50], ACE++ [22], and the vanilla FLUX.1-Dev [18] backbone. The examples cover typical instruction-based edits such as style changes, attribute modifications, and cases involving visible text (*e.g.*, a toy holding a "CARE" label).

Qualitatively, CARE-Edit (especially the masked variant) follows the textual instructions while better preserving unedited content and fine-grained structures. Compared to SOTA methods, our method produces fewer spurious background changes and sharper, more localized boundaries at the edited regions. These trends are consistent with the quantitative gains reported in the main paper.

B.2. Subject-driven Contextual Image Editing

In this subsection, we provide more results on subject-driven contextual editing, primarily on DreamBench++ [31]. The goal is not only to preserve subject identity (*e.g.*, a particular person, pet, or product) but also to compose the subject into new contexts with complex surroundings and interactions.

A key motivation behind our design is that, for this class of tasks, it is often difficult to resolve the *relative size and placement* of the reference objects in the base image using

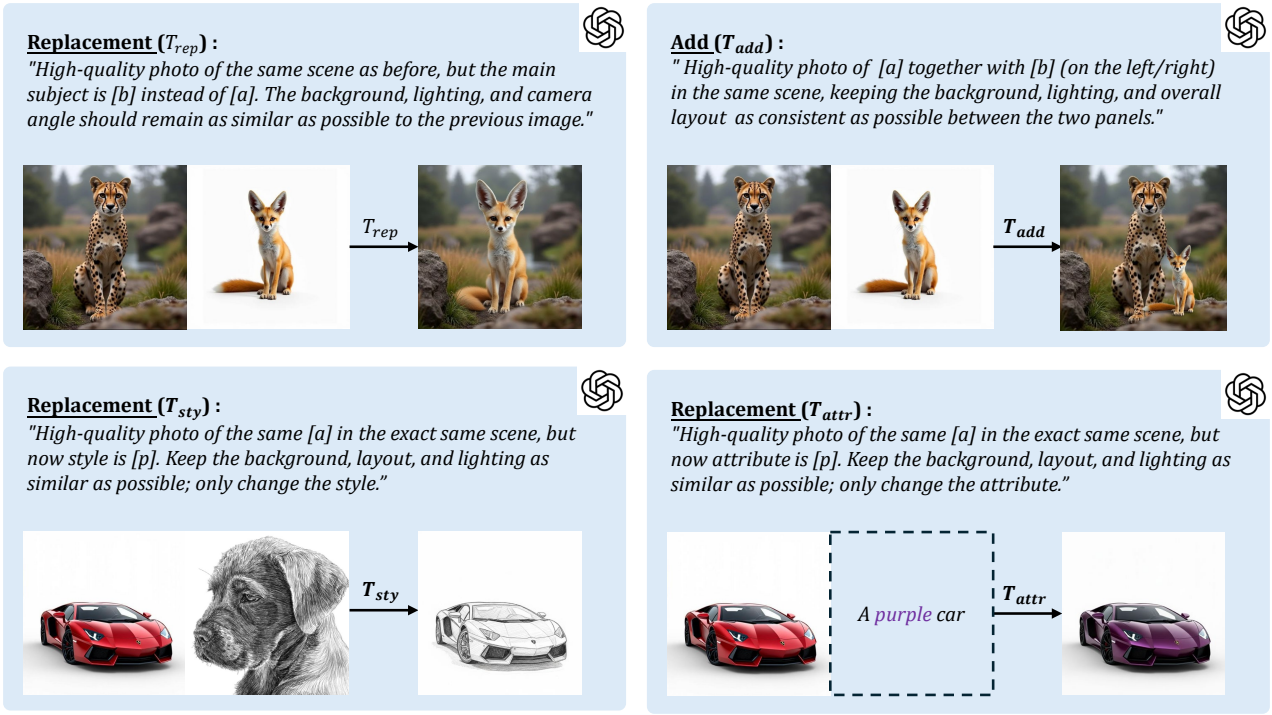

Figure 10. **Task-Specific GPT Prompt Templates.** We visualize the templates used to construct training triplets for: (i) *Replacement* (Top-Left); (ii) *Addition* (Top-Right); and (iii) *Style/Attribute Change* (Bottom). By constraining the LLM to modify specific slots (*e.g.*, subject identity) while holding scene descriptions constant, it ensures the resulting image pairs possess consistent backgrounds.

text prompts and the backbone model alone. Instructions such as "The man is holding a camera." or "Add a Rubik's Cube next to the sneaker." do not uniquely determine how large the inserted object should be or where it should appear. To address this, CARE-Edit incorporates a user-provided mask as an additional control signal. Even when the mask is coarse, it specifies the intended location and approximate size of the edited content, disambiguating the spatial relationship between the reference objects and the base subject.

