BRIDGE: BACKGROUND ROUTING AND ISOLATED DISCRETE GATING FOR COARSE-MASK LOCAL EDITING

Anonymous authors
Paper under double-blind review

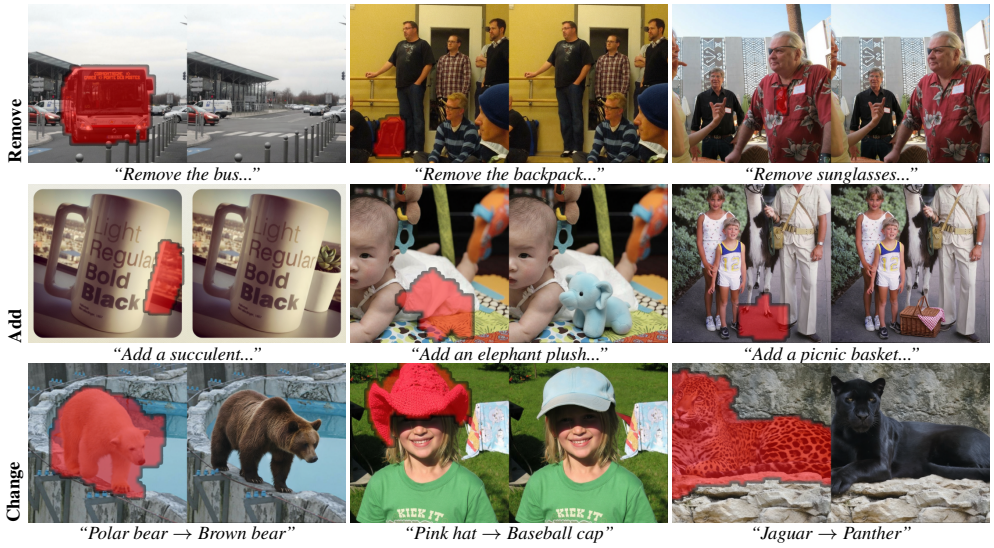


Figure 1: **BRIDGE produces high-quality coarse-mask local edits while preserving seamless background fusion.** All examples in this figure are generated by our final BRIDGE model. Across remove, add, and change edits, the generated results remain well integrated with the surrounding scene while allowing the edited subject to depart from the rough mask shape when needed. Together, these examples illustrate the central goal of coarse-mask local editing: maintain background consistency without inheriting mask-shape bias.

ABSTRACT

Coarse-mask local editing must preserve the background while allowing new content to depart from the mask contour. Existing mask-conditioned editors often entangle these goals and treat the coarse mask as a shape prior. We introduce **BRIDGE**, a dual-path architecture with a full-canvas Main Path, a local Subject Path, and a learned binary gate that routes subject-token positional embeddings between subject-centric and background-anchored coordinates. On Qwen-Image-Edit-2511, BRIDGE keeps the mask outside the DiT backbone and trains both targets with independent noise. On BRIDGE-Bench, it achieves 0.503 local SigLIP2-T, 0.902 DINO, and 0.175 DreamSim, outperforming ACE++ and FLUX.1-Fill in local precision. We then instantiate BRIDGE on FLUX.2 [klein] 9B Base with full-transformer training and subject-reference conditioning. Dense BBox supervision preserves Sub-image quality; optional mask-cropped inference removes out-of-mask Sub tokens, tightening user support without training a sparse model that overfits the mask silhouette. On supported ICE-Bench tasks, FLUX.2 BRIDGE averages 0.5941 over four local edits and 0.6254 over subject-guided edits, leading the compared open-source models on the latter. Across Qwen and FLUX.2, BRIDGE separates support localization from subject geometry and improves local generation while preserving the surrounding scene.

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

In real local image editing, users rarely provide object-accurate masks. They draw rough scribbles or boxes to indicate where an edit should occur. The mask must therefore localize the operation without prescribing the geometry of the result, while the editor must preserve the surrounding scene. We call the failure to satisfy this separation *mask-shape bias*: a model treats a localization hint as the target object contour. This produces a Two-Zone Constraint: the background should remain stable, whereas the editable region should follow the instruction and may depart from the mask boundary. Figs. 1 and 2 show the resulting tension in remove, add, and change edits.

Existing methods obtain locality mainly through explicit mask conditioning or inference-time attention control. Inpainting and control-branch editors such as FLUX (Black Forest Labs, 2024), ACE++ (Mao et al., 2025), BrushNet (Ju et al., 2024), and PowerPaint-style task prompting (Zhuang et al., 2024) encode the mask as latent tokens, auxiliary features, or visual context. This provides strong spatial support, but also allows an imprecise input contour to influence object shape. Training-free methods such as Prompt-to-Prompt (Hertz et al., 2023), Null-text Inversion (Mokady et al., 2023), MasaCtrl (Cao et al., 2023), and Plug-and-Play diffusion features (Tumanyan et al., 2023) manipulate attention without retraining, but large geometric changes depend on schedules and attention heuristics. Neither route explicitly separates the support used for localization from the coordinate system used to generate a new subject.

A diagnostic Qwen-Image experiment (Wu et al., 2025) provides the key design clue. Assigning the same positional embedding (PE) to different visual tokens causes them to reuse similar image context, while a cross-region attention mask suppresses this coupling. PE assignment and attention connectivity therefore jointly control where a token borrows context and how strongly that context is used. Based on this observation, we propose **BRIDGE** (Background Routing and Isolated Discrete Gating for Coarse-Mask Local Editing). BRIDGE processes a full-canvas Main Path and a local Subject Path in one DiT. The mask defines the Subject Path support and external latent-blending region; a learned binary gate routes each subject token between subject-centric and background-anchored PEs. Boundary tokens can borrow background coordinates for fusion, while interior tokens retain the geometric freedom needed to create new structure.

Our first instantiation adapts Qwen-Image-Edit-2511 (Wu et al., 2025) with rank-512 LoRA (Hu et al., 2022) and 13.31M GateBlock parameters across 60 layers. The Main and Subject targets receive independently sampled noise, and the final edited image is decoded from the Main Path. On BRIDGE-Bench, this model reaches 0.503 local SigLIP2-T, 0.902 local DINO, and 0.175 local DreamSim. It also transfers zero-shot to MagicBrush and ICE-Bench, providing evidence that the learned routing is useful beyond the constructed benchmark.

We then ask whether the same separation transfers to a different backbone and to subject-reference editing. The FLUX.2 [klein] 9B Base (Black Forest Labs, 2026) instantiation trains the full transformer with layer-wise gates and conditions on a generated subject reference. Its Sub branch uses an aligned crop. Training only mask-selected Sub tokens produced less coherent Sub images, color drift for small supports, and strong mask-shape overfitting. We therefore retain dense BBox supervision for Sub-image quality and optionally crop Sub tokens by the user mask only at inference. Tokens outside the mask are removed together with their PE candidates, tightening spatial support without teaching a sparse training model to fill the mask silhouette. The Main image is decoded independently, so the cropped Sub preview is a diagnostic rather than the final output.

The two backbones are evaluated under separate protocols. The Qwen model is compared on BRIDGE-Bench, MagicBrush (Zhang et al., 2023), and ICE-Bench (Pan et al., 2025). The FLUX.2 model is evaluated on supported ICE-Bench mask tasks using rewritten self-contained prompts and subject references when provided. It averages 0.5941 over four supported local-edit tasks and 0.6254 over subject-guided editing, leading the compared open-source models on the latter group. These results connect the original Qwen routing mechanism with the later FLUX.2 extension while keeping their training and evaluation settings distinct.

Our contributions are threefold:

- We formulate mask-shape bias as a failure to separate spatial support from output geometry and introduce BRIDGE, which combines Main/Subject paths with learned token-level PE routing.

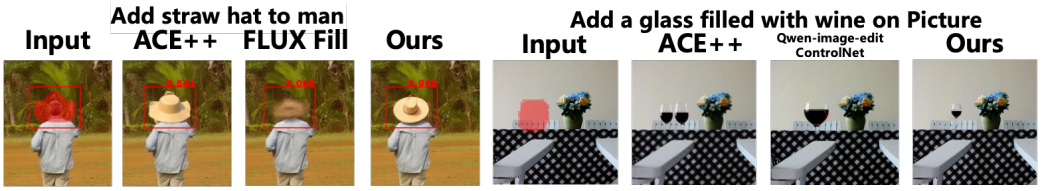


Figure 2: **Mask-shape bias in coarse-mask local editing.** The left four examples are from the test split of our constructed dataset, and the right four examples are from the public ICE-Bench benchmark. On the left, ACE++ produces a jagged hat boundary that follows the rough mask, FLUX.1-Fill fails to complete the edit, and our method **BRIDGE** generates a natural round hat. On the right, **BRIDGE** generates a single cup as requested, whereas ACE++ produces two cups and Q-Control (DiffSynth-Studio, 2025) enlarges the cup unrealistically, showing that mask-conditioned baselines often overfit the wide mask instead of following the instruction.

- We instantiate BRIDGE on both Qwen-Image-Edit-2511 and FLUX.2, and introduce dense BBox training with optional mask-cropped Sub inference for the FLUX.2 subject-reference setting.
- We evaluate the Qwen and FLUX.2 models under separate benchmark protocols, showing improved local precision, cross-benchmark transfer, and competitive open-source ICE-Bench results.

2 Related Work

2.1 Instruction-Guided Image Editing

Generative image editing has advanced rapidly with diffusion models (Ho et al., 2020; Rombach et al., 2022; Peebles & Xie, 2023). Early text-guided methods such as InstructPix2Pix (Brooks et al., 2023), DiffEdit (Couairon et al., 2023), and Prompt-to-Prompt (Hertz et al., 2023) enabled global style changes and local attribute edits from text descriptions. More recent models such as ACE (Han et al., 2025) and ACE++ (Mao et al., 2025) introduce multimodal conditioning for more context-aware editing. These models improve instruction following, but they do not directly address the coarse-mask regime where the mask is only a spatial hint and the target object geometry must deviate from the marked contour.

2.2 Attention and Layout Control in Diffusion

Training-based layout control learns spatial conditioning from annotated layouts. LayoutDiffusion (Zheng et al., 2023) trains a layout-to-image diffusion model with object-aware cross-attention that aligns bounding boxes and generated objects. BRIDGE instead targets coarse-mask local editing: its full-canvas and bbox-local paths are jointly trained, and a per-layer gate learns how subject-token coordinates interact with source-image context.

Training-free methods manipulate attention maps or positional embeddings in pretrained diffusion models. Prompt-to-Prompt (Hertz et al., 2023) and Plug-and-Play (Tumanyan et al., 2023) use cross-attention or self-attention features for spatial and semantic control. BoxDiff (Xie et al., 2023) and LoCo (Zhao et al., 2025) impose localized constraints for layout adherence, while MasaCtrl (Cao et al., 2023), PosBridge (Xiong et al., 2025), and RoPECraft (Gokmen et al., 2025) manipulate attention keys or positional embeddings to transfer structure, identity, or motion. BRIDGE learns a token-level PE routing policy from the current visual features instead of applying a fixed routing rule.

2.3 Inpainting and Generative Filling

Inpainting models are a strong baseline for local editing. LaMa (Suvorov et al., 2022) established high-resolution completion, while diffusion-based inpainters such as FLUX.1-Fill (Black Forest Labs, 2024) and BrushNet (Ju et al., 2024) improve photorealistic filling under mask constraints. PowerPaint-style task prompting (Zhuang et al., 2024) further improves instruction alignment by

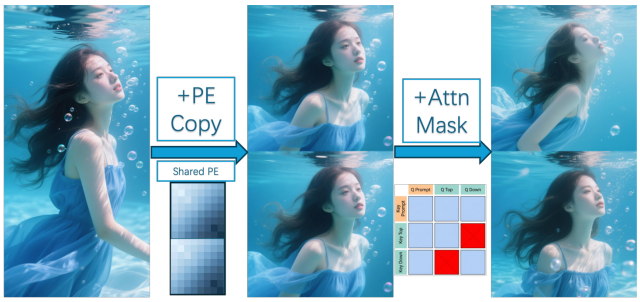


Figure 3: **Diagnostic Qwen-Image (Wu et al., 2025) experiment.** Positional embeddings determine which image context a token reuses. Sharing a PE across different visual tokens yields duplicated or highly similar content, while a cross-region attention mask suppresses this aliasing. This motivates BRIDGE's token-level PE switch for controlling where subject tokens borrow context from.

conditioning on textual edits. Recent works like OminiControl (Tan et al., 2025) introduce universal spatial control via in-context learning by concatenating image and mask inputs. In these families, the mask is typically compressed together with image content through a VAE or related encoder and then provided to the DiT as latent tokens, feature maps, or in-context visual conditions. Treating a coarse user mask as a geometric boundary can couple the final object shape to an irregular support region. BRIDGE assigns the mask to external support construction and blending, while BridgePath layout and PE routing govern geometry formation inside the DiT.

3 METHOD

BRIDGE addresses coarse-mask local editing by separating two decisions that inpainting models often entangle: where the edit is allowed to occur, and how the new content should use the surrounding context. The mask defines the bbox-aligned Subject Path support and the external latent-blending region. BridgePath layout and learned positional-embedding (PE) routing control generation inside the DiT. We first formalize this separation through the Two-Zone Constraint, then introduce BridgePath generation, the Discrete Geometric Gate, the training objective, and the data-construction pipeline.

3.1 PROBLEM FORMULATION AND DIAGNOSTIC INSIGHT

Given a source image I, an instruction y, and a coarse user mask M, local editing can be described by two zones. The background zone $\Omega_b = 1 - M$ should preserve appearance, texture, and layout. The editing zone $\Omega_e = M$ should satisfy the instruction while allowing the generated object to take a natural shape that may differ from the mask contour. In implementation, the coarse mask is converted to an axis-aligned patch-grid bounding box for Subject Path support, and the original mask remains available for optional blending. The mask localizes the user's intent while allowing the generated geometry to differ from its contour.

We use a diagnostic experiment on Qwen-Image (Wu et al., 2025) to motivate PE routing. We assign the same PE to distinct visual tokens and observe that they tend to generate duplicated or highly similar content, as illustrated in Fig. 3. This indicates that PE can act as a context selector: changing a token's coordinates changes which visual context the pre-trained generator is inclined to reuse. At the same time, applying a cross-region attention mask suppresses this duplication, showing that PE assignment and attention connectivity jointly determine both the source and the strength of contextual reuse. BRIDGE builds on this observation by training LoRA adapters and a token-level PE switch, so Subject Path tokens can decide where to borrow background context from and how strongly to use that context during generation.

3.2 BRIDGEPATH GENERATION AND PE ROUTING

BRIDGE uses a BridgePath architecture processed by one multimodal DiT backbone (Esser et al., 2024). The full-canvas Main Path generates the final edited image, while the bbox-local Subject Path generates the editable content. At inference, the two paths start from independently sampled Gaussian

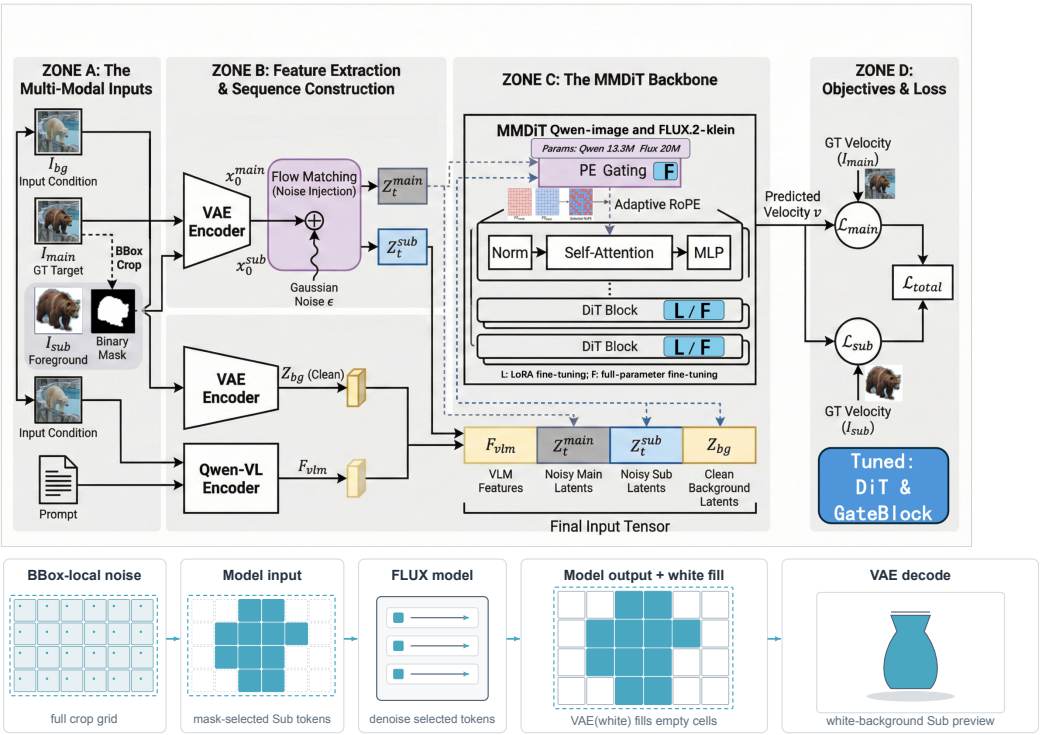


Figure 4: **BRIDGE overview and optional mask-cropped Sub inference. Top:** The Main Path preserves full-canvas context, while the Subject Path generates the editable region and the Discrete Geometric Gate routes subject-token positional embeddings. **Bottom:** During optional mask-cropped inference, Gaussian noise is initialized on the bbox-local Sub grid and only mask-selected tokens are retained and denoised. The sparse output is restored to its original crop-grid positions, and missing cells receive the VAE latents of an all-white crop before decoding. This operation belongs to the BRIDGE inference path and is independent of the specific backbone. The white-background Sub preview is diagnostic; the final edited image is decoded independently from the Main branch.

noise. Their tokens are concatenated with source-image condition tokens in the visual sequence and jointly processed with instruction tokens in each DiT block. Attention and PE routing determine how subject tokens use or separate from source-image context. Fig. 4 summarizes the architecture; Appendix A gives the training-field and supervision mapping.

The lower panel of Fig. 4 shows BRIDGE's optional mask-cropped Sub inference path. Appendix G.2 details the training observations and inference design evaluated in the cross-backbone extension.

The Subject Path uses background-anchored coordinates when scene alignment is useful and subject-centric coordinates when new structure is required. The learned gate makes this choice separately for each subject token and DiT layer.

3.3 Discrete Geometric Gating

Complete isolation can yield floating or poorly integrated objects, while always using background coordinates can reintroduce mask-shape bias. We therefore introduce a Discrete Geometric Gate that predicts a binary routing variable for subject tokens in each DiT block.

For layer l, let $H^l = [Z^l_{\text{main}}; Z^l_{\text{sub}}; Z^l_{\text{cond}}] \in \mathbb{R}^{N_l \times D}$ denote the visual sequence containing Main, Subject, and source-image condition tokens, with $D = 3072$. Each layer has an independent GateBlock h^l_ϕ. The GateBlock projects the visual sequence to a 64-dimensional hidden space, applies a single-layer Transformer encoder, and predicts a scalar logit for each subject-token index $i \in \mathcal{S}_l$,

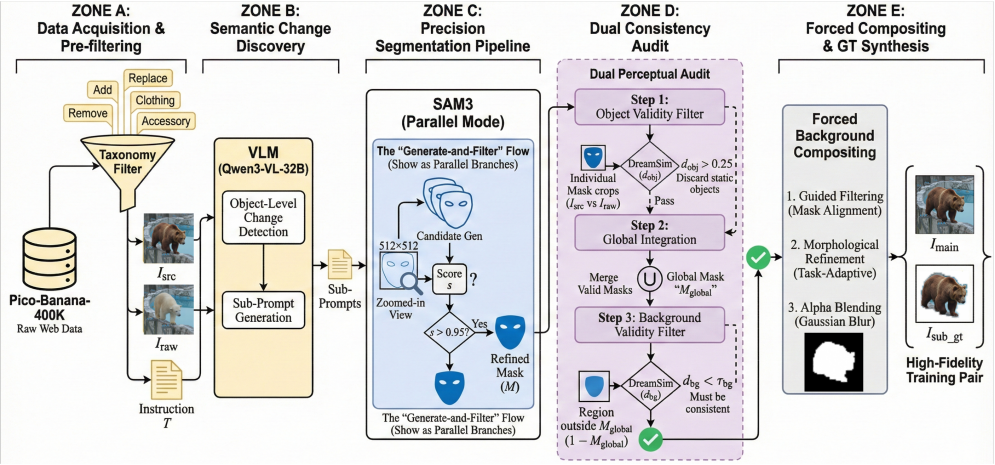


Figure 5: **Automated data pipeline for BRIDGE-Bench.** A VLM discovers local edit concepts, SAM proposes candidate masks, DreamSim filters for sufficient foreground change and limited background drift, and forced compositing restores the untouched scene. The resulting records provide the full edited target, source-image condition, subject target, and support mask used by the two-path training pipeline.

where $\mathcal{S}_l$ is the token range of the bbox-local Subject Path. The routing probabilities are

$$p_i^l = \sigma\big(h_\phi^l(H^l)_i\big), \quad i \in \mathcal{S}_l, \tag{1}$$

The threshold stays fixed at 0.5, with entropy and sparsity regularization weights set to zero. Its final linear head is zero-initialized. The hard forward decision and straight-through value are

$$G_{\text{hard},i}^l = \mathbf{1}\big[h_\phi^l(H^l)_i \geq 0\big] = \mathbf{1}\big[p_i^l \geq 0.5\big], \tag{2}$$

$$G_{\text{ST},i}^l = G_{\text{hard},i}^l + p_i^l - \text{stopgrad}(p_i^l). \tag{3}$$

The forward value remains binary, while the sigmoid probability supplies the surrogate gradient (Bengio et al., 2013). The effective PE for subject tokens is

$$PE_{\text{eff},i}^l = G_{\text{ST},i}^l\, PE_{\text{base},i} + (1 - G_{\text{ST},i}^l)\, PE_{\text{swap},i}, \quad i \in \mathcal{S}_l, \tag{4}$$

where PE_{base} denotes the original Subject Path coordinates and PE_{swap} denotes background-anchored coordinates copied from the corresponding Main Path support. Thus, a forward value of 1 favors structural freedom, while 0 favors scene fusion. Zero logits give $p = 0.5$, and the inclusive threshold initially selects PE_{base}. Across a 60-layer DiT, the GateBlocks add about 13.31M parameters, compared with roughly 1.13B parameters for a ControlNet-style block copy under the same base-model dimensions. Additional implementation details are provided in Appendix A.

3.4 Training Objective

BRIDGE is trained with the flow-matching objective used by modern DiT generators (Esser et al., 2024). Let $Z_0 = z_0^{\text{main}} \oplus z_0^{\text{sub}}$ be the concatenated clean targets: z_0^{main} encodes the full background-preserved edited image, and z_0^{sub} encodes the foreground-on-white target crop within the Subject Path support. We form $Z_1 = z_1^{\text{main}} \oplus z_1^{\text{sub}}$ by sampling z_1^{main} and z_1^{sub} independently from matching Gaussian distributions. For $t \sim \mathcal{U}(0, 1)$,

$$Z_t = tZ_1 + (1 - t)Z_0. \tag{5}$$

The network predicts the velocity field over the concatenated Main/Subject sequence, conditioned on the instruction embedding T, timestep t, and source-image condition z_{edit}:

$$\mathcal{L}_{fm} = \mathbb{E}\left[\|v_\theta(Z_t, T, t, z_{edit}) - (Z_1 - Z_0)\|_2^2\right]. \tag{6}$$

We optimize the LoRA adapters (Hu et al., 2022) and GateBlocks end-to-end as a joint generation problem over the Main and Subject paths. The user mask supplies bbox construction and inference-time blending support. Appendix A specifies the complete tensor and supervision flow.

Table 1: BRIDGE-Bench test-dataset results. Local metrics use the edit bounding box; lower DreamSim is better and all other metrics are higher-is-better.

Method	Local Precision				Global Quality (BRIDGE-Bench)			
	Sig-T↑	DINO↑	DSim↓	Sig-I↑	DINO↑	DSim↓	Sig-I↑	Sig-T↑
ACE++ (Mao et al., 2025)	0.390	0.852	0.303	0.856	0.982	0.081	0.954	0.705
FLUX.1-Fill (Black Forest Labs, 2024)	0.262	0.807	0.321	0.835	0.978	0.076	0.947	0.632
BRIDGE	**0.503**	**0.902**	**0.175**	**0.907**	**0.988**	**0.054**	**0.965**	**0.776**

Table 2: MagicBrush zero-shot results on full images using current-turn local descriptions (current protocol; all ↑). SigLIP2-T is the model-native sigmoid score defined in Eq. 8.

Method	Final Turn					All Turn				
	CLIP-I	CLIP-T Local	DINOv3	SigLIP2-I	SigLIP2-T Local	CLIP-I	CLIP-T Local	DINOv3	SigLIP2-I	SigLIP2-T Local
ACE++	0.8909	0.2819	0.9464	0.8036	0.3612	0.9362	0.2824	0.9706	0.8871	0.4177
FLUX.1-Fill	0.8979	0.2701	0.9503	0.8058	0.2691	**0.9466**	0.2738	0.9742	0.8969	0.3399
Q-Control	0.8984	0.2934	0.9565	**0.8446**	0.4797	0.9425	0.2908	0.9765	0.9067	0.5081
BRIDGE	**0.9070**	**0.2959**	**0.9587**	0.8438	**0.5584**	0.9463	**0.2925**	**0.9775**	**0.9070**	**0.5593**

3.5 Data construction pipeline

Training requires examples where the foreground changes while the background remains stable. We build an internally processed dataset from Pico-Banana-400K (Qian et al., 2026): a VLM discovers local edit concepts, SAM proposes candidate masks (Carion et al., 2026), DreamSim filters for sufficient object change and limited background drift (Fu et al., 2023), and forced compositing preserves the untouched scene. Fig. 5 summarizes this construction flow. The processed pool defines the training split and the BRIDGE-Bench test dataset; full thresholds, target construction, and evaluation preprocessing are provided in Appendix A.

4 Experiments

We evaluate local precision, global quality, transfer, ablations, and cost.

4.1 Qwen-BRIDGE Results

Setup. Our processed Pico-Banana-400K (Qian et al., 2026) pool contains 42,425 training pairs, and the BRIDGE-Bench test dataset contains 1,444 pairs. Both splits follow the same filtering and compositing pipeline. We also evaluate zero-shot transfer on MagicBrush (Zhang et al., 2023) and ICE-Bench (Pan et al., 2025). External baselines use released checkpoints and settings; internal ablations share the Qwen-Image-Edit-2511 (Wu et al., 2025) backbone, LoRA configuration, data, optimizer, hardware, and budget. We report FLUX.1-Fill (Black Forest Labs, 2024), ACE/ACE++ (Han et al., 2025; Mao et al., 2025), UltraEdit (Zhao et al., 2024), Q-Control (DiffSynth-Studio, 2025), DINOv3 (Siméoni et al., 2025), SigLIP2 (Tschannen et al., 2025), DreamSim (Fu et al., 2023), and CLIP (Radford et al., 2021). Local BRIDGE-Bench metrics use the edit bounding box; full protocols and training details are in the Appendix.

BRIDGE-Bench. On the BRIDGE-Bench test dataset, BRIDGE leads all reported local metrics in Table 1, including SigLIP2-T (0.503), DINO (0.902), and DreamSim (0.175).

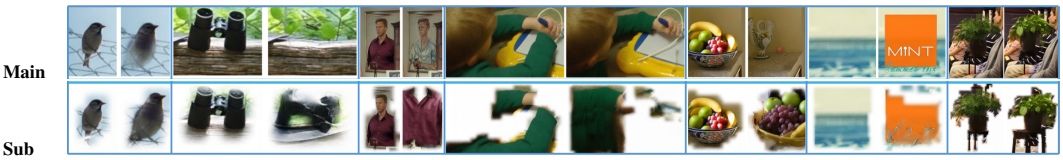


Figure 6: **Visual identity of the routing ablation.** The top and bottom rows show Main and Sub outputs, respectively. Within each example pair, Full BRIDGE is on the left and BridgePath with fixed PE is on the right. Adaptive routing yields sharper Sub content in these examples, consistent with Table 4.

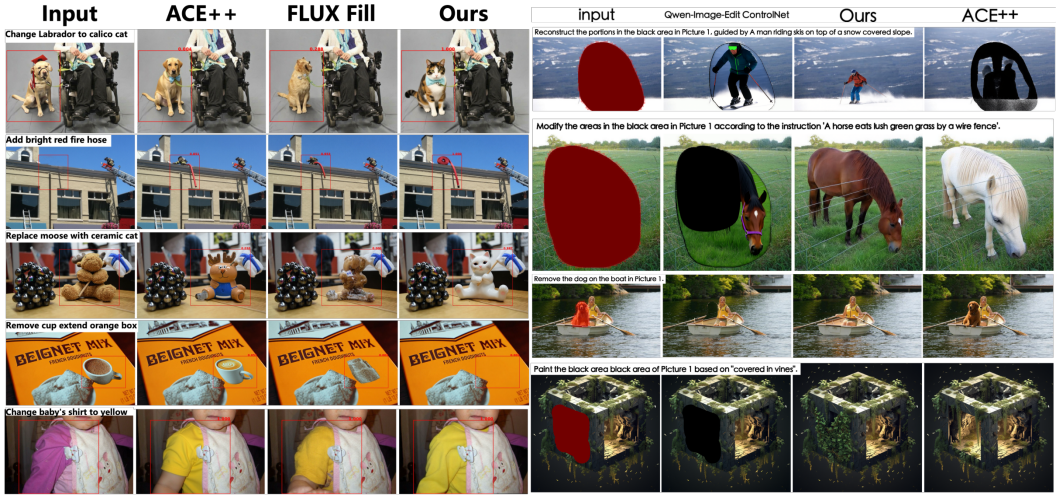


Figure 7: **BRIDGE improves prompt-consistent local generation under coarse masks.** Left: our constructed test split. Right: ICE-Bench examples. Compared with Q-Control and ACE++, BRIDGE more often follows prompts, generates plausible local geometry, and blends edits cleanly into the scene.

Table 3: ICE-Bench results grouped by input/prompt protocol; comparisons are made within each block. The left block uses the orange data set, where we replace `<SOURCE>` with "Picture 1" and describe the masked region as the black area. The right block uses the original ICE-Bench inputs and prompts, which may include `<Mask>` and `<SOURCE>`.

Orange data set (modified inputs/prompts)					**Original ICE-Bench inputs/prompts**				
Model	AES↑	IMG↑	PF↑	SRC↑	Model	AES↑	IMG↑	PF↑	SRC↑
ACE++	0.480	**0.575**	0.575	0.925	ACE	0.480	0.487	**0.643**	0.916
Q-Control	0.472	0.569	**0.588**	0.937	ACE++	**0.481**	**0.577**	0.616	0.930
BRIDGE	**0.494**	0.565	0.582	**0.942**	UltraEdit	0.455	0.458	0.442	**0.951**

Table 4: BRIDGE-Bench test-dataset ablation. The mask-free row disables external blending ($\alpha = 0$); trained variants share the backbone, LoRA, data, optimizer, and budget.

	Global Quality				**Local Precision**			
Config	DINO ↑	DreamSim ↓	SigLIP2-I ↑	SigLIP2-T ↑	SigLIP2-T ↑	DINO ↑	DreamSim ↓	SigLIP2-I ↑
Mask-free Baseline ($\alpha = 0$)	0.977	0.087	0.949	0.754	0.453	0.878	0.227	0.890
Baseline	0.986	0.065	0.957	0.773	0.464	0.870	0.239	0.881
w/o Discrete Gate	0.979	0.089	0.946	**0.784**	0.463	0.875	0.237	0.885
BRIDGE	**0.988**	**0.054**	**0.965**	0.776	**0.503**	**0.902**	**0.175**	**0.907**

4.2 FLUX.2 Cross-Backbone Results

We instantiate BRIDGE on FLUX.2 [klein] 9B Base (Black Forest Labs, 2026) with full-transformer training, layer-wise gates, and subject-reference conditioning. Main predicts the full canvas and Sub uses an aligned crop. BBox supervision retains Sub quality; optional inference-time token cropping tightens mask support (Fig. 4). Appendix G.3 reports the irregular-mask stress test and indexes the full STE animations in the supplement.

For ICE-Bench (Pan et al., 2025), GPT-6 Luna rewrites placeholder-based instructions into self-contained prompts. We evaluate the step-1,400 Dense/BBox checkpoint with CFG 4 and 50 steps; Appendix G.4 reports the protocol and per-task scores. Task-by-task visual results from the same Dense/BBox run are shown in Appendix G.7.

Table 5: **ICE-Bench mask-group comparison.** Original-model scores are reconstructed from ICE-Bench (Pan et al., 2025); later scores are from Emu3.5 (Cui et al., 2025). † Ours was not trained for Inpainting (Task 17), Outpainting (Task 18), or Face Swap (Task 31); these are diagnostic. For Task 17, the model receives the original source image and a separate `SourceMask`; Luna rewrites the prompt from the source and mask. Details are in the Appendix.

Open-source models

Original ICE-Bench models

Model	19–22	17	18	29–30	31
ACE++ (Mao et al., 2025)	0.5830	0.6430	**0.6230**	0.6066	0.5396
ACE (Han et al., 2025)	0.5695	0.5954	0.5528	0.5722	0.4907
OmniGen (Xiao et al., 2025)	0.5364	0.5454	0.5733	0.5759	0.5324
Ours (FLUX.2 BBox-Sub)†	**0.5941**	**0.6506**	0.5801	**0.6254**	**0.5700**

Later open-source models

Model	17–22	29–30	31
Qwen-Edit-2509 (Wu et al., 2025)	0.587	0.595	0.539
Emu3.5 (Cui et al., 2025)	**0.613**	0.624	0.528
Ours (FLUX.2 BBox-Sub)†	0.6012	**0.6254**	**0.5700**

Closed models: architectures and training recipes undisclosed

Closed model	17–22	29–30	31
Gemini 2.5 Flash Image Preview (Cui et al., 2025)	0.615	0.617	0.565
Gemini 2.5 Flash Image (Cui et al., 2025)	**0.617**	**0.627**	**0.595**

Among compared open-source models, FLUX.2 BRIDGE leads subject-guided editing (Tasks 29–30, 0.6254) and the diagnostic Face Swap evaluation (0.5700). Its four supported local-edit tasks average 0.5941. Later papers report only the combined Tasks 17–22 score; Gemini references are separated as closed models. With the prompt fixed, BRIDGE remains reliable across Local Subject Addition mask variants; in the Add-a-glass case, it consistently inserts a plausible glass, while ACE++ is inconsistent and Q-Control fails (Appendix G.6).

5 Conclusion

BRIDGE separates spatial support from subject geometry with full-canvas Main and local Subject paths plus learned token-level PE routing. Main preserves the scene, Subject generates local content, and optional mask-cropped inference tightens irregular support. Qwen results improve local precision and transfer to MagicBrush and ICE-Bench, while the FLUX.2 instantiation extends the design to full-transformer, subject-reference training.

The model currently assumes a mask. Without one, full-canvas Subject support makes both paths generate complete canvases and reduces quality relative to the base model. Disabling Subject does not restore mask-free generation: in our no-Sub diagnostic, Main showed no visible prompt-driven change. Future work will unify masked editing and mask-free generation in one model.

AI USE STATEMENT

Generative AI tools were used in the research workflow for dataset construction and experimental preprocessing. A vision-language model assisted local-edit concept discovery, caption generation, and semantic filtering. Qwen-Image-Edit-2511 generated subject-reference conditions used to construct the FLUX training data. GPT-6 Luna rewrote ICE-Bench prompts into model-compatible natural-language instructions using the original source images, binary masks, and available reference images. Generative AI tools were also used for related-work retrieval and discovery and for coding assistance, including code creation, editing, debugging, testing, and verification. The authors reviewed the AI-assisted data transformations, prompts, code, and experimental artifacts and take responsibility for the final paper, claims, code, and results.

REPRODUCIBILITY STATEMENT

Sections 3–4 and the appendix document the model, training objective, data construction, fixed training configuration, evaluation protocols, ablations, and supporting results.

REFERENCES

O. Avrahami, O. Fried, and D. Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42(4):149:1–149:11, 2023. doi: 10.1145/3592450.

S. Bai, Y. Cai, R. Chen, K. Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Y. Bengio, N. Léonard, and A. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Black Forest Labs. FLUX.1 Fill [dev]. Official model documentation, 2024. URL https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev.

Black Forest Labs. FLUX.2 [klein] 9B Base model card. Hugging Face model card, 2026. URL https://huggingface.co/black-forest-labs/FLUX.2-klein-base-9B. Accessed 2026-09-26.

T. Brooks, A. Holynski, and A. A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *CVPR*, 2023.

M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, pp. 22560–22570, 2023.