Figure 16 illustrates this design with representative cases such as "The man is holding a cup.", "Add a watch next to the drink.", and "Add a toy bear next to the cat.". In all these examples, CARE-Edit produces edits where the inserted objects have plausible geometry and scale, while the main subject's identity, pose, and global lighting are preserved. This mask-guided formulation enables reliable subject-driven contextual editing in scenarios such as personalized product shots and multi-object layout design, where precise control over relative size and placement is crucial.

To further evaluate subject-driven contextual editing under more challenging compositions, we provide an extended teaser-style comparison in Figure 11. These examples require the model to jointly preserve the identity of the reference subject, fit the geometry and illumination of the base image, and follow the user's instructions the same time. Compared with strong open- and closed-source baselines, CARE-Edit produces more coherent compositions with better subject fidelity and fewer failure cases such as role reversing between the base and reference inputs, imbalanced object proportion, identity drift, and cut-and-paste artifacts.

B.3. More Results on Diverse Editing Tasks

In this concluding subsection, we present extensive visual evidence of CARE-Edit across different editing tasks and summarize how they map to practical usage scenarios. Figures 17–19, together with Figure 12, show extended qualitative results for object removal, object addition and replacement, and style transfer.

Object Removal. Given an input image, the model removes the selected object and synthesizes background content consistent with the surrounding regions. Figure 17 shows that our CARE-Edit can inpaint relatively large masked areas without obvious seams or blur, while leaving unedited regions nearly unchanged.

We additionally study robustness to imperfect user masks. Figure 12 shows that CARE-Edit remains effective even when the removal region is specified using coarse masks rather than a precise object mask. The model can still re-

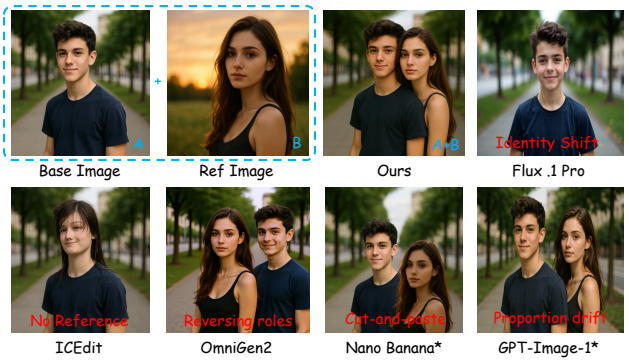

Figure 11. **Extended teaser comparison on challenging subject-driven contextual editing.** We compare CARE-Edit with more representative open- and closed-source (*) baselines on cases that require preserving reference-subject identity while composing it into a new scene specified by the base image and instruction. Competing methods may suffer from failures such as identity drift, role reversal between the base and reference inputs, imbalanced proportion, or visible copy-and-paste artifacts. In contrast, CARE-Edit better maintains subject fidelity and overall scene coherence.

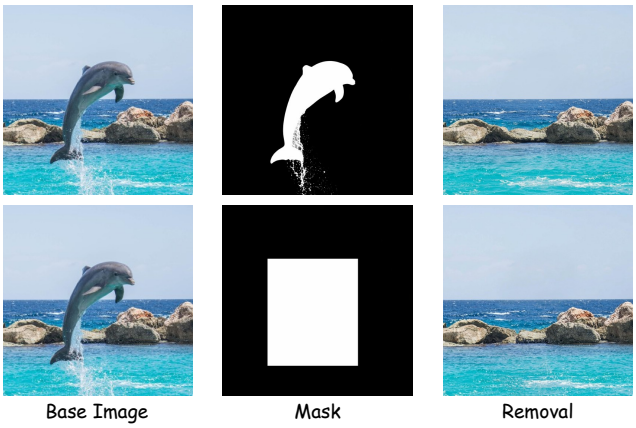

Figure 12. **Robust object removal under different mask qualities.** We compare CARE-Edit under coarse and more accurate user masks for object removal. Even with approximate masks, the model can successfully remove the target object and inpaint plausible background content with smooth local transitions. This highlights the robustness of CARE-Edit to coarse user annotations.

move the target object and impaint plausible background content with smooth transitions. A tighter mask further improves local texture recovery and boundary cleanliness. However, CARE-Edit can exploit more accurate spatial guidance without becoming fragile to approximate user input.

Addition and Replacement. The user provides a short text instruction (*e.g.*, "Add ...", "Replace ...") and a coarse mask indicating desired location and approximate size of the edited object. Figure 18 shows that CARE-Edit uses this mask to control scale and placement, filling the region with an object that matches the text and blends with the scene.

Style Transfer. The image content is largely preserved, while global appearance is modified according to a target style. Figure 19 shows that CARE-Edit maintains scene structure and object boundaries, avoiding severe detail loss.

These results show that CARE-Edit can handle removal, addition, replacement, and stylistic changes.