N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Rädle, T. Afouras, E. Mavroudi, K. Xu, et al. SAM 3: Segment anything with concepts. In *ICLR*, 2026.

G. Couairon, J. Verbeek, H. Schwenk, and M. Cord. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *ICLR*, 2023.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. Emu3.5: Native multimodal models are world learners, 2025.

A. Defazio, X. A. Yang, H. Mehta, K. Mishchenko, A. Khaled, and A. Cutkosky. The Road Less Scheduled. In *NeurIPS*, 2024.

DiffSynth-Studio. DiffSynth-Studio/Qwen-Image-Blockwise-ControlNet-Inpaint. Official model documentation, 2025. URL https://modelscope.cn/models/DiffSynth-Studio/Qwen-Image-Blockwise-ControlNet-Inpaint.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling Rectified Flow Transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *NeurIPS*, 2023.

A. B. Gokmen, Y. Ekin, B. B. Bilecen, and A. Dundar. RoPECraft: Training-free motion transfer with trajectory-guided RoPE optimization on diffusion transformers. In *NeurIPS*, 2025.

Z. Han, Z. Jiang, Y. Pan, J. Zhang, C. Mao, C. Xie, Y. Liu, and J. Zhou. ACE: All-round creator and editor following instructions via Diffusion Transformer. In *ICLR*, 2025.

K. He, J. Sun, and X. Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1397–1409, 2013.

A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-Prompt image editing with Cross-Attention Control. In *ICLR*, 2023.

J. Ho and T. Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-Rank Adaptation of large language models. In *ICLR*, 2022.

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 *ECCV*, pp. 150–168. Springer, 2024.

J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang. MUSIQ: Multi-scale image quality transformer. In *ICCV*, pp. 5148–5157, 2021.

S. Lin, B. Liu, J. Li, and X. Yang. Common diffusion noise schedules and sample steps are flawed. In *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5404–5411, 2024. URL https://openaccess.thecvf.com/content/WACV2024/html/Lin_Common_Diffusion_Noise_Schedules_and_Sample_Steps_Are_Flawed_WACV_2024_paper.html.

C. Mao, J. Zhang, Y. Pan, Z. Jiang, Z. Han, Y. Liu, and J. Zhou. ACE++: Instruction-based image creation and editing via Context-Aware Content Filling. In *Proc. of IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1979–1987, 2025.

K. Mishchenko and A. Defazio. Prodigy: An expeditiously adaptive parameter-free learner. In *ICML*, 2024.

R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023.

Y. Pan, X. He, C. Mao, Z. Han, Z. Jiang, J. Zhang, and Y. Liu. ICE-Bench: A unified and comprehensive benchmark for image creating and editing. In *ICCV*, pp. 16586–16596, 2025.

W. Peebles and S. Xie. Scalable diffusion models with transformers. In *ICCV*, pp. 4195–4205, 2023.

Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-Banana-400K: A large-scale dataset for text-guided image editing. In *CVPR*, pp. 37226–37235, 2026.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2149–2159, 2022.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. OminiControl: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14940–14950, 2025.

M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1921–1930, 2023.

C. Wu, J. Li, J. Zhou, J. Lin, K. Gao, K. Yan, S.-m. Yin, S. Bai, X. Xu, Y. Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified image generation. In *CVPR*, pp. 13294–13304, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xiao_OmniGen_Unified_Image_Generation_CVPR_2025_paper.html.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. BoxDiff: Text-to-image synthesis with training-free box-constrained diffusion. In *ICCV*, pp. 7452–7461, 2023.

Peilin Xiong, Junwen Chen, Honghui Yuan, and Keiji Yanai. PosBridge: Multi-view Positional Embedding transplant for identity-aware image editing. In *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24–27, 2025*. BMVA, 2025. URL https://bmva-archive.org.uk/bmvc/2025/assets/papers/Paper_348/paper.pdf.

K. Zhang, L. Mo, W. Chen, H. Sun, and Y. Su. MagicBrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS*, 2023.

H. Zhao, X. Ma, L. Chen, S. Si, R. Wu, K. An, P. Yu, M. Zhang, Q. Li, and B. Chang. UltraEdit: Instruction-based fine-grained image editing at scale. In *NeurIPS*, 2024.

Peiang Zhao, Han Li, Ruiyang Jin, and S Kevin Zhou. LoCo: Training-free layout-to-image synthesis with localized constraints. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9481–9490, 2025.

Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. LayoutDiffusion: Controllable diffusion model for layout-to-image generation. In *CVPR*, pp. 22490–22499, 2023.

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 *ECCV*, pp. 195–211. Springer, 2024.

A Implementation Details

Training Details. The reported Qwen BRIDGE model uses the pre-trained **Qwen-Image-Edit-2511** transformer (Wu et al., 2025). Rank-**512** LoRA (Hu et al., 2022) adapts selected attention, image/text feed-forward, modulation, and input/output projection layers, while the **GateBlocks** are trained from scratch. The final linear head of each GateBlock is zero-initialized. We use the **Prodigy** (Mishchenko & Defazio, 2024) optimizer with a Schedule-Free learning-rate policy (Defazio et al., 2024) (initial LR=1.0). Training uses **5** $\times$ **NVIDIA RTX PRO 6000 Blackwell (96GB)** GPUs for **30K micro-steps**, corresponding to about **1.875K optimizer updates**. Gradient accumulation is 16 per process, giving an effective total batch size of $5 \times 16 = 80$ and about 3.54 passes over the 42,425-pair training set ($1{,}875 \times 80/42{,}425 \approx 3.54$). The final fine-tuning checkpoint contains **4,725,211,136** LoRA parameters and **13,313,340** GateBlock parameters. Classifier-Free Guidance (CFG) training (Ho & Salimans, 2021) drops the conditioning text with probability 10%. Before deployment, the LoRA weights are fused into the backbone. Runtime overhead then comes from BridgePath, primarily the Subject Path.

Training Data Flow and Path Supervision. Each processed record separates the source-image condition from the two generation targets. The `image` field is the full background-preserved edited target for the Main Path. The `edit_image` field supplies the source image to the Qwen visual-conditioning stream. Forced compositing uses source-guided boundary refinement, erosion for removal, dilation for addition and replacement, 33×33 Gaussian smoothing, and alpha blending. The refined soft support extracts the edited foreground onto white; its crop is rounded outward to a 16-pixel grid and encoded by the frozen VAE as `sub_input_latents` for Subject Path supervision. Table 6 summarizes the complete mapping.

Table 6: Training-data and tensor flow for the reported Qwen BRIDGE model. Main and Subject targets receive independently sampled Gaussian noise at the same sampled timestep.

Field/tensor	Semantic and spatial content	Training role	Inference role
`prompt` / T	Edit instruction; text sequence	Joint DiT conditioning	Joint DiT conditioning
`edit_image` / z_{edit}	Source image; full image	Clean visual-condition tokens	Source-image condition
`image` / z_0^{main}	Background-preserved edited target; full canvas	Main target with noise z_1^{main}	Replaced by Main noise
`ref_gt` / z_0^{sub}	Edited foreground on white; 16-pixel-aligned support crop	Subject target with noise z_1^{sub}	Replaced by Subject noise
`back_mask`	Edit support; full image and aligned bbox	Defines Subject support	Defines support and blending
Main output	Final edited image; full canvas	Main flow-matching prediction	Returned and decoded
Subject output	Local subject representation; bbox-local canvas	Subject flow-matching prediction	Decoded for diagnostics

The Main and Subject targets are noised separately: $z_1^{\text{main}} \sim \mathcal{N}(0, I)$ and $z_1^{\text{sub}} \sim \mathcal{N}(0, I)$ are independent draws. At each DiT block, instruction tokens are jointly processed with a visual sequence ordered as Main tokens, Subject tokens, and source-image condition tokens. The GateBlock observes this visual sequence and produces routing values used for Subject-token PE selection. At inference, the pipeline returns the decoded Main Path as the edited image; the Subject Path is also decoded for analysis. External background blending is applied to the Main latent outside the selected support after each denoising update.

Controlled Ablation Budget. All trained variants in Table 4 use the same Qwen-Image-Edit-2511 backbone, rank-512 LoRA configuration, BRIDGE training set, optimizer and hyperparameters, 5 $\times$ 96GB hardware configuration, and 30K-micro-step budget. The Baseline uses a single generation path. The "w/o Discrete Gate" variant adds BridgePath with fixed positional assignments, and BRIDGE adds adaptive per-token PE routing.

Supplementary Ablation Study. Table 7 presents the complete supplementary ablation study on BRIDGE-Bench. It adds the earlier single-path data control to the four inference, architecture, and routing configurations reported in the main-paper ablation.

Table 7: Complete supplementary ablation study on BRIDGE-Bench. The table combines the earlier single-path data control with the mask-free inference, single-path, fixed-routing, and adaptive-routing configurations. Lower DreamSim is better; all other metrics are higher-is-better.

	Global Quality				Local Precision			
Configuration	DINO ↑	DreamSim ↓	SigLIP2-I ↑	SigLIP2-T ↑	SigLIP2-T ↑	DINO ↑	DreamSim ↓	SigLIP2-I ↑
Earlier single-path data control	0.986	0.057	0.959	0.712	0.402	0.864	0.263	0.873
Mask-free Baseline ($\alpha = 0$)	0.977	0.087	0.949	0.754	0.453	0.878	0.227	0.890
Baseline	0.986	0.065	0.957	0.773	0.464	0.870	0.239	0.881
w/o Discrete Gate	0.979	0.089	0.946	**0.784**	0.463	0.875	0.237	0.885
BRIDGE	**0.988**	**0.054**	**0.965**	0.776	**0.503**	**0.902**	**0.175**	**0.907**

The PE diagnostic uses Qwen-Image with 40 inference steps, seed 0, and CFG 4. The reported Qwen BRIDGE results use Qwen-Image-Edit-2511 with the training configuration above, CFG 2, and latent-blending strength $\alpha = 0.1$ unless a benchmark-specific protocol states otherwise. MagicBrush and ICE-Bench follow their respective protocols below.

Table 8 gives the parameter breakdown of the final fine-tuning checkpoint.

Table 8: Parameter breakdown of the final BRIDGE fine-tuning checkpoint. The four LoRA rows sum to the reported LoRA total.

Component	**Parameters**
Attention-projection LoRA	1,509,949,440
Image/text feed-forward LoRA	1,887,436,800
Modulation-projection LoRA	1,321,205,760
Input/output-projection LoRA	6,619,136
LoRA total	**4,725,211,136**
GateBlocks	13,313,340
All trained parameters	**4,738,524,476**

STE parameter accounting across backbones. The straight-through routing module has a backbone-dependent parameterization. The counts below include only learned STE/GateBlock parameters; the threshold buffer τ is not learnable and is therefore excluded from the STE total. If checkpoint tensor storage is counted instead, the single scalar buffer adds one element. For a model with L transformer layers, the learned parameter count is

$$P_{\text{Qwen}}(L) = 221{,}889L, \qquad P_{\text{FLUX.2 Klein}}(L) = 624{,}001L.$$

Thus, including the scalar τ buffer in checkpoint tensor storage gives 13,313,341 elements for Qwen-Image and 19,968,033 for either FLUX.2 Klein variant.

Table 9 summarizes the backbone-specific learned parameter counts.

Table 9: Backbone-specific parameter accounting for the straight-through routing module. Sparse/Mask and Dense/BBox FLUX.2 variants use the same GateBlock shape and therefore have the same parameter count.

Backbone	**Layers**	**Input dim.**	**STE dim.**	**Per-layer**	**STE total**
Qwen-Image	60	3,072	64	221,889	13,313,340
FLUX.2 Klein Sparse/Mask	32	4,096	128	624,001	19,968,032
FLUX.2 Klein Dense/BBox	32	4,096	128	624,001	19,968,032

GateBlock Architecture. At layer l, the gate observes $H^l = [Z^l_{\text{main}}; Z^l_{\text{sub}}; Z^l_{\text{cond}}] \in \mathbb{R}^{N_l \times 3072}$, the visual sequence containing Main, Subject, and source-image condition features. Each of the 60 DiT layers has its own GateBlock with independent parameters. A GateBlock projects from 3072 to 64 channels, applies a single-layer Transformer encoder, and predicts scalar logits whose Subject-token entries control PE routing. Subject tokens occupy the bbox-local range on the Subject

Path canvas. The hard forward decision uses the inclusive 0.5 threshold, and the sigmoid probability supplies the straight-through gradient (Bengio et al., 2013).

Gate Initialization and Visual-Ablation Identity. Zero-initialized gate heads produce $p = 0.5$. With the inclusive threshold, the initial hard forward route selects PE_{base}; the straight-through estimator changes gradients only. Figure 6 labels the Main/Subject output strips and maps the left/right image in every pair to Full BRIDGE/fixed-PE BridgePath.

Inference Configuration. The evaluation settings are as follows:

- **Inference Parameters:** Our default evaluation uses a classifier-free guidance (CFG) (Ho & Salimans, 2021) scale of 2.0. MagicBrush follows a separate protocol: the main-paper BRIDGE result uses CFG$= 4$ with mask-based blending, and Table 12 reports additional model and inference settings, including the Qwen-Image-edit ablation baseline with CFG$= 4$. To mitigate the potential saturation and over-exposure artifacts derived from high guidance scale, we apply the CFG Rescaling trick (Lin et al., 2024). Specifically, we rescale the guided noise prediction to match the norm of the conditional noise prediction: $\epsilon_{pred} \leftarrow \epsilon_{pred} \cdot \frac{\|\epsilon_{pos}\|}{\|\epsilon_{pred}\|}$. Here, ϵ_{pos} denotes the conditional noise prediction. Baseline models (e.g., FLUX (Black Forest Labs, 2024)) are evaluated using their respective official default inference settings.

- **Background Preservation:** The user mask determines the support for latent blending (Avrahami et al., 2023). Unless otherwise noted, experiments use bounding-box support M_{bbox}; the main MagicBrush result uses mask-based support, and Table 12 includes a bbox-support variant. The default blending strength is $\alpha = 0.1$, where larger α retains more source-image background. For bbox support, each denoising step blends the Main latent outside the bounding box $(1 - M_{bbox})$ with source-image latents noised to the same timestep:

$$z_{t-1}^{blend} = z_{t-1} \cdot M_{bbox} + \big((1-\alpha) \cdot z_{t-1} + \alpha \cdot z_{t-1}^{orig}\big) \cdot (1 - M_{bbox}) \tag{7}$$

  where z_{t-1}^{orig} denotes the source-image latents corrupted to timestep $t-1$. This operation preserves background context while allowing generation throughout the bbox support.

Figure 3 Diagnostic Setting. Figure 3 uses the official DiffSynth-Studio Qwen-Image Quick Start setting: model `Qwen/Qwen-Image`, prompt "精致肖像，水下少女，蓝裙飘逸，发丝轻扬，光影透澈，气泡环绕，面容恬静，细节精致，梦幻唯美。", seed 0, 40 inference steps, and the pipeline-default CFG of 4.0. The panels vary only PE assignment and cross-region attention. The main BRIDGE setting uses CFG$= 2$, while MagicBrush uses CFG$= 4$.

Main-Setting Grid Search for CFG and Blending Alpha. We additionally performed a grid search over CFG $\in \{1, 2, 4\}$ and blending $\alpha \in \{0.0, 0.1, 0.4, 0.7, 1.0\}$ for the main setting. The table image in Fig. 8 reports the aggregated metrics for all 15 configurations, and the line plot in Fig. 9 highlights the two criteria we used most directly for selection: local DINO within the edit bounding box and global DreamSim. Among the tested settings, CFG$= 2$ with $\alpha = 0.1$ simultaneously achieves the best local DINO (0.5317) and the best global DreamSim (0.7963), which is why we use this pair for the main experiments. MagicBrush uses the separate inference setting stated above.

CFG	Alpha	DINO	DreamSim	Local DINO	Local DreamSim	Selected
1	0.0	0.610	0.817	0.499	0.799	
1	0.1	0.610	0.819	0.500	0.799	
1	0.4	0.620	0.811	0.489	0.801	
1	0.7	0.619	0.813	0.487	0.799	
1	1.0	0.619	0.812	0.496	0.814	
2	0.0	0.631	0.811	0.500	0.827	
2	0.1	0.631	0.796	0.532	0.801	**Yes**
2	0.4	0.632	0.797	0.528	0.797	
2	0.7	0.632	0.801	0.528	0.801	
2	1.0	0.629	0.798	0.527	0.799	
4	0.0	0.630	0.802	0.528	0.819	
4	0.1	0.632	0.802	0.531	0.810	
4	0.4	0.631	0.802	0.529	0.800	
4	0.7	0.631	0.805	0.528	0.780	
4	1.0	0.632	0.800	0.530	0.802	

Figure 8: **Grid-search summary for the main setting.** We compare 15 combinations of CFG and latent-blending strength α. The highlighted row marks the selected configuration used in the main experiments.

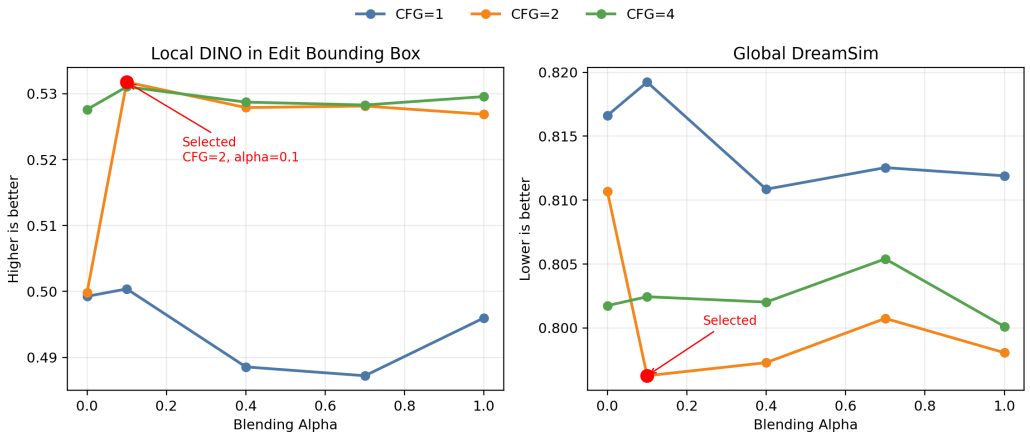


Figure 9: **Selection rationale for CFG and blending alpha in the main setting.** The left panel plots local DINO within the edit bounding box, and the right panel plots global DreamSim. The selected setting, CFG$=2$ and $\alpha = 0.1$, is highlighted in red because it provides the strongest joint trade-off between local edit fidelity and global perceptual consistency among the tested combinations.

Historical MagicBrush Global-Description Results. Table 10 reports the earlier protocol that pairs outputs with global descriptions. Main-paper Table 2 uses current-turn local descriptions.

Table 10: Historical zero-shot results on the MagicBrush test set using global descriptions. Final Turn and All Turn are reported separately, and Avg. Rank averages the ranks over the 10 displayed metrics.

Method	Final Turn					All Turn					Avg. Rank↓
	L1↓	L2↓	CLIP-I↑	DINO↑	CLIP-T↑	L1↓	L2↓	CLIP-I↑	DINO↑	CLIP-T↑	Avg. Rank
ACE++	0.081	0.032	0.891	0.817	0.307	0.047	0.017	0.936	0.894	0.306	3.7
FLUX.1-Fill	**0.075**	**0.025**	0.898	0.839	0.297	0.044	**0.013**	**0.947**	**0.919**	0.300	2.1
Q-Control	0.078	0.030	0.898	**0.840**	0.322	0.044	0.016	0.943	0.912	0.315	2.2
BRIDGE	0.076	0.032	**0.907**	0.839	**0.324**	**0.043**	0.017	0.946	0.910	**0.315**	**2.0**

MagicBrush Current-Protocol Evaluation. Main-paper Table 2 evaluates full output images paired with current-turn local descriptions. Table 11 evaluates a common tight mask-defined crop for

every method and target. The crop evaluation uses this shared local image domain and mask-based blending.

Table 11: Additional MagicBrush evaluation on the common tight edit-region crop (all ↑). Main-paper Table 2 evaluates full images; this table evaluates the shared local crop.

Split	Method	CLIP-I	CLIP-T Local	SigLIP2-I	SigLIP2-T Local
Final	ACE++	0.7978	0.2607	0.7559	0.1864
	FLUX.1-Fill	0.8054	0.2580	0.7743	0.1397
	Q-Control	0.8150	0.2684	0.7932	0.2935
	BRIDGE	**0.8215**	**0.2727**	**0.8011**	**0.3377**
All	ACE++	0.8113	0.2590	0.7797	0.1713
	FLUX.1-Fill	0.8228	0.2591	0.7999	0.1522
	Q-Control	0.8317	0.2676	0.8127	0.2985
	BRIDGE	**0.8383**	**0.2715**	**0.8208**	**0.3240**

Additional MagicBrush Inference Variants. Table 12 summarizes supplementary MagicBrush evaluations by model and inference setting. The first row is the Qwen-Image-Edit ablation baseline evaluated with CFG$= 4$ and mask blending. The other two rows are BRIDGE variants comparing CFG$= 2$ with mask blending and CFG$= 4$ with bbox blending.

Table 12: Supplementary MagicBrush model and inference settings.

Model / Setting	CFG	Blend	Final Turn					All Turn				
			L1↓	L2↓	CLIP-I↑	DINO↑	CLIP-T↑	L1↓	L2↓	CLIP-I↑	DINO↑	CLIP-T↑
Ablation baseline (Qwen-Image-edit)	4	mask	0.0808	0.0324	**0.9020**	0.8188	0.3230	0.0468	0.0178	**0.9418**	0.8953	0.3140
BRIDGE (ours)	2	mask	0.0815	0.0329	0.9006	0.8177	0.3224	0.0472	0.0181	0.9412	0.8949	0.3136
BRIDGE (ours)	4	bbox	0.0838	0.0355	0.8997	0.8174	**0.3246**	0.0484	0.0195	0.9385	0.8910	**0.3155**

B Data Curation and BRIDGE-Bench Construction

Raw Corpus and Processed Split. Our raw source corpus is derived from Pico-Banana-400K (Qian et al., 2026), a collection of approximately 400,000 text-guided image-editing examples based on real-world OpenImages photographs. Its dual-model pipeline uses a multimodal model to generate edits and a reasoning model for quality auditing, spanning a taxonomy of 35 edit types. We select five categories that require structural modification: Category Replacement, Object Addition, Object Removal, Clothing Edit, and Accessory Modification. We create the paper's train/test split after the filtering and compositing steps described below.

Precision Segmentation and Filtering. Raw generative outputs typically lack precise ground-truth segmentation masks. To address this, we re-engineered the Segment Anything Model 3 (SAM3) (Carion et al., 2026) into an efficient parallel generate-and-filter pipeline. Driven by fine-grained sub-prompts generated by a Vision-Language Model (Qwen3-VL-32B (Bai et al., 2025)), SAM3 generates candidate masks. The VLM then performs a "Zoomed-in Analysis" on cropped mask regions to assign semantic confidence scores, retaining only highly aligned mask proposals with a strict confidence threshold of $s \geq 0.95$.

Dual Audit and Background Consistency Verification. To reduce background drift and improve background consistency, we implement a dual audit:

- **Individual Mask Filtering (Object Change):** We compute the DreamSim (Fu et al., 2023) distance between the source and target cropped regions. We enforce a minimum threshold of $d_{obj} > 0.25$. Regions failing to meet this threshold represent negligible perceptual changes and are discarded as false positives.
- **Global Background Audit:** We compute the union of all valid masks to form a global foreground mask. The background DreamSim distance (calculated on the inverted mask

region) must be below a cut-off threshold ($d_{bg} < 0.6$), which discards samples with large global shifts outside the intended edit region.

Finally, for all retained samples, we apply a "Forced Compositing" strategy using guided filters (He et al., 2013) and morphological operations (e.g., erosion for removal, dilation for addition) paired with Gaussian alpha blending. This process weakens explicit shape boundaries, better simulates messy human user scribbles, and makes the retained background nearly identical at the pixel level ($L_1 \approx 0$).

Curating the BRIDGE-Bench Evaluation Set. We construct BRIDGE-Bench from the cleaned pool using a stricter background threshold (e.g., DreamSim $D_{bg} < 0.4$) and task-class balancing; the remaining pairs form the training split. We audit the evaluation pairs using seam quality along the expanded mask boundary and fine-grained background perceptual distance.

Figure 10 shows selected processed examples from our internal dataset after the full filtering and compositing pipeline.



Figure 10: **Selected processed examples from our internal dataset.** These examples illustrate the instruction-guided local edits retained after semantic filtering, background auditing, and forced background compositing.

Matched Common-ID Evaluation. Table 13 reports a matched evaluation using the sample identifiers available for all three methods. Each metric is aggregated over the same identifier set within its evaluator. The numerical differences from main-paper Table 1 arise from this common-ID restriction.

Table 13: Matched common-ID evaluation on BRIDGE-Bench. The metric order matches main-paper Table 1.

	Local Precision				Global Quality			
Method	**Sig-T↑**	**DINO↑**	**DSim↓**	**Sig-I↑**	**DINO↑**	**DSim↓**	**Sig-I↑**	**Sig-T↑**
ACE++	0.392	0.852	0.303	0.856	0.982	0.082	0.954	0.705
FLUX.1-Fill	0.262	0.807	0.321	0.835	0.978	0.076	0.947	0.631
BRIDGE	**0.507**	**0.903**	**0.173**	**0.908**	**0.988**	**0.053**	**0.965**	**0.778**

For naming clarity, the reported SigLIP2-T values are model-native sigmoid scores,

$$\mathrm{sigmoid}(\cos(\widehat{e}_I, \widehat{e}_T)\ \exp(s) + b)\,, \tag{8}$$

with the model's learned log-scale parameter s and bias b; they are not raw embedding cosines. We use the abbreviation "SigLIP2-T" for this score throughout the paper.

C Additional ICE-Bench Results

Data Preprocessing and Prompt Standardization. We evaluate 1,171 ICE-Bench samples across five local-editing tasks with the following modified-input and standardized-prompt protocol:

- **Prompt template:** Each instruction begins with "Picture 1 is the image to modify.", and <SOURCE> is replaced with "Picture 1".
- **Region terminology:** Inpainting and Removal instructions replace "mask" references with "black area"; Addition and Text Render instructions remove those references.
- **Removal inputs:** For *Local Subject Removal* and *Local Text Removal*, pixels inside the source mask are set to black before inference, matching the removal-specialized input format.

Score Definition. For ICE-Bench local editing tasks (Tasks 17–22) (Pan et al., 2025), we map raw metrics into four normalized dimensions, where musiq_koniq follows MUSIQ (Ke et al., 2021):

$$S_{\text{AES}} = \frac{\text{aes_v2.5}}{10}, \quad S_{\text{IMG}} = \frac{\text{musiq_koniq}}{100}, \tag{9}$$

$$S_{\text{PF}} = \frac{2 \cdot \text{CLIP-cap} + \text{VLLM-QA}}{2}, \quad S_{\text{SRC}} = \frac{\text{CLIP-src} + (1 - \text{L1-src})}{2}. \tag{10}$$

The final score for each task is computed as:

$$\text{TaskScore} = 0.3S_{\text{AES}} + 0.3S_{\text{IMG}} + 0.3S_{\text{PF}} + 0.1S_{\text{SRC}}. \tag{11}$$

We report results on five tasks (Tasks 17, 19–22); Task 18 (Outpainting) is excluded in our evaluation. The reported overall score is the equal-weight macro average over these tasks rather than a pooled sample average.

Results Analysis. As shown in Table 14, our method achieves the highest average score in **Aesthetic Quality (AES)** and competitive performance in **Source Consistency (SRC)**. A high AES score (e.g., in Task 17 and Task 19) indicates that our generated results possess superior visual appeal and composition, which is critical for user preference in real-world editing scenarios. High SRC scores (e.g., 0.942 average) demonstrate that BRIDGE effectively preserves the non-edited background regions, a key requirement for local editing tasks. In Task 17 (Inpainting) and Task 19 (Local Subject Addition), our method obtains strong aesthetic and source consistency metrics (see Tables 15 and 16), suggesting that BridgePath geometric guidance helps balance local editing and background preservation.

Table 14: Averaged dimension scores on ICE-Bench local editing (Tasks 17, 19–22). Baseline metrics are taken from the original ICE-Bench paper (Pan et al., 2025).

Method	AES	IMG	PF	SRC
ACE++ (Mao et al., 2025)	0.481	**0.577**	0.616	0.930
ACE (Han et al., 2025)	0.480	0.487	**0.643**	0.916
UltraEdit (Zhao et al., 2024)	0.455	0.458	0.442	**0.951**
BRIDGE	**0.494**	0.565	0.582	0.942

Per-task Raw Metrics. Tables 15, 16, 17, 18, and 19 report the raw metrics (before normalization) for each evaluated task. Baseline rows are taken from ICE-Bench (Pan et al., 2025).

Table 15: ICE-Bench Task 17 (Inpainting) detailed metrics. Baseline metrics are taken from the original ICE-Bench paper (Pan et al., 2025).

Method	Aesthetic	Imaging	CLIP-cap	VLLM-QA	CLIP-src	L1-src
ACE++ (Mao et al., 2025)	5.064	61.661	0.272	0.910	0.776	0.016
ACE (Han et al., 2025)	4.878	51.793	0.269	0.833	0.785	0.024
UltraEdit (Zhao et al., 2024)	3.817	46.284	0.250	0.180	**0.952**	0.019
BRIDGE	**5.214**	**61.685**	**0.278**	**0.953**	0.789	**0.013**

Table 16: ICE-Bench Task 19 (Local Subject Addition) detailed metrics. Baseline metrics are taken from the original ICE-Bench paper (Pan et al., 2025).

Method	Aesthetic	Imaging	CLIP-cap	VLLM-QA	CLIP-src	L1-src
ACE++ (Mao et al., 2025)	5.014	**62.083**	0.268	**0.785**	0.894	0.018
ACE (Han et al., 2025)	4.965	51.704	0.272	0.555	0.897	0.029
UltraEdit (Zhao et al., 2024)	4.881	47.855	0.275	0.555	0.909	0.021
BRIDGE	**5.146**	61.239	**0.278**	0.426	**0.937**	**0.013**

Table 17: ICE-Bench Task 20 (Local Subject Removal) detailed metrics. Baseline metrics are taken from the original ICE-Bench paper (Pan et al., 2025).

Method	Aesthetic	Imaging	CLIP-cap	VLLM-QA	CLIP-src	L1-src
ACE++ (Mao et al., 2025)	**5.061**	**61.614**	0.229	0.312	**0.901**	0.017
ACE (Han et al., 2025)	4.996	47.011	0.258	**0.757**	0.852	0.024
UltraEdit (Zhao et al., 2024)	4.858	48.748	0.226	0.287	0.888	0.018
BRIDGE	5.009	57.203	**0.262**	0.748	0.861	**0.013**

Table 18: ICE-Bench Task 21 (Local Text Render) detailed metrics. Baseline metrics are taken from the original ICE-Bench paper (Pan et al., 2025).

Method	Aesthetic	Imaging	CLIP-cap	VLLM-QA	CLIP-src	L1-src
ACE++ (Mao et al., 2025)	4.231	**43.276**	0.277	**0.834**	0.899	0.012
ACE (Han et al., 2025)	4.275	43.159	0.276	0.791	0.860	0.016
UltraEdit (Zhao et al., 2024)	4.506	38.887	0.277	0.098	0.946	0.014
BRIDGE	**4.527**	43.224	**0.279**	0.307	**0.957**	**0.006**

Table 19: ICE-Bench Task 22 (Local Text Removal) detailed metrics. Baseline metrics are taken from the original ICE-Bench paper (Pan et al., 2025).

Method	Aesthetic	Imaging	CLIP-cap	VLLM-QA	CLIP-src	L1-src
ACE++ (Mao et al., 2025)	4.694	**59.636**	0.260	0.704	0.905	0.017
ACE (Han et al., 2025)	**4.896**	49.766	**0.273**	**0.801**	0.888	0.033
UltraEdit (Zhao et al., 2024)	4.665	47.294	0.264	0.714	0.910	0.023
BRIDGE	4.820	58.985	0.271	0.653	**0.937**	**0.012**

Direct Evaluation Reports for Q-Control, ACE++, and BRIDGE. The direct Q-Control, ACE++, and BRIDGE evaluations use the modified-input and standardized-prompt protocol defined above. Table 20 reports per-task final scores and the five-task average, and Table 21 reports the averaged mapped dimensions. Tables 22, 23, 24, 25, and 26 provide the corresponding per-task metrics.

Table 20: ICE-Bench final scores for the three evaluated methods under the standardized protocol. Task 18 (Outpainting) is not included in this evaluation.

Run	T17	T19	T20	T21	T22	Avg.
Q-Control	0.601768	0.613342	0.591195	0.514650	0.590880	0.582367
ACE++	0.642074	0.607212	0.526383	0.557288	0.574833	0.581558
BRIDGE	**0.656471**	0.581629	**0.605180**	0.492648	**0.597094**	**0.586604**

Table 21: Averaged mapped dimension scores for the three evaluated methods on ICE-Bench Tasks 17 and 19–22.

Run	AES	IMG	PF	SRC	Avg. Score
Q-Control	0.472404	0.568606	0.587753	0.937379	0.582367
ACE++	0.479863	**0.575003**	0.575187	0.925421	0.581558
BRIDGE	**0.494323**	0.564674	0.582213	**0.942415**	**0.586604**

Table 22: Per-method metrics for ICE-Bench Task 17 (Inpainting).

Run	Aesthetic	Imaging	CLIP-cap	VLLM-QA	CLIP-src	L1-src
Q-Control	4.353530	61.674982	0.267331	0.765586	0.838662	0.016663
ACE++	5.069124	61.198315	0.276025	0.897756	0.793418	0.014730
BRIDGE	**5.214464**	**61.685315**	**0.277598**	**0.952618**	0.789001	**0.012815**

Table 23: Per-method metrics for ICE-Bench Task 19 (Local Subject Addition).

Run	Aesthetic	Imaging	CLIP-cap	VLLM-QA	CLIP-src	L1-src
Q-Control	5.114608	**62.391083**	0.276779	0.641148	0.887129	0.016649
ACE++	5.075284	61.490830	0.272050	0.631579	0.897898	0.015308
BRIDGE	**5.145709**	61.239281	**0.278256**	0.425837	**0.936747**	**0.013002**

Table 24: Per-method metrics for ICE-Bench Task 20 (Local Subject Removal).

Run	Aesthetic	Imaging	CLIP-cap	VLLM-QA	CLIP-src	L1-src
Q-Control	**5.021658**	58.158275	0.260421	0.633663	**0.873895**	0.015996
ACE++	4.940362	**61.914121**	0.229052	0.207921	0.867261	0.016733
BRIDGE	5.009282	57.203025	**0.262418**	**0.747525**	0.861470	**0.012711**

Table 25: Per-method metrics for ICE-Bench Task 21 (Local Text Render).

Run	Aesthetic	Imaging	CLIP-cap	VLLM-QA	CLIP-src	L1-src
Q-Control	4.330330	**44.016858**	**0.281760**	0.490798	0.901521	0.010693
ACE++	4.266104	43.993627	0.275505	**0.803681**	0.893160	0.010754
BRIDGE	**4.526936**	43.223927	0.278827	0.306748	**0.956576**	**0.006425**

Table 26: Per-method metrics for ICE-Bench Task 22 (Local Text Removal) under the modified-input protocol.

Run	Aesthetic	Imaging	CLIP-cap	VLLM-QA	CLIP-src	L1-src
Q-Control	4.800064	58.061950	0.270548	0.632653	**0.949094**	0.016509
ACE++	4.642299	58.904480	0.262020	0.581633	0.875389	0.015392
BRIDGE	**4.819754**	**58.985408**	**0.271069**	**0.653061**	0.937104	**0.011801**

D ADDITIONAL QUALITATIVE RESULTS

We provide additional qualitative results on BRIDGE-Bench across several editing types. In particular, we highlight cases where the model generates **sharp, detailed structures** that are less constrained by the initial mask shape. When the mask does not match the object's natural geometry, inpainting baselines often produce blurry or box-like artifacts. BRIDGE can synthesize more organic shapes (e.g., hair strands, animal limbs, complex textures) while maintaining background consistency in these examples.

D.1 ADDITIONAL RESULTS: REMOVE

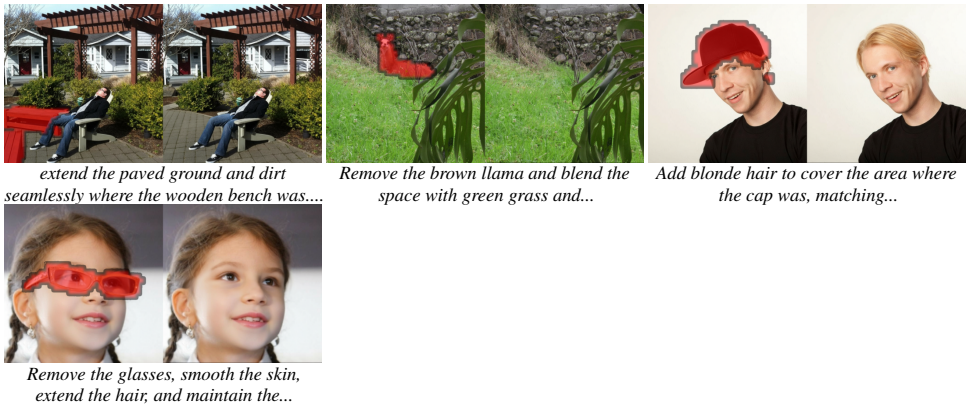


Figure 11: Additional results for **Remove**.