C. Analysis of Expert Lattent Attention Maps

To complement the qualitative results and provide a mechanistic understanding of CARE-Edit, we conduct a deep diagnostic empirical analysis of the model's internal expert learning behavior. A core hypothesis of this work is that different editing conditions (*e.g.*, text semantics vs. spatial masks vs. reference style) impose different learning dynamics on a shared backbone. CARE-Edit resolves this via an explicit routing of heterogenous, specialized experts.

The main paper only visualizes the *Base* Expert due to the space limitations. To validate that these experts indeed evolve distinct, complementary roles rather than collapsing into a uniform average, we visualize the attention maps of all three condition-aware experts, (i) *Base*, (ii) *Mask*, and (iii) *Reference*, throughout the training process. Figure 13 illustrates the evolution of these attention map distributions at different training iterations ($T = 0$, $T = 30$K, $T = 70$K, and $T = 100$K). This visualization effectively opens the black box of these experts, revealing how the model learns to disentangle complex editing objectives over time.

Base Expert: The Global Anchor. As shown in Figure 13, the *Base* Expert (Top Row) maintains a robust, spatially widespread activation pattern across the entire training trajectory. Even at late training stages ($T = 100$K), its attention map covers the majority of the image canvas with high intensity. This confirms its role as the task-agnostic anchor, which is responsible for preserving the intrinsic structure, lighting, and layout of the original image, while incorporating conditional information. By handling the global coherence, the *Base* Expert frees the other experts to focus purely on differential changes, ensuring that the unedited regions remain perceptually and semantically consistent with base images.

Mask Expert: Spatial and Geometric Specialization. The *Mask* Expert (mid-row in Figure 13) displays the most dramatic evolution, demonstrating the emergence of spatial intelligence. At early training stages ($T = 0$), the attention is diffuse, noisy, and object-unaware. However, as training progresses through the mid-phase ($T = 30$K to 70K), the entire attention becomes aggressively focused, concentrating strictly within and immediately around the user-provided input editing regions. By $T = 100$K, the *Mask* Expert ex-

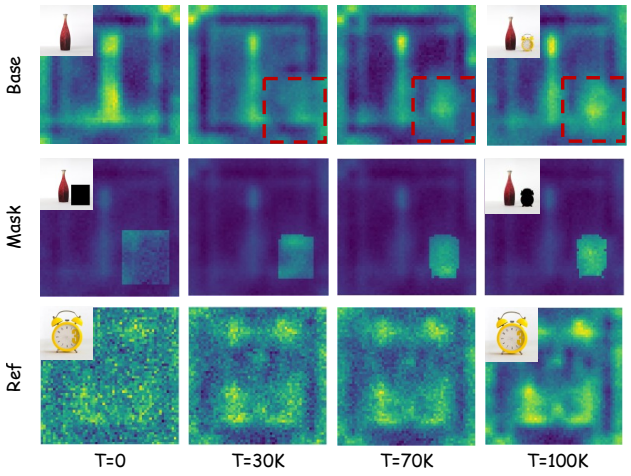

Figure 13. **Expert Specialization During Training.** We visualize the latent attention maps for the *Base*, *Mask*, and *Reference* experts at increasing training iterations (T). **(Top) Base Expert:** Maintains consistent, global activation throughout training ($T = 0 \rightarrow 100\text{K}$), acting as a foundation to preserve image structure. **(Middle) Mask Expert:** Learns to progressively suppress background signals, evolving from a noisy initialization to a highly localized, structure-aware attention map that precisely targets the edit region. **(Bottom) Reference Expert:** Gradually increases engagement in regions requiring semantic or stylistic modification. This distinct separation of concerns confirms that CARE-Edit effectively disentangles conflicting editing signals into specialized processing pathways.

hibits fine-grained, binary-like activation boundaries that align perfectly with the intended edit objects. This trajectory indicates that the *Mask* Expert successfully learns to exploit the *Mask Repaint* module's signals, delegating geometric restructuring (*e.g.*, object removal, shape modification) exclusively to this expert while suppressing its influence on the original background to prevent potential leakage or artifacts.

Reference Expert: Semantic and Stylistic Injection. The *Reference* Expert (Bottom Row in Figure 13) exhibits a distinct pattern of "semantic sparsity." Unlike the Mask expert, which aligns with geometry, the Reference expert aligns with content relevant to the style or identity transfer. Initially inactive, its attention grows as the model learns to map features from the reference image encoder (Z_r) to the generated latent space. At convergence stage ($T = 100\text{K}$), we observe heightened activation in regions that require texture synthesis or photometric adjustment (*e.g.*, the surface of an object changing material, or the entire scene during style transfer). More importantly, its activation map is orthogonal to the Base expert. It injects fine-grained appearance cues (*e.g.*, color, texture) without overwriting the original structural geometry maintained by the Base and Mask experts.

Overall, these distinct activation signatures observed in

Figure 14. **Layer-wise activation analysis.** We visualize the average activation of the routed experts across DiT blocks. Earlier blocks focus on the *Mask* Expert, reflecting the importance of spatial grounding and boundary localization in the early stage of denoising. As depth increases, the model relies relatively more on the *Reference* Expert for appearance refinement, while the *Base* Expert remains consistently active as a global anchor for layout preservation and scene coherence. The layer-wise activation shrifts further prove the condition-aware specialization of CARE-Edit.