D.2 Additional Results: Add



Figure 12: Additional results for **Add**.

D.3 Additional Results: Replace

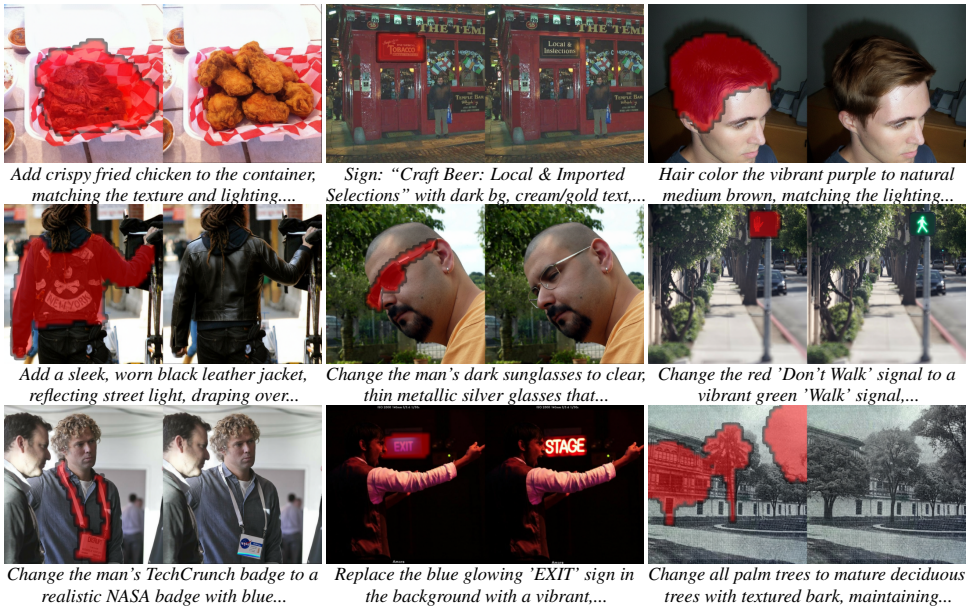


Figure 13: Additional results for **Replace** (1/2).

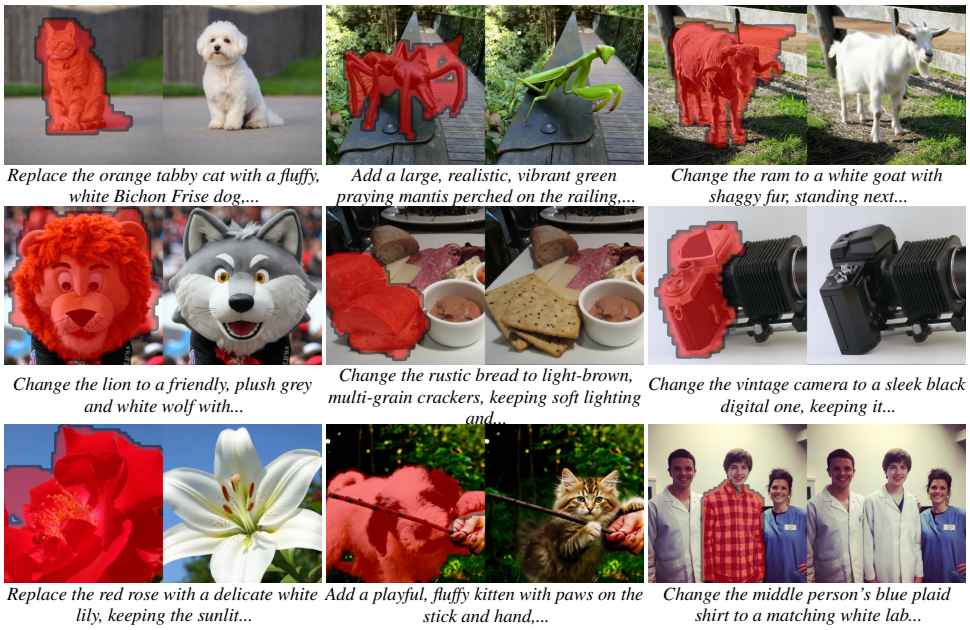


Figure 14: Additional results for **Replace** (2/2).

E SELECTED INSTRUCTION EXAMPLES

To provide further insight into our VLM-driven automated pipeline, we list several representative "Result-Oriented Dense Captions" generated by Qwen3-VL during our data curation process. To address keyword bias and token limits in imperative instructions, these dense captions (incorporating both global and local scene descriptions) ensure that the text aligns closely with the intended visual outcome.

- **Category Replacement:**
  - *Instruction:* "Change the horse to a majestic white unicorn with shimmering coat, silver mane and tail, and an elegant golden horn."
  - *Global Caption:* "A rider in equestrian gear gently touches a majestic white unicorn with a shimmering coat, silver mane and tail, and an elegant golden horn."
  - *Local Caption:* "a white majestic unicorn with silver mane and tail, and an elegant golden horn"
- **Object Addition:**
  - *Instruction:* "Add a perfectly halved, marinated soft-boiled egg (ajitsuke tamago) to the bowl, nestled gently into the broth among the noodles and existing toppings on the left side, ensuring its creamy, golden-orange yolk is visible and its glossy, light brown marinated white reflects the warm overhead light..."
  - *Global Caption:* "A steaming bowl of ramen with noodles, seaweed, mushrooms, cilantro, and a perfectly halved marinated soft-boiled egg on the left, its golden yolk and glossy white reflecting warm light, all under shallow depth of field."
  - *Local Caption:* "A halved marinated soft-boiled egg with creamy golden-orange yolk and glossy light brown white, nestled in broth on the left, reflecting warm overhead light."
- **Object Removal:**
  - *Instruction:* "Remove the child and pony, replacing them with lush green grass matching the existing lawn."
  - *Global Caption:* "A grassy lawn with palm fronds on the left and green barriers in the background, uniformly covered in lush green grass."
  - *Local Caption:* "Lush green grass seamlessly matching the existing lawn, replacing any non-grass elements."
- **Complex Scene Integration:**
  - *Instruction:* "Integrate a sleek, dark metallic condenser microphone on a subtle stand to the right of the woman, positioned slightly in front of her mouth to suggest an interview or performance, ensuring its surface reflects the warm, soft lighting present in the scene and casts a soft shadow consistent with the existing environment."
  - *Global Caption:* "A woman with curly hair smiles into a sleek dark metallic condenser microphone on a subtle stand, positioned slightly in front of her mouth, under warm soft lighting with reflective surfaces and soft shadows."
  - *Local Caption:* "A sleek dark metallic condenser microphone on a subtle stand, reflecting warm soft light and casting a soft shadow, positioned slightly in front of the woman's mouth."
- **Structural Accessory Modification:**
  - *Instruction:* "Change the product box to a modern, glossy, luxurious package with a sophisticated logo, keeping the hand holding it in place."
  - *Global Caption:* "A hand holds a modern, glossy, luxurious blue product box with a sophisticated silver logo, set against a wooden table background."
  - *Local Caption:* "a sleek, glossy blue box with a refined silver circular logo and elegant text, exuding luxury and sophistication"

F SUPPORT AND TRADE-OFF ANALYSIS

Inference Cost. We measured complete-pipeline inference after model loading on an NVIDIA RTX PRO 6000 Blackwell Max-Q using BF16, no offload, 1024×1024 resolution, 50 steps, CFG 4, seed 0, an exact 25% edit region, and one full warmup per process. The matched Qwen and BRIDGE runs use the same fused rank-512 LoRA. Table 27 reports latency and peak process VRAM.

Table 27: End-to-end inference measurement at 1024×1024 with an exact 25% edit region. Latency is mean ± standard deviation and excludes model loading. Qwen and BRIDGE use the same fused rank-512 LoRA; Q-Control uses its released base checkpoint.

Method	n	Latency (s)	vs. Qwen	Peak VRAM (GiB)	vs. Qwen
Qwen baseline	5	91.985 ± 0.205	—	57.354	—
BRIDGE	3	111.649 ± 0.193	+21.38%	57.379	+0.04%
Q-Control	5	112.311 ± 1.908	+22.10%	59.482	+3.71%

For fixed conditioning lengths, Subject Path token count grows approximately with edit area; at full-image editing, it approaches one additional image-token grid. The percentage overhead therefore also depends on the full text/image-conditioning sequence length. The measured BRIDGE latency increase has a 21.07%–21.64% bootstrap 95% interval, while its 26 MiB peak-process increase is consistent with the 13.31M BF16 GateBlock parameters. Dynamic token cropping and packed-latent execution remain directions for reducing redundant Subject Path computation.

G FLUX.2 CROSS-BACKBONE EXTENSION DETAILS

Training records and branch targets. The FLUX.2 extension uses 27,834 training records. Table 28 distinguishes the full-canvas target, source condition, subject-reference condition, and Sub-branch crop target. The subject-reference images were generated with Qwen-Image-Edit-2511 (Wu et al., 2025) and selected for subject identity and visual quality. They condition the model and do not serve as the Sub-branch target.

Table 28: Image roles in the FLUX.2 subject-driven training records.

Record field	Training role
`image`	Full edited target for the Main branch.
`edit_image`	Full-canvas source/background image condition.
`generated_subject_image`	Generated and selected subject-reference image condition.
`sub`	Aligned crop of the edited subject used as the Sub-branch target.
`back_mask`	Spatial support used to derive the bounding box and positional correspondence.

For the evaluated Dense/BBox variant, the Sub branch retains every token in the bounding rectangle derived from `back_mask`. Main and Sub targets use independently sampled Gaussian noise. The visual sequence is ordered as noisy Main tokens, noisy Sub tokens, source/background condition tokens, and, when retained by subject-image dropout, subject-reference condition tokens. The routing module observes the joint visual hidden states. The backbone-specific STE parameter counts are reported in Table 9. The Main and Sub flow-matching losses are combined with equal default weight. The evaluated checkpoint was trained for 1,400 optimizer steps on eight GPUs with a per-GPU micro-batch size of two and gradient accumulation of eight, giving an effective global batch size of 128.

G.1 TRAINING DYNAMICS OF DENSE/BBOX AND SPARSE/MASK RUNS

Figure 15 reports the complete logged training curves supplied for the Dense/BBox and Sparse/Mask checkpoints. The light-cyan curve is the eight-GPU Dense/BBox run (`use_sparse_sub_branch=false`); the dark-green curve is the six-GPU Sparse/Mask run (`use_sparse_sub_branch=true`). Both use equal Main and Sub loss weights. The plots

record total, Main, and Sub losses, selected Sub tokens, gradient norm, and fractional data passes. Both loss trajectories decrease and remain finite over the recorded interval. Because the runs use different GPU counts and effective global batches, equal optimizer steps do not correspond to equal sample exposure; the curves document optimization behavior and are not an equal-budget quality comparison.

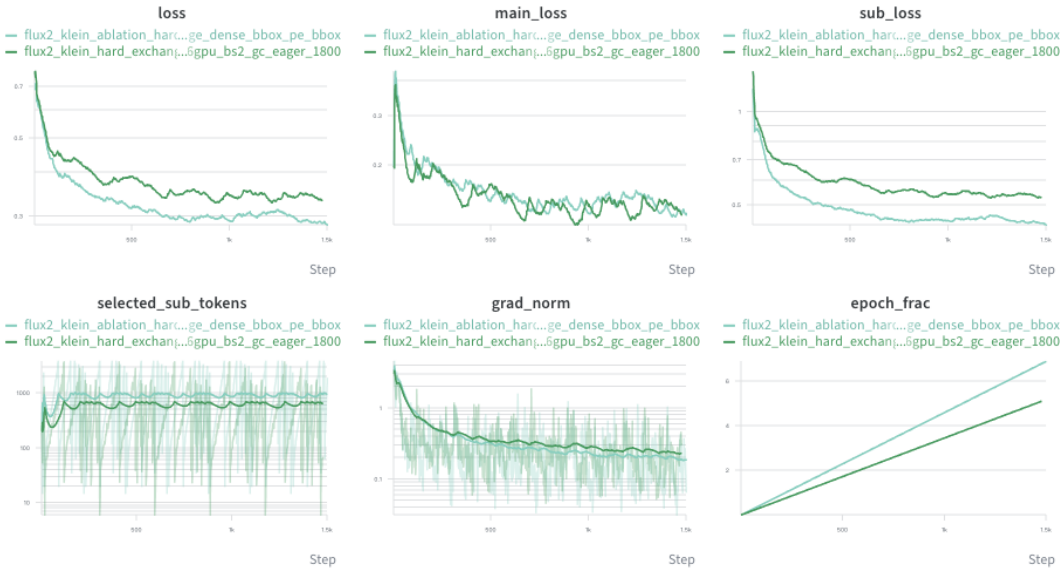


Figure 15: **Training dynamics for the two FLUX.2 checkpoints.** Light cyan denotes Dense/BBox training and dark green denotes Sparse/Mask training. From top left to bottom right: total loss, Main loss, Sub loss, selected Sub-token count, gradient norm, and fractional data passes. The different sample-exposure rates reflect the eight-GPU and six-GPU training configurations.

G.2 Why BBox Training Uses Optional Mask-Cropped Inference

The dense and sparse variants address different failure modes. With a dense BBox Sub branch, every cell in the bounding rectangle receives a Sub token. This preserves a complete local latent grid and produces stronger Sub-image quality, but a non-rectangular user mask leaves rectangular support in regions where the user did not request new content. The model can consequently generate throughout the bounding rectangle.

We initially addressed this mismatch by training a Sparse/Mask model whose Sub sequence retained only mask-selected cells. Development comparisons revealed three recurring problems: the decoded Sub image was less coherent than that of the BBox-trained model, small supports exhibited visible color drift, and the model learned to fill the supplied mask shape. The last behavior reintroduced the mask-shape dependence that BRIDGE is designed to reduce. These observations motivated retaining dense BBox supervision during training, as illustrated in Fig. 16.

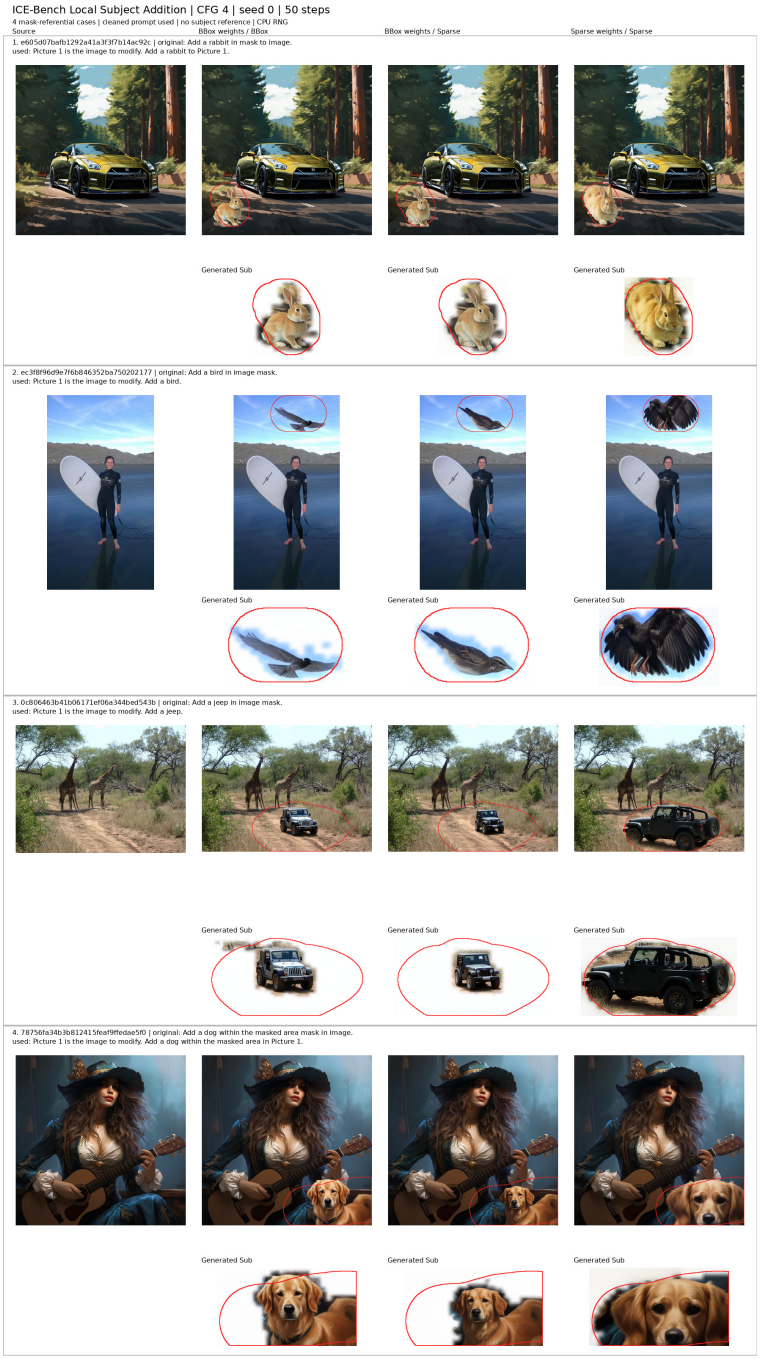


Figure 16: **Qualitative motivation for BBox training.** Four ICE Addition cases at CFG $= 4$ compare the Dense/BBox and separately trained Sparse/Mask variants. In the first example, the mask retains only a small number of Sub tokens; the decoded Sparse/Mask Sub image exhibits an incorrect color distribution and closely fills the supplied mask contour, revealing severe mask-shape overfitting. The remaining examples show the same quality and support-control trade-off. The final design therefore retains BBox training and applies mask cropping only at inference with the same BBox-trained checkpoint.

The final support-control design applies sparsification at inference to the same BBox-trained checkpoint. It removes Sub tokens outside the user mask and restricts the corresponding PE candidate pairs to the retained cells. In the no-Sub diagnostic, changing the prompt did not visibly alter the source

background when the Sub branch was absent. Removing outside-mask Sub support therefore makes those locations approximate the no-Sub condition while the retained local tokens continue to generate the requested subject. This combines the subject quality of BBox training with tighter user-mask support and avoids training the model to reproduce the mask silhouette.