Figure 13 validate the efficacy of our *Condition-Aware Routing* design. Instead of forcing a single set of weights to compromise between preserving identity and changing style, CARE-Edit dynamically distributes the workload: the *Mask* expert handles the "where," the *Reference* expert handles the "what" (appearance), and the *Base* expert ensures global consistency of the image to be edited. This learned specialization serves as the key factor enabling our CARE-Edit to minimize task interference from mulitple inputs and thus achieve high-fidelity editing in the challening multi-condition scenarios.

D. Analysis of Task-Expert Activation

Beyond the iteration-evolving attention maps, we also analyze how expert activation changes across routed DiT blocks. Figure 14 provides a layer-wise view of how CARE-Edit activates among specialized experts. Earlier routed blocks place stronger emphasis on the *Mask* Expert, which is consistent with the need to establish edit location and boundary geometry at the beginning of denoising. As depth increases, the model relies relatively more on the *Reference* Expert to refine appearance related details, while the *Base* Expert remains steadily active across depth to preserve global structure. This layer-dependent activation analysis provides additional evidence that CARE-Edit performs dynamic condition-aware routing rather than static multi-branch fusion during training.

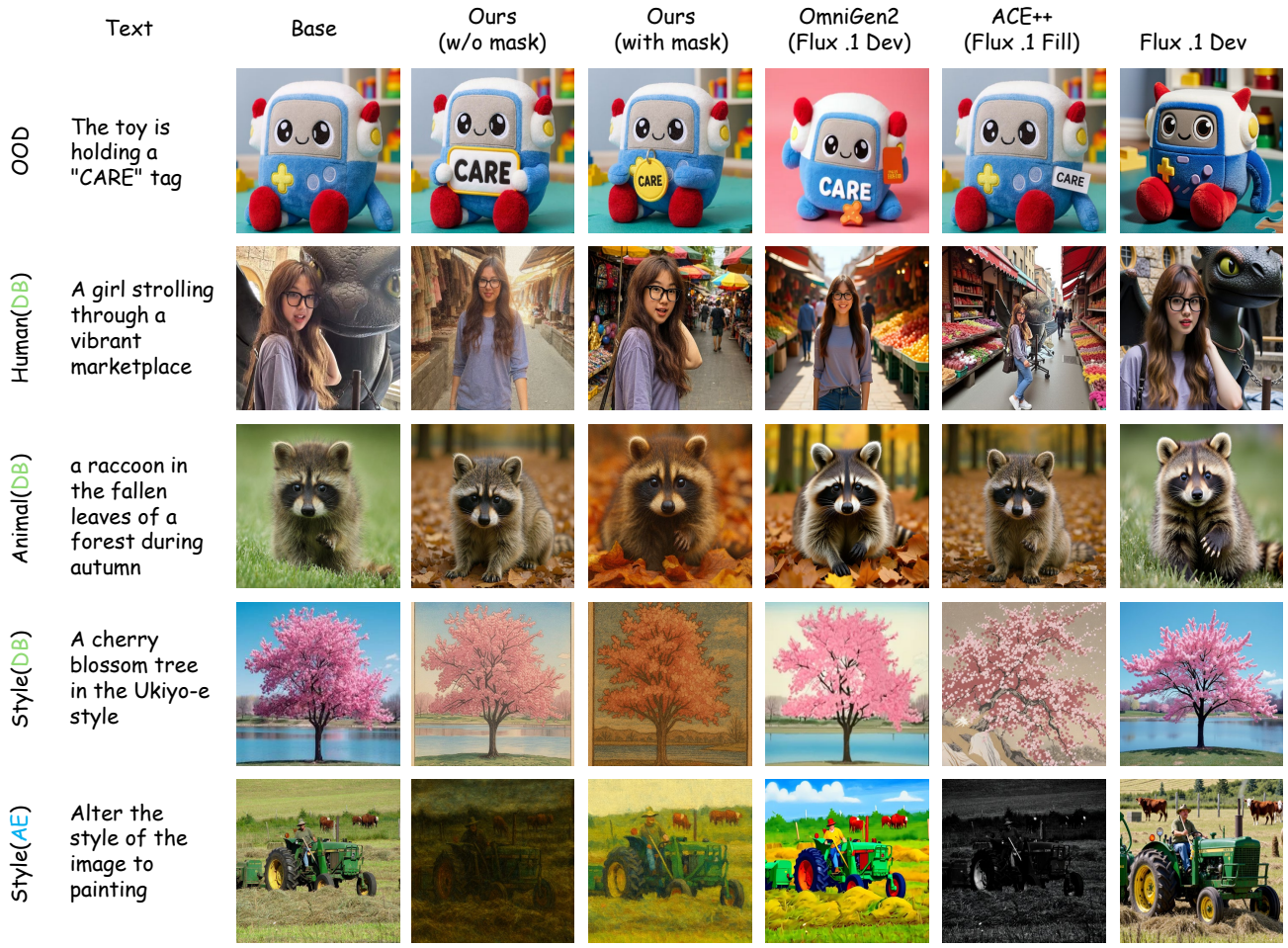

Figure 15. **Instruction-based Editing on EMU-Edit and MagicBrush.** We compare CARE-Edit against existing unified editors [47, 50]. Rows exhibit challenging scenarios such as text rendering ("CARE tag"), global style transfer ("Ukiyo-e"), and complex object insertions. **(1) Text and Geometric Fidelity (Row 1):** The geometric rigidity of the tag and the legibility of the text are paramount. While ACE++ and OmniGen2 correctly interpret the semantic intent, they suffer from *structural drift*, resulting in warped tag boundaries and deformed glyphs. CARE-Edit, particularly the masked variant, utilizes the Mask Expert to enforce spatial constraints, producing orthogonal tag geometry and crisp, readable text. **(2) Identity Preservation (Row 2):** CARE-Edit preserves the subject's facial identity and hair texture significantly better than baselines, which tend to over-blend the subject into the crowd (identity dilution) or generate a generic face. **(3) Style Disentanglement (Row 4 & 5):** In the *"Ukiyo-e"* and *"Painting"* style transfer tasks, baselines often hallucinate new objects or flatten the entire scene into a texture map. CARE-Edit disentangles style from structure. The *Base Expert* maintains the complex branching of the cherry blossom tree and the mechanical details of the tractor, while *Reference Expert* strictly applies the artistic texture to the environments.

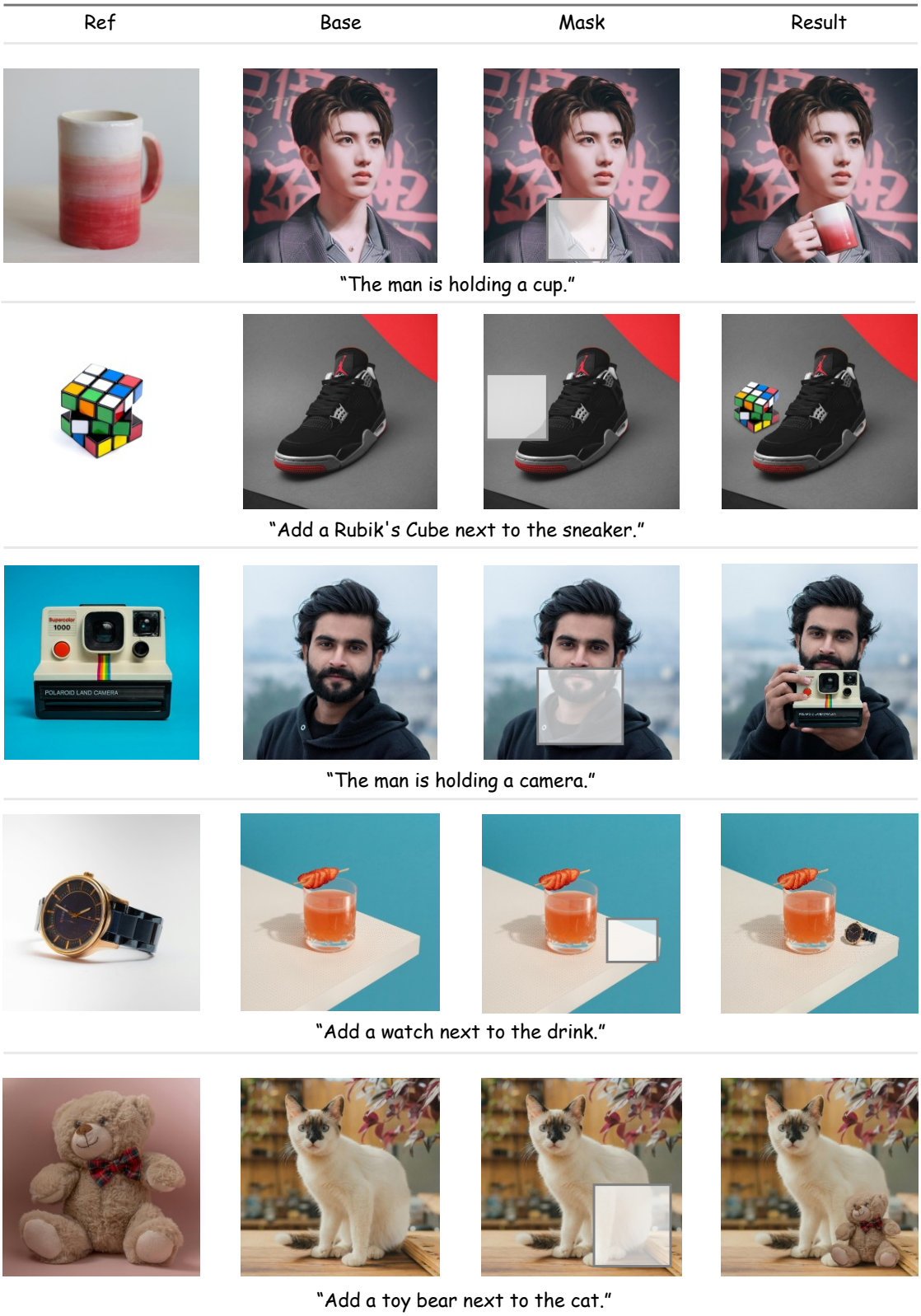

Figure 16. **Complex Contextual Editing Results.** Multi-condition examples requiring harmonization of subject identity, mask constraints, and text prompts. CARE-Edit successfully composes subjects into disparate environments while respecting the user-provided spatial layout.

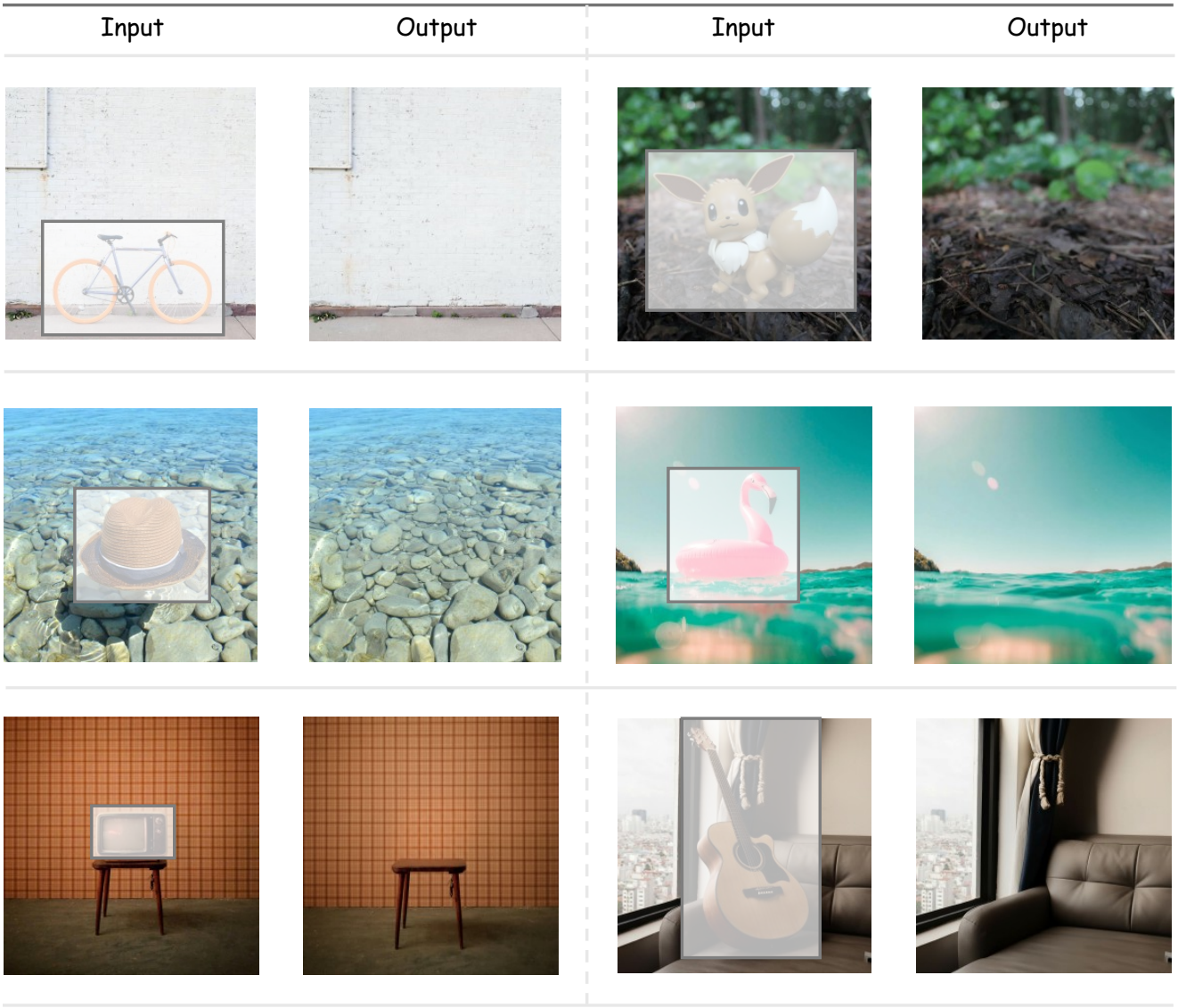

Figure 17. **Object Removal Results.** This task requires the model to remove the foreground object and hallucinate a plausible background (*e.g.*, sofa fabric, wall patterns) that is consistent with surrounding scene context. **(i) Top and Middle:** CARE-Edit successfully handles stochastic textures, such as natural water ripples and uneven stone surfaces, filling the void with spatially coherent content. **(ii) Bottom-Left:** A stress test for structural consistency. Removing the television requires reconstructing the rigid grid pattern of the wallpaper. CARE-Edit accurately hallucinates the missing tiles, maintaining the correct perspective and alignment without the blurring or geometric warping.