G.3 Coercive-Mask Stress Test and STE Routing Dynamics

We further evaluate the support-control mechanism on a web-sourced qualitative stress case with an intentionally irregular, coercive mask. Both inference arms in Fig. 17 use the same BBox-trained checkpoint, source image, prompt, subject references, submitted mask, initial noise, seed 0, CFG 4, and 50 denoising steps. Latent blending is disabled in both arms. Dense/BBox inference retains all 1,302 Sub tokens inside the enclosing rectangle, whereas mask-cropped inference retains the 618 mask-selected tokens and restricts the PE candidate pairs to the same support. The decoded Sub views make the support mismatch visible: the dense route allocates generation capacity throughout the rectangle, while the cropped route confines active Sub content to the user-selected cells. Earlier Qwen experiments used inference-time α blending for background preservation; this comparison disables blending and isolates Sub-token support control.

Table 29 provides the exact relative paths in the anonymous supplementary archive. Reviewers can begin with the output comparison, then inspect the complete mask-cropped animation, and use the remaining records for the matched Dense/BBox control or layer-specific views.

Table 29: Reviewer file map for the anonymous token-control supplementary archive. Paths are relative to the archive root. Layer indices run from `00` to `31`; denoising-step indices run from `00` to `49`.

Relative path	Content
`README.md`	Quick-start reading order, protocol, file map, and anonymization record.
`comparison.png`	Same-checkpoint Dense/BBox versus mask-cropped Main and decoded-Sub comparison.
`mask_cropped/ste_masks/run/ste_layer_exchange_timestep_grid.gif`	Primary 32-layer, 50-step STE animation for mask-cropped Sub inference.
`dense/ste_masks/run/ste_layer_exchange_timestep_grid.gif`	Matched 32-layer, 50-step Dense/BBox routing control.
`mask_cropped/ste_masks/run/ste_layer_exchange_by_timestep/layer_XX_exchange.gif`	Animation for one selected transformer layer.
`mask_cropped/ste_masks/run/ste_layer_exchange_by_timestep/layer_XX/step_YY.png`	Raw hard-forward routing mask for one layer and denoising step.
`experiment.json`	Effective CFG, step count, seed, checkpoint role, and no-blending setting.
`verification.json`	Matched-condition checks and the 1,302/618 Sub-token counts.

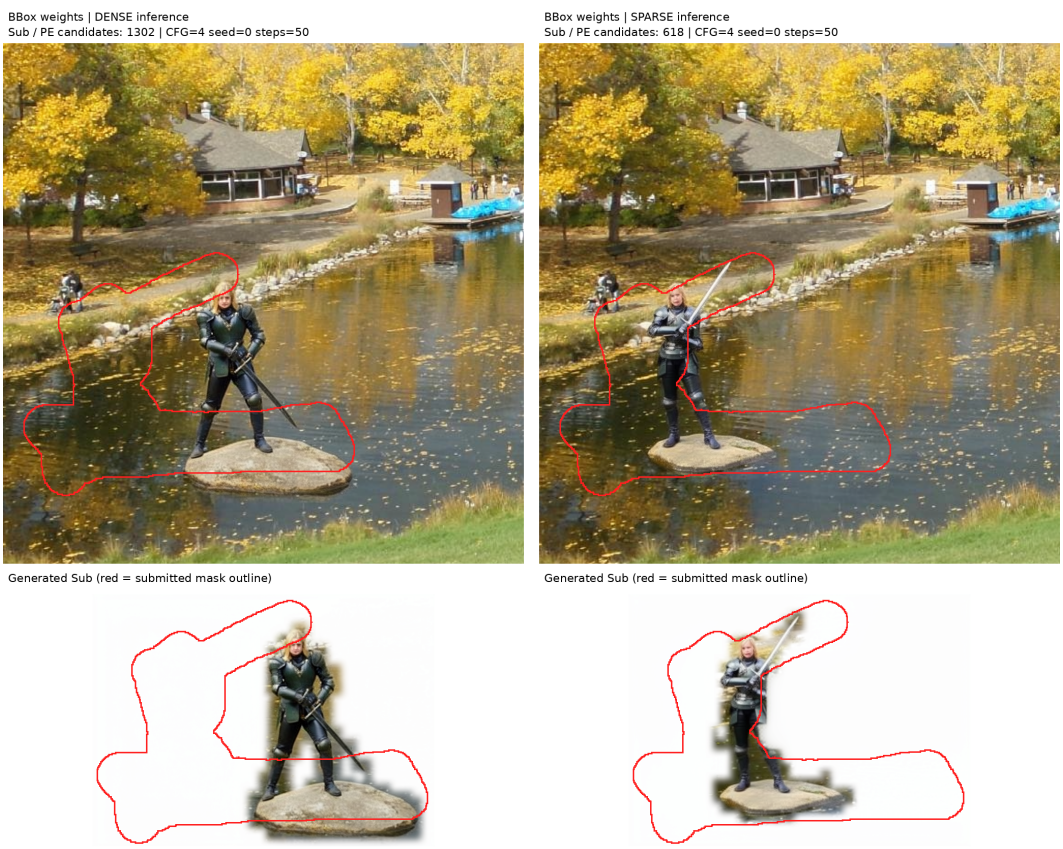


Figure 17: **Same-checkpoint coercive-mask stress test.** The left column uses Dense/BBox inference with 1,302 Sub tokens; the right column applies mask-cropped inference with 618 retained tokens. Both use the BBox-trained checkpoint and identical generation settings without latent blending. The red curve is the submitted mask boundary and is shown only for diagnosis. The top row is the decoded Main output and the bottom row is the decoded Sub output.

Figure 18 visualizes the hard-forward STE decisions for all 32 transformer layers at denoising step 25 in the mask-cropped arm. White cells use the paired Main-image positional embedding, while black cells retain the original Sub positional embedding. Mask cropping first determines which Sub tokens exist; the learned gate then selects the positional source independently for every retained token and layer. The anonymous supplementary archive includes the full 50-frame layer-by-timestep GIFs for both Dense/BBox and mask-cropped inference, together with the individual frames, aggregate exchange maps, and sanitized configuration records; Table 29 gives their exact relative paths.

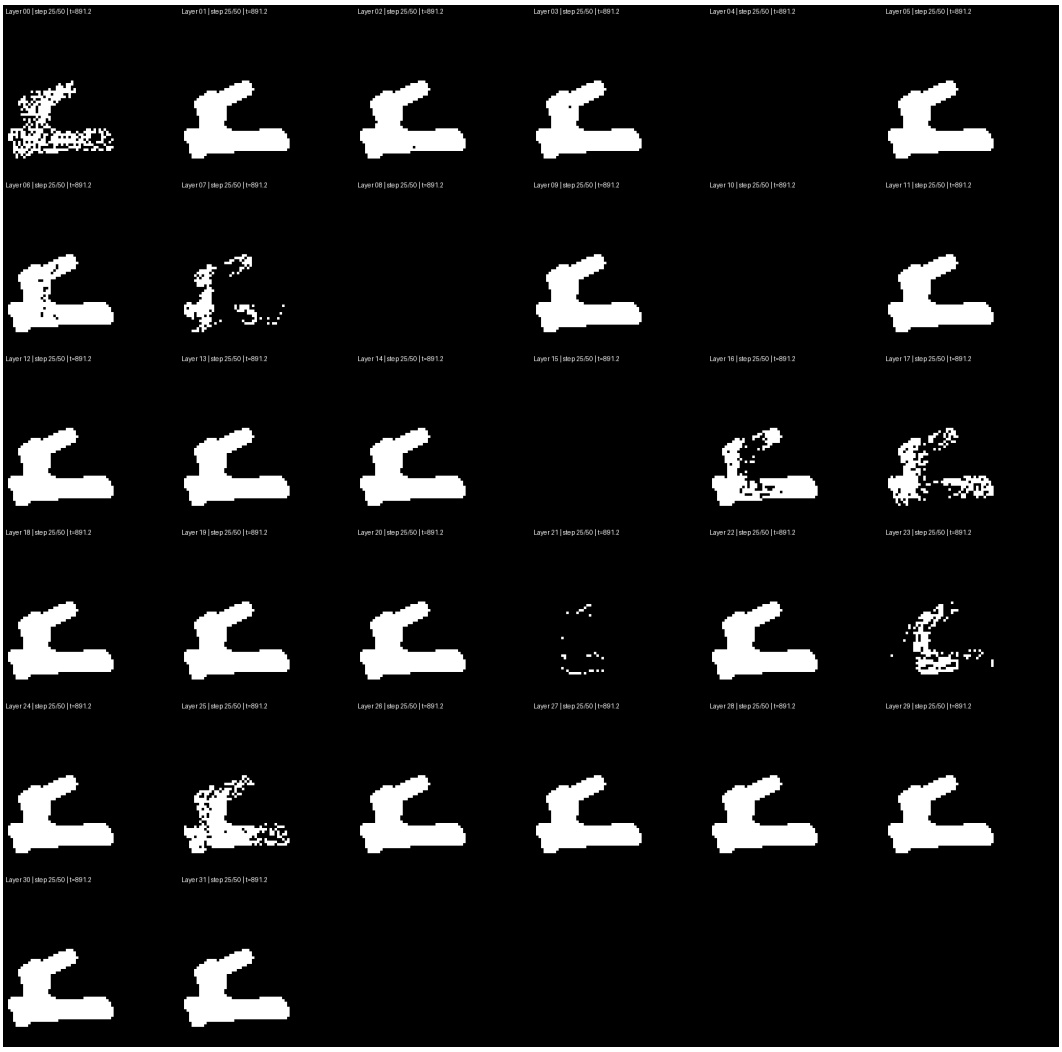


Figure 18: **STE routing on mask-cropped Sub tokens.** Hard-forward PE exchange decisions for all 32 layers at denoising step 25 of 50. White denotes exchange to the paired Main-image PE and black denotes retention of the Sub PE. The complete 50-frame temporal evolution is provided in the anonymous supplementary archive at the exact path listed in Table 29.

The ICE-Bench scores in Tables 5 and 30 use the full Dense/BBox inference path (`use_sparse_sub_branch=false` with BBox PE exchange). The optional mask-cropped mode uses the same checkpoint and is reported separately as a support-control analysis.

G.4 ICE-Bench Protocol and Score Provenance

Prompt-normalized ICE-Bench protocol. The evaluation covers Tasks 17–22 and 29–31, the nine ICE-Bench tasks that provide `SourceMask` (Pan et al., 2025). GPT-6 Luna receives the original instruction, source image, mask, and any available subject-reference image, and rewrites the instruction into a self-contained prompt. "Picture 1" denotes the source image and "Picture 2" denotes the reference when present. The rewriter does not receive the target image or `TargetCaption`. The source image remains unmodified for every task, including Task 17, and `SourceMask` is supplied as a separate spatial input. The Dense/BBox checkpoint uses CFG 4, 50 steps, seed 0, global Sub spatial IDs, and no latent blending. Official scoring uses the original `Instruction`, `TargetCaption`, task-to-metric mapping, and aggregation code.

Mask-group model comparisons. The columns in Table 5 summarize all ICE-Bench tasks with a `SourceMask`: Tasks 17–22, 29–30, and 31, following the corresponding groups in Emu3.5 Table 13 (Cui et al., 2025). Its upper block contains open-source models, while the lower block reports the proprietary Gemini references. FLUX.2 BRIDGE was not trained for Tasks 17, 18, or 31. Tasks 19–22 average 0.5941; later papers report only their combined Tasks 17–22 aggregate. The Task 29–30 value 0.6254 is the simple mean of Subject-guided Inpainting and Virtual Try On, and Task 31 contributes the Face Swap score 0.5700 directly.

The ICE-Bench original-paper block is computed from the approximate per-task scores in Table 30: Tasks 17–22 and Tasks 29–30 are simple group means, while Task 31 is copied directly. In the original-model block of Table 5, Tasks 17 and 18 are reported separately and the `19-22` column averages the four trained local-edit tasks. The later open-source and closed-model values are transcribed from Emu3.5 Table 13. FLUX.2 BRIDGE was evaluated on these nine `SourceMask` tasks, so the comparison reports the three corresponding mask groups.

Per-task results. Table 30 separates the current FLUX.2 run from external reference values. The FLUX.2 values are computed from the per-example ICE-Bench records. The baseline values are reconstructed from the rounded component metrics in the ICE-Bench supplementary material using its official scoring formula, so their displayed four-decimal task scores are approximate. Inpainting, Outpainting, and Face Swap are completed diagnostic runs because the model was not trained for these tasks; all three are excluded from the supported six-task mean.

Table 30: Per-task ICE-Bench scores for tasks with a `SourceMask`. Tasks 17, 18, and 31 are diagnostics because FLUX.2 BRIDGE was not trained for Inpainting, Outpainting, or Face Swap. ACE++, ACE, OmniGen, and UltraEdit are external reference values reconstructed from rounded metrics reported by ICE-Bench (Pan et al., 2025); they are not new local reruns.

ICE-Bench task	Ours (FLUX Subgroup)	ACE++	ACE	OmniGen	UltraEdit
17: Inpainting (diagnostic)	**0.6506**	0.6430	0.5954	0.5454	0.4520
18: Outpainting (diagnostic)	0.5801	**0.6230**	0.5528	0.5733	0.4784
19: Local Subject Addition	**0.6373**	0.6286	0.5623	0.5482	0.5501
20: Local Subject Removal	**0.5906**	0.5464	0.5733	0.5631	0.4963
21: Local Text Render	0.5564	**0.5593**	0.5514	0.4708	0.4462
22: Local Text Removal	0.5922	**0.5977**	0.5910	0.5635	0.5625
29: Subject-guided Inpainting	**0.5733**	0.5406	0.5081	0.4936	—
30: Virtual Try On	**0.6775**	0.6725	0.6362	0.6581	—
31: Face Swap (diagnostic)	**0.5700**	0.5396	0.4907	0.5324	—
Nine-task mean (incl. diagnostics)	**0.6031**	0.5945	0.5624	0.5498	—
Supported six-task mean	**0.6045**	0.5908	0.5704	0.5496	—

G.5 Per-Task ICE-Bench Component Comparisons

Tables 31–39 compare the FLUX Subgroup run with the models reported for each corresponding task in the official ICE-Bench supplementary tables (Pan et al., 2025). Task 17–22 include ACE, OmniGen, ACE++, and UltraEdit. The official Task 29–31 tables report ACE, OmniGen, and ACE++ only, so UltraEdit is not added to those comparisons. Published component metrics are transcribed at their reported three-decimal precision. Their TaskScore values are recomputed from those rounded components with the official `cal_scores.py` formula and are therefore approximate. The Ours TaskScore values use the unrounded aggregate metrics from the current run. Bold marks the best displayed result in each column; lower L1-src is better and all other metrics are higher-is-better.

Table 31: ICE-Bench component metrics for Task 17 (Inpainting); diagnostic because this task was not used for task-specific training.

Method	Aesthetic	Imaging	CLIP-cap	VLLM-QA	CLIP-src	L1-src	TaskScore
ACE	4.878	51.793	0.269	0.833	0.785	0.024	0.5954
OmniGen	4.545	59.264	0.238	0.524	0.734	0.108	0.5454
ACE++	5.064	**61.661**	0.272	0.910	0.776	**0.016**	0.6430
UltraEdit	3.817	46.284	0.250	0.180	**0.952**	0.019	0.4520
Ours (FLUX Subgroup)	**5.316**	60.457	**0.277**	**0.923**	0.785	0.019	**0.6506**

Table 32: ICE-Bench component metrics for Task 18 (Outpainting); diagnostic because this task was not used for task-specific training.

Method	Aesthetic	Imaging	CLIP-cap	VLLM-QA	CLIP-src	L1-src	TaskScore
ACE	5.514	50.403	0.287	0.376	0.891	0.017	0.5528
OmniGen	5.442	**65.758**	0.265	0.326	0.802	0.114	0.5733
ACE++	**5.644**	64.156	**0.289**	**0.531**	0.908	**0.010**	**0.6230**
UltraEdit	4.498	43.968	0.274	0.220	**0.945**	0.018	0.4784
Ours (FLUX Subgroup)	5.200	61.141	0.283	0.408	0.942	0.049	0.5801

Table 33: ICE-Bench component metrics for Task 19 (Local Subject Addition).

Method	Aesthetic	Imaging	CLIP-cap	VLLM-QA	CLIP-src	L1-src	TaskScore
ACE	4.965	51.704	0.272	0.555	0.897	0.029	0.5623
OmniGen	4.584	58.911	0.249	0.479	0.814	0.066	0.5482
ACE++	5.014	**62.083**	0.268	0.785	0.894	0.018	0.6286
UltraEdit	4.881	47.855	0.275	0.555	0.909	0.021	0.5501
Ours (FLUX Subgroup)	**5.170**	59.979	**0.279**	**0.823**	**0.919**	**0.015**	**0.6373**

Table 34: ICE-Bench component metrics for Task 20 (Local Subject Removal).

Method	Aesthetic	Imaging	CLIP-cap	VLLM-QA	CLIP-src	L1-src	TaskScore
ACE	4.996	47.011	**0.258**	**0.757**	0.852	0.024	0.5733
OmniGen	4.792	54.320	0.238	0.658	0.787	0.061	0.5631
ACE++	5.061	**61.614**	0.229	0.312	0.901	0.017	0.5464
UltraEdit	4.858	48.748	0.226	0.287	0.888	0.018	0.4963
Ours (FLUX Subgroup)	**5.158**	57.486	0.255	0.614	**0.913**	**0.014**	**0.5906**

Table 35: ICE-Bench component metrics for Task 21 (Local Text Render).

Method	Aesthetic	Imaging	CLIP-cap	VLLM-QA	CLIP-src	L1-src	TaskScore
ACE	4.275	43.159	0.276	0.791	0.860	0.016	0.5514
OmniGen	4.015	42.527	0.261	0.380	0.815	0.066	0.4708
ACE++	4.231	**43.276**	0.277	0.834	0.899	0.012	**0.5593**
UltraEdit	**4.506**	38.887	0.277	0.098	**0.946**	0.014	0.4462
Ours (FLUX Subgroup)	4.230	41.024	**0.278**	**0.865**	0.873	**0.010**	0.5564

Table 36: ICE-Bench component metrics for Task 22 (Local Text Removal).