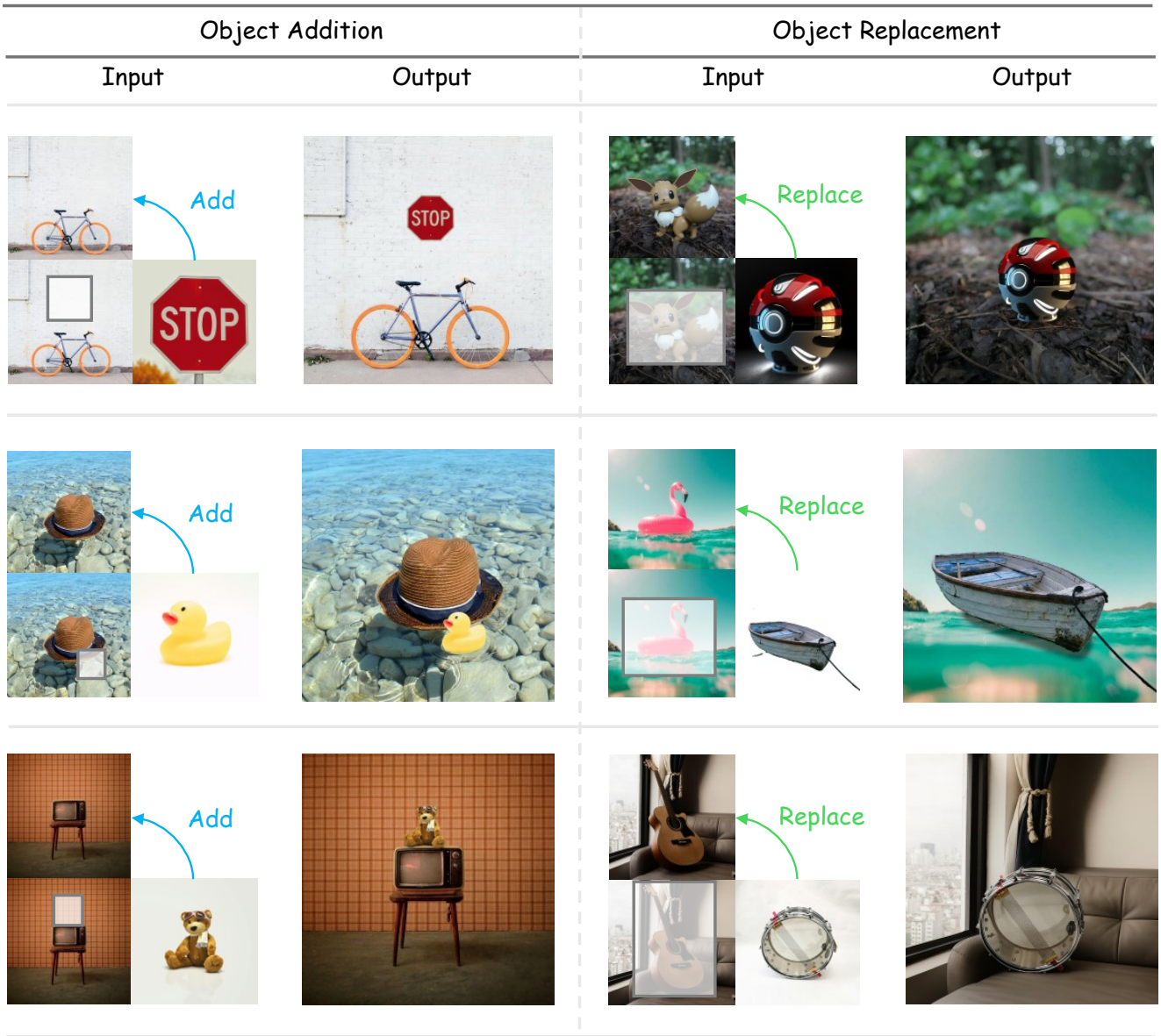

Figure 18. **Object Addition and Replacement Results.** CARE-Edit demonstrates precise control over scene composition, inserting or swapping objects while rigorously adhering to environmental constraints. **(i) Object Addition (Left):** CARE-Edit introduces new elements that respect physical laws. Note that how the added rubber duck (middle) is generated with accurate water reflections and surface interaction. **(ii) Object Replacement (Right):** CARE-Edit handles drastic changes in structure and material while maintaining lighting consistency. In the top-right example, replacing a furry pokemon creature with a pokemon ball (metallic sphere), CARE-Edit correctly renders specular highlights and casts realistic shadows onto the complex dirt terrain, ensuring the newly added object sits naturally within the depth of field.