Method	Aesthetic	Imaging	CLIP-cap	VLLM-QA	CLIP-src	L1-src	TaskScore
ACE	**4.896**	49.766	**0.273**	**0.801**	0.888	0.033	0.5910
OmniGen	4.487	56.398	0.246	0.674	0.793	0.097	0.5635
ACE++	4.694	**59.636**	0.260	0.704	0.905	0.017	**0.5977**
UltraEdit	4.665	47.294	0.264	0.714	0.910	0.023	0.5625
Ours (FLUX Subgroup)	4.864	57.414	0.273	0.638	**0.945**	**0.015**	0.5922

Table 37: ICE-Bench component metrics for Task 29 (Subject-guided Inpainting).

Method	Aesthetic	Imaging	CLIP-cap	VLLM-QA	DINO-ref	CLIP-src	L1-src	TaskScore
ACE	4.812	52.544	0.197	0.171	0.562	0.766	0.015	0.5081
OmniGen	4.459	59.995	0.186	0.093	0.555	0.642	0.149	0.4936
ACE++	4.835	**63.419**	0.186	0.257	0.563	0.753	0.040	0.5406
Ours (FLUX Subgroup)	**5.031**	60.533	**0.228**	**0.436**	**0.580**	**0.833**	**0.014**	**0.5733**

Table 38: ICE-Bench component metrics for Task 30 (Virtual Try On).

Method	Aesthetic	Imaging	CLIP-cap	VLLM-QA	DINO-ref	CLIP-src	L1-src	TaskScore
ACE	4.837	64.723	0.231	0.629	0.751	**0.889**	**0.006**	0.6362
OmniGen	4.696	73.313	0.235	0.722	0.744	0.847	0.058	0.6581
ACE++	4.577	**73.525**	0.243	0.804	**0.763**	0.882	0.029	0.6725
Ours (FLUX Subgroup)	**4.930**	70.693	**0.248**	**0.825**	0.760	0.878	0.010	**0.6775**

Table 39: ICE-Bench component metrics for Task 31 (Face Swap); diagnostic because this task was not used for task-specific training.

Method	Aesthetic	Imaging	CLIP-cap	VLLM-QA	Face-ref	CLIP-src	L1-src	TaskScore
ACE	4.983	56.985	0.232	0.400	0.250	**0.763**	0.018	0.4907
OmniGen	4.309	64.021	0.217	0.484	**0.477**	0.661	0.112	0.5324
ACE++	**5.034**	**64.963**	0.231	0.442	0.378	0.760	0.054	0.5396
Ours (FLUX Subgroup)	4.854	64.283	**0.255**	**0.716**	0.391	0.757	**0.014**	**0.5700**

ACE++ reproduction. We additionally reran the released ACE++ checkpoints on the same `SourceMask` task set. The run uses the original ICE-Bench source images, the task-appropriate released local-editing, subject, and portrait LoRAs, 50 denoising steps, guidance scale 50, repainting scale 1.0, and seed 42. We then apply the same ICE-Bench evaluator and official task-score aggregation. Table 40 compares this rerun with scores reconstructed from the rounded component metrics published in the ICE-Bench supplement (Pan et al., 2025). The rerun is close in the nine-task macro average (0.5941 versus 0.5945), while individual tasks differ; Outpainting is 0.0125 lower and Face Swap is 0.0279 higher.

Table 40: ACE++ reproduction on the nine ICE-Bench `SourceMask` tasks. Published values are approximate composite scores reconstructed from rounded component metrics. Δ is reproduction minus published.

Task	Published	Reproduced	Δ
17: Inpainting	0.6430	0.6389	-0.0041
18: Outpainting	0.6230	0.6106	-0.0125
19: Local Subject Addition	0.6286	0.6224	-0.0062
20: Local Subject Removal	0.5464	0.5398	-0.0065
21: Local Text Render	0.5593	0.5667	$+0.0074$
22: Local Text Removal	0.5977	0.5870	-0.0108
29: Subject-guided Inpainting	0.5406	0.5592	$+0.0187$
30: Virtual Try On	0.6725	0.6544	-0.0181
31: Face Swap	0.5396	0.5675	$+0.0279$
Nine-task macro	0.5945	0.5941	-0.0005

Relative to the ACE++ reference, FLUX.2 BRIDGE is higher on supported Tasks 19, 20, 29, and 30, and on diagnostic Tasks 17 and 31. The two supported-task deficits are 0.0029 on Local Text Render and 0.0055 on Local Text Removal, while the largest deficit, 0.0429, occurs on the unsupported Outpainting task. This task-level pattern explains why the Outpainting-inclusive Task 17–22 group in Table 5 is treated as a diagnostic aggregate.

G.6 Local Subject Addition Mask Intervention

We vary the mask while keeping the source image, Luna-rewritten prompt, seed, and sampling configuration fixed within each model. The four conditions are the original mask, an equal-area ellipse, a smaller mask, and a larger mask. Figures 19–22 compare BRIDGE, ACE++, and Q-Control (Qwen-Image-Edit with Blockwise ControlNet Inpaint) (DiffSynth-Studio, 2025); the red contour marks the input mask. These examples diagnose response to mask support and are reported separately from the ICE-Bench task score.

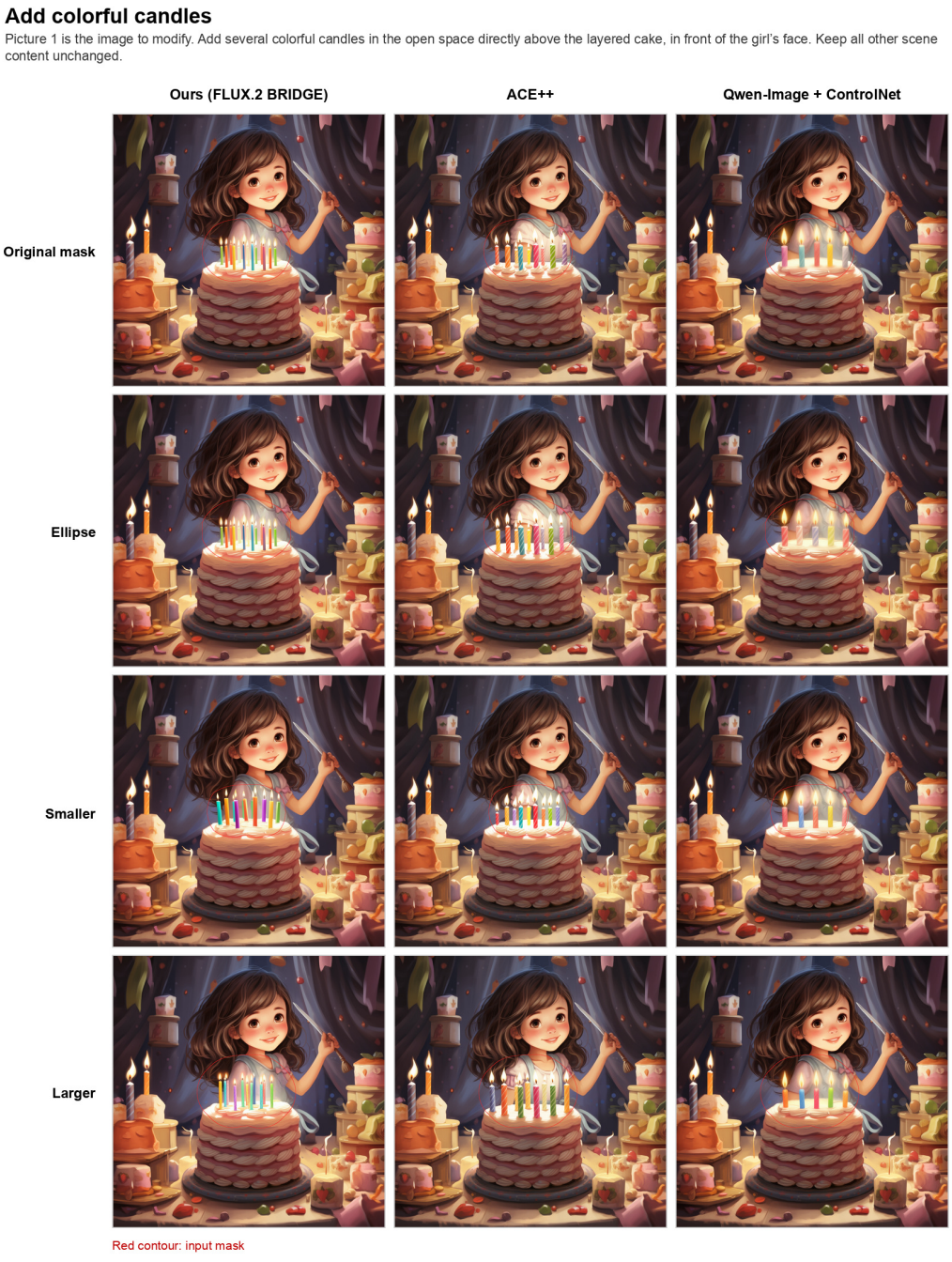


Figure 19: **Local Subject Addition mask intervention (1/4): colorful candles.** Each column uses the same prompt across the four mask conditions.

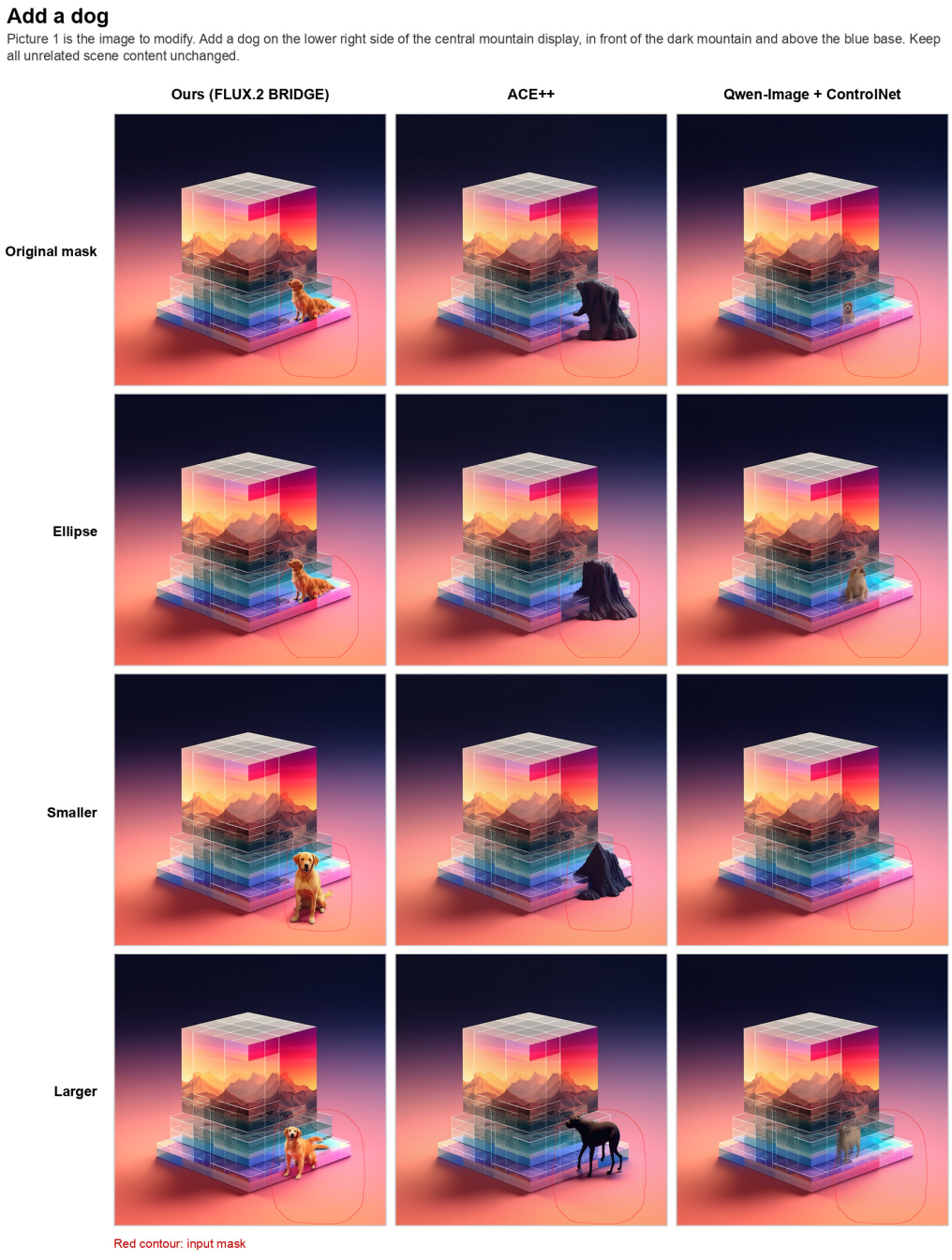


Figure 20: **Local Subject Addition mask intervention (2/4): dog.** The red contour shows the mask supplied to each model.

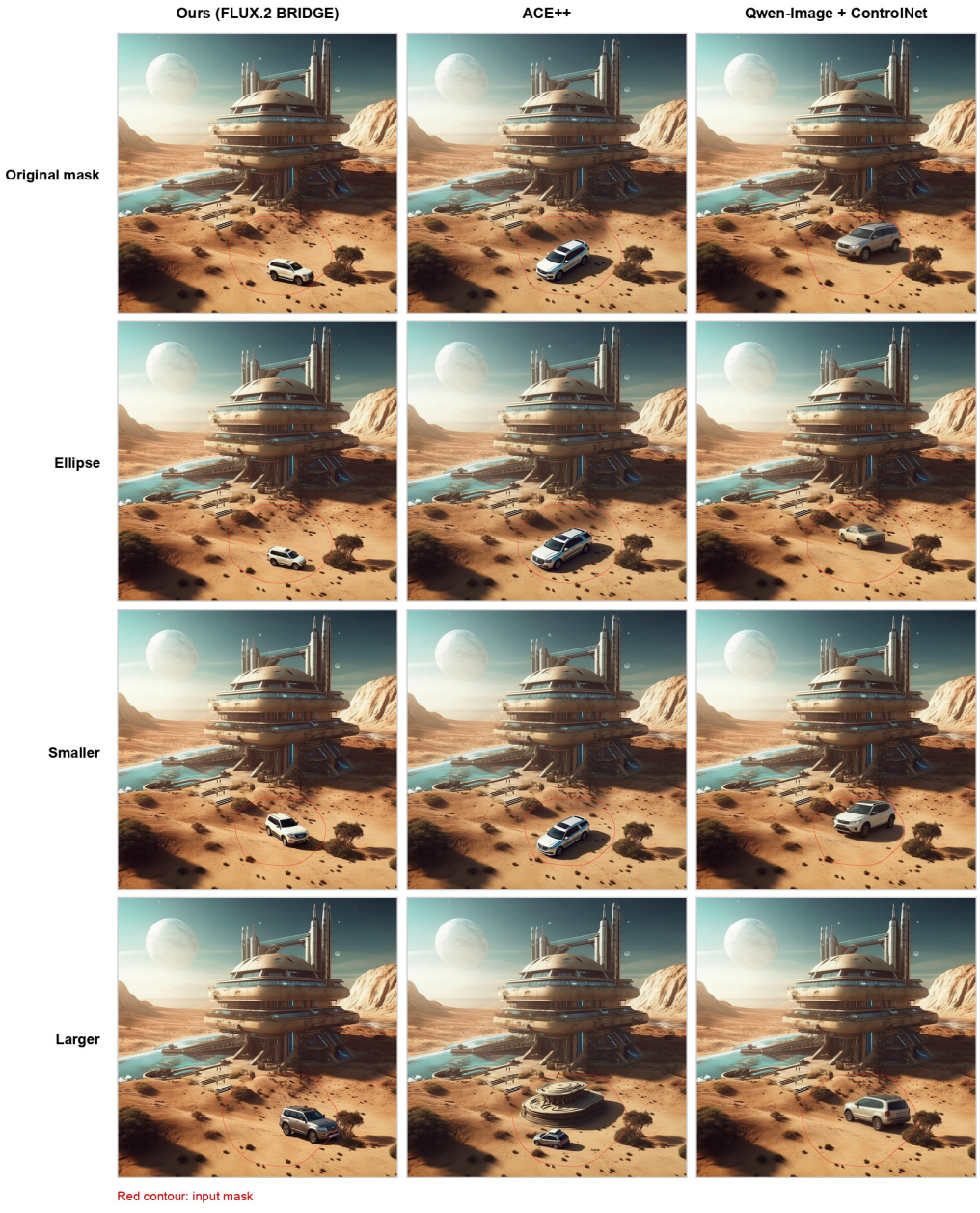


Figure 21: **Local Subject Addition mask intervention (3/4): SUV.** Mask area changes are shown separately from the equal-area ellipse.

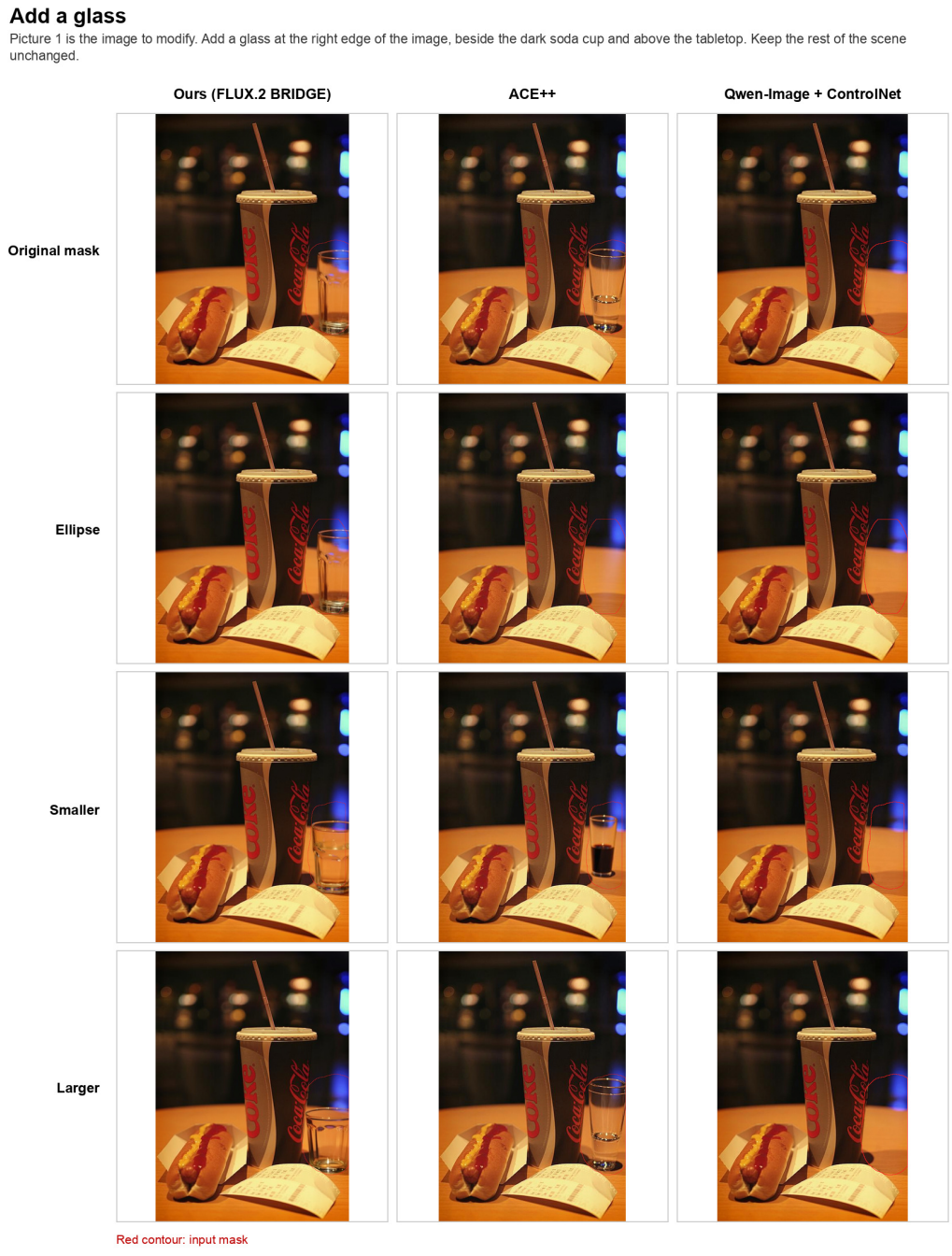


Figure 22: **Local Subject Addition mask intervention (4/4): glass.** BRIDGE consistently inserts a plausible glass across the displayed masks. ACE++ is inconsistent, producing missing or malformed glasses, while Q-Control fails to insert the requested object.