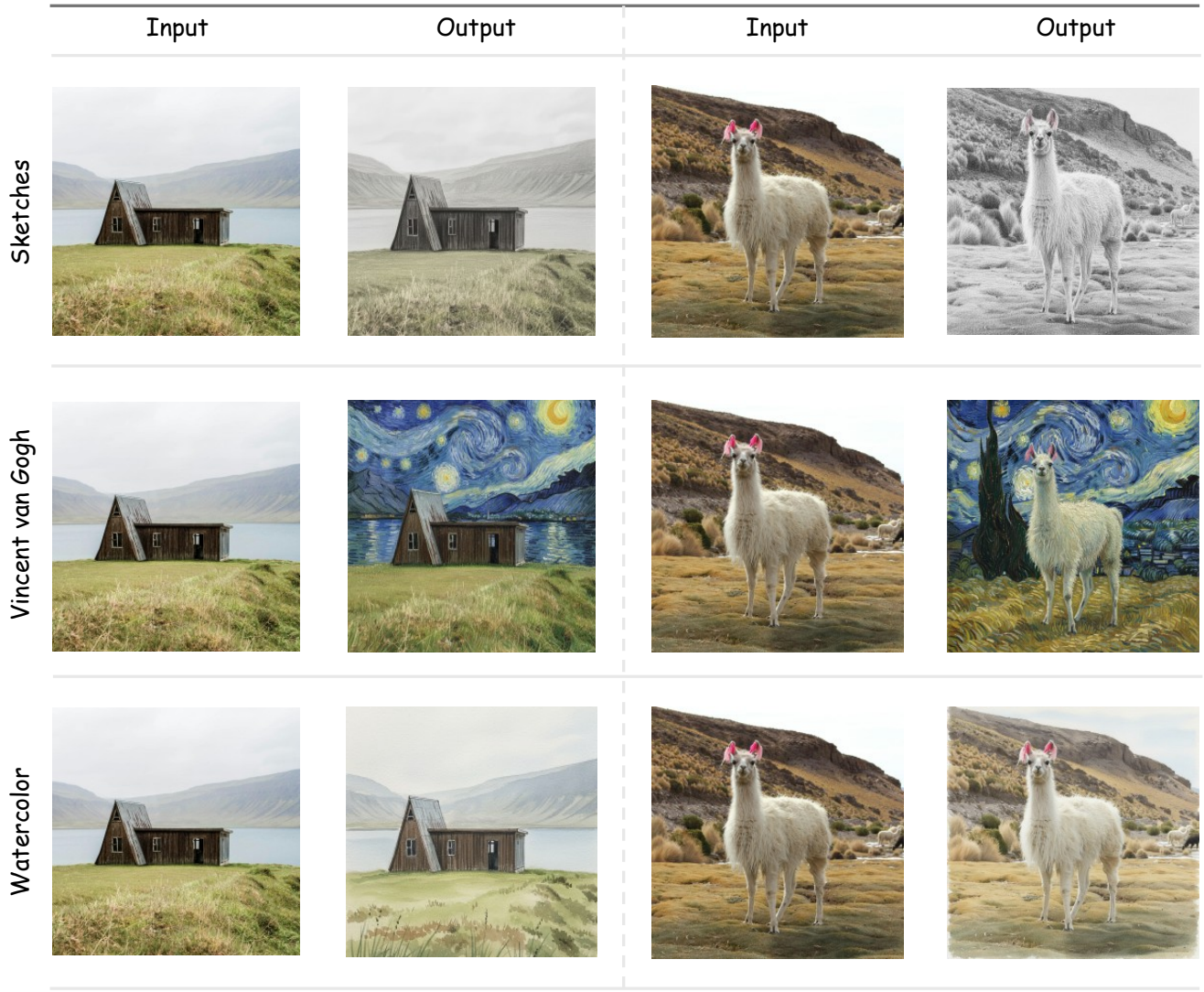

Figure 19. **Reference-guided Style Transfer with Structure Preservation.** A key advantage of CARE-Edit is the ability to decouple style from structure via expert routing. Unlike holistic transfer methods that often deform the underlying geometry, our approach injects the target aesthetic while strictly anchoring the scene layout. **(i) Top:** The results show that CARE-Edit's *Reference Expert* successfully translates the scene into the swirling impasto of Van Gogh or a watercolor wash, yet the *Base Expert* ensures the architectural rigidity of the cabin, preserving the straight lines of the roof and window frames. **(ii) Bottom:** The subject's fur texture is re-rendered to match the artistic medium, demonstrating CARE-Edit's successful fine-grained textural adaptation without distorting the animal's original silhouette or pose.