G.7 ICE-Bench Qualitative Showcase

The following figures show five selected outputs for each ICE-Bench task from the same FLUX.2 Dense/BBox run used for the quantitative table. Each row in the source showcase contains the source image, an optional subject reference, and the generated image with the 1-pixel red mask boundary.

Prompts were rewritten with GPT-6 Luna and evaluated with CFG $= 4$; the diagnostic tasks are included for completeness and are not evidence of task-specific training.

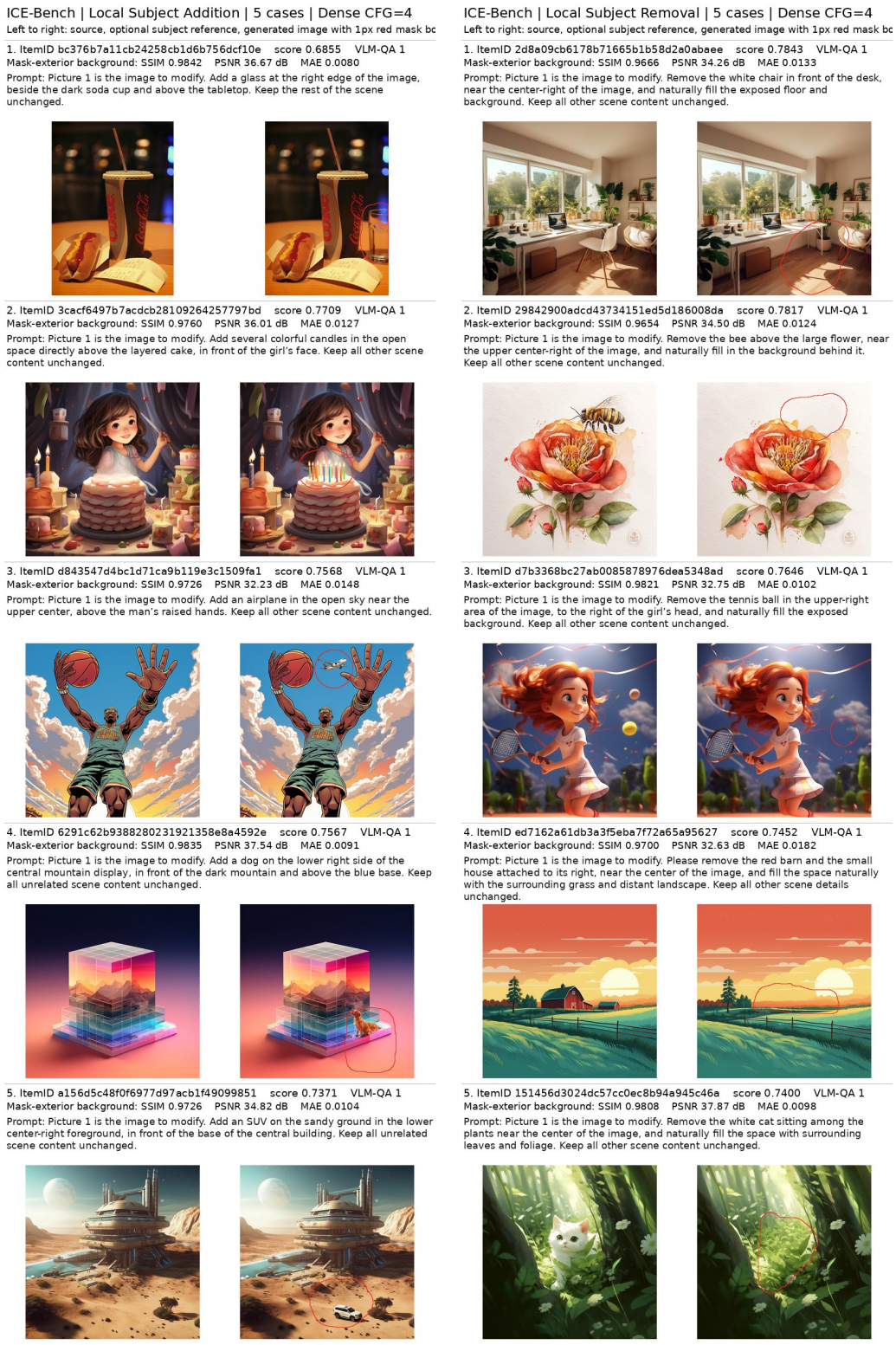


Figure 23: **ICE-Bench qualitative showcase (1/6).** Two selected five-case panels from the same Dense/BBox run. The embedded panels retain the rewritten prompt, optional reference, generated image, and red mask boundary.

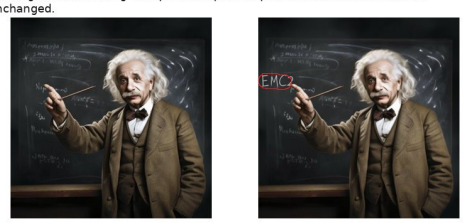
ICE-Bench | Local Text Render | 5 cases | Dense CFG=4

Left to right: source, optional subject reference, generated image with 1px red mask bc

1. ItemID 6a388b11d38ada02b8b06f79278f5a1a score 0.6751 VLM-QA 1
Mask-exterior background: SSIM 0.9717 PSNR 34.75 dB MAE 0.0115
Prompt: Picture 1 is the image to modify. Add the word “Sakura” on the train’s front, centered across the lower pink panel between the two headlights. Keep all other scene content unchanged.



2. ItemID 688434a83185831b972c94b045490a1f score 0.6610 VLM-QA 1
Mask-exterior background: SSIM 0.9857 PSNR 38.16 dB MAE 0.0084
Prompt: Picture 1 is the image to modify. Position the text “EMC2” in the upper-left area of the chalkboard, just below its top edge and to the left of the writing teacher’s pointing hand, matching the specified spot. Keep all unrelated scene content unchanged.



3. ItemID 652487f4e129a728f9731445a25d6d0c score 0.6572 VLM-QA 1
Mask-exterior background: SSIM 0.9797 PSNR 37.53 dB MAE 0.0097
Prompt: Picture 1 is the image to modify. Arrange for “12” to reside in the upper-left wall area, just below the small round fixture and left of the left lamp, matching the scene’s style. Keep all unrelated content unchanged.



4. ItemID d081e4e23e0056f2d561304de053ac13 score 0.6540 VLM-QA 1
Mask-exterior background: SSIM 0.9832 PSNR 36.91 dB MAE 0.0102
Prompt: Picture 1 is the image to modify. Overlay the text “STOP” in the center of the large white ring on the back wall. Keep all unrelated scene content unchanged.



5. ItemID fc0edb0b83eba7943dfcefce1e3b6d28 score 0.6493 VLM-QA 1
Mask-exterior background: SSIM 0.9782 PSNR 36.54 dB MAE 0.0101
Prompt: Picture 1 is the image to modify. Embed the lowercase text “shop” on the dark upper-right cabinet panel, above the shelves and to the right of the central divider. Keep all other scene content unchanged.



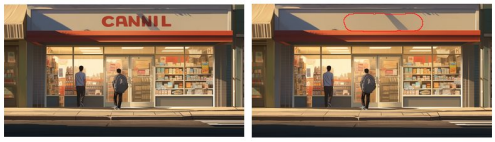
ICE-Bench | Local Text Removal | 5 cases | Dense CFG=4

Left to right: source, optional subject reference, generated image with 1px red mask bc

1. ItemID cb7bdb92b40b70a3b2bf761930b40bfb score 0.7563 VLM-QA 1
Mask-exterior background: SSIM 0.9815 PSNR 38.22 dB MAE 0.0089
Prompt: Picture 1 is the image to modify. Remove all typography from the central marquee sign, including the large lettering and the smaller text on both panels, and restore the sign’s underlying surface. Keep all unrelated scene content unchanged.



2. ItemID 245cf81d060baaea729e0ed06ee48cfd score 0.6708 VLM-QA 1
Mask-exterior background: SSIM 0.9784 PSNR 35.90 dB MAE 0.0118
Prompt: Picture 1 is the image to modify. Remove the red “CANNUL” lettering on the sign near the top center and restore the sign’s underlying surface, keeping the rest of the scene unchanged.



3. ItemID 2439a995f497c23352156871ea0155f5 score 0.7364 VLM-QA 1
Mask-exterior background: SSIM 0.9614 PSNR 35.23 dB MAE 0.0118
Prompt: Picture 1 is the image to modify. Remove the word “S A A R E” from the red sign above the left storefront opening, and restore the sign’s red surface in its place. Keep all other scene content unchanged.



4. ItemID ff2ab3061e675362606debe5b7377c12 score 0.7328 VLM-QA 1
Mask-exterior background: SSIM 0.9662 PSNR 33.58 dB MAE 0.0128
Prompt: Picture 1 is the image to modify. Remove the small stylized text on the light gray sign near the upper left of the bar, above the bottle shelves, and restore the sign’s plain surface. Keep the rest of the scene unchanged.



5. ItemID 6d4dd693f0ac9844eb16a70576bc08bf score 0.7252 VLM-QA 1
Mask-exterior background: SSIM 0.9815 PSNR 35.34 dB MAE 0.0109
Prompt: Picture 1 is the image to modify. Remove the “ALL AM PS” text from the red sign at the bottom center, and restore the sign’s plain red surface. Keep all other scene content unchanged.



Figure 24: **ICE-Bench qualitative showcase (2/6).** Two selected five-case panels from the same Dense/BBox run.

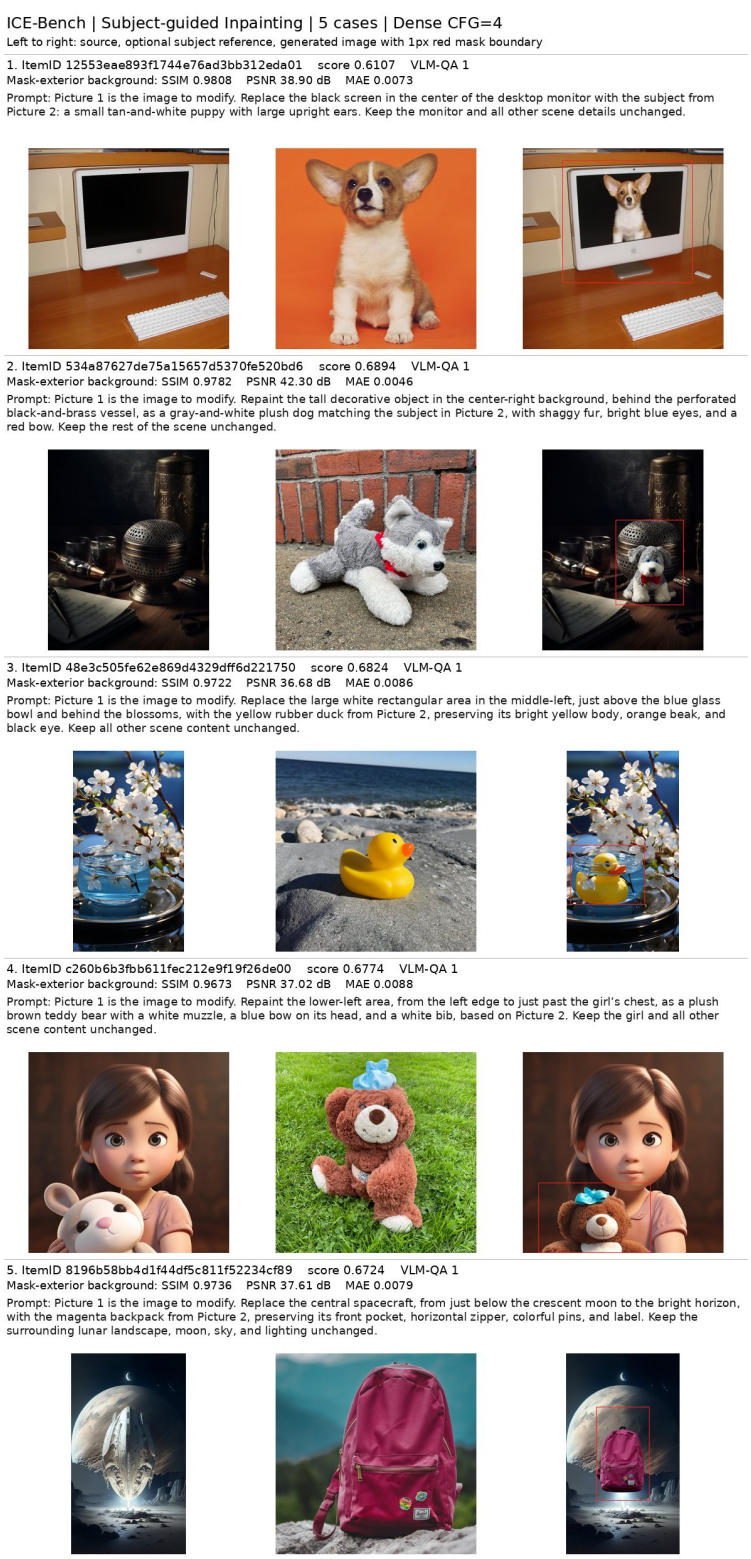


Figure 25: **ICE-Bench qualitative showcase (3/6).** One selected five-case panel from the same Dense/BBox run.

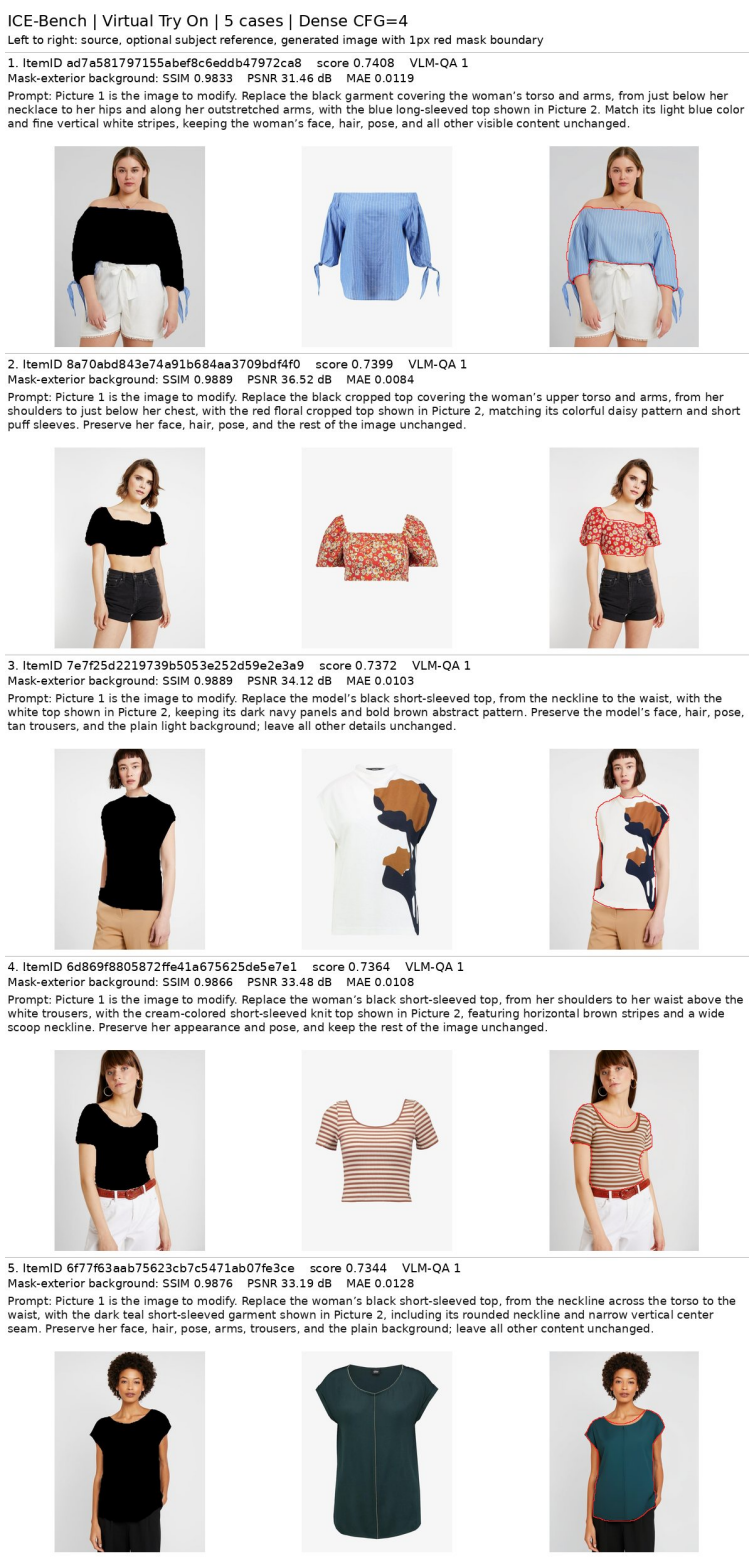


Figure 26: **ICE-Bench qualitative showcase (4/6).** One selected five-case panel from the same Dense/BBox run.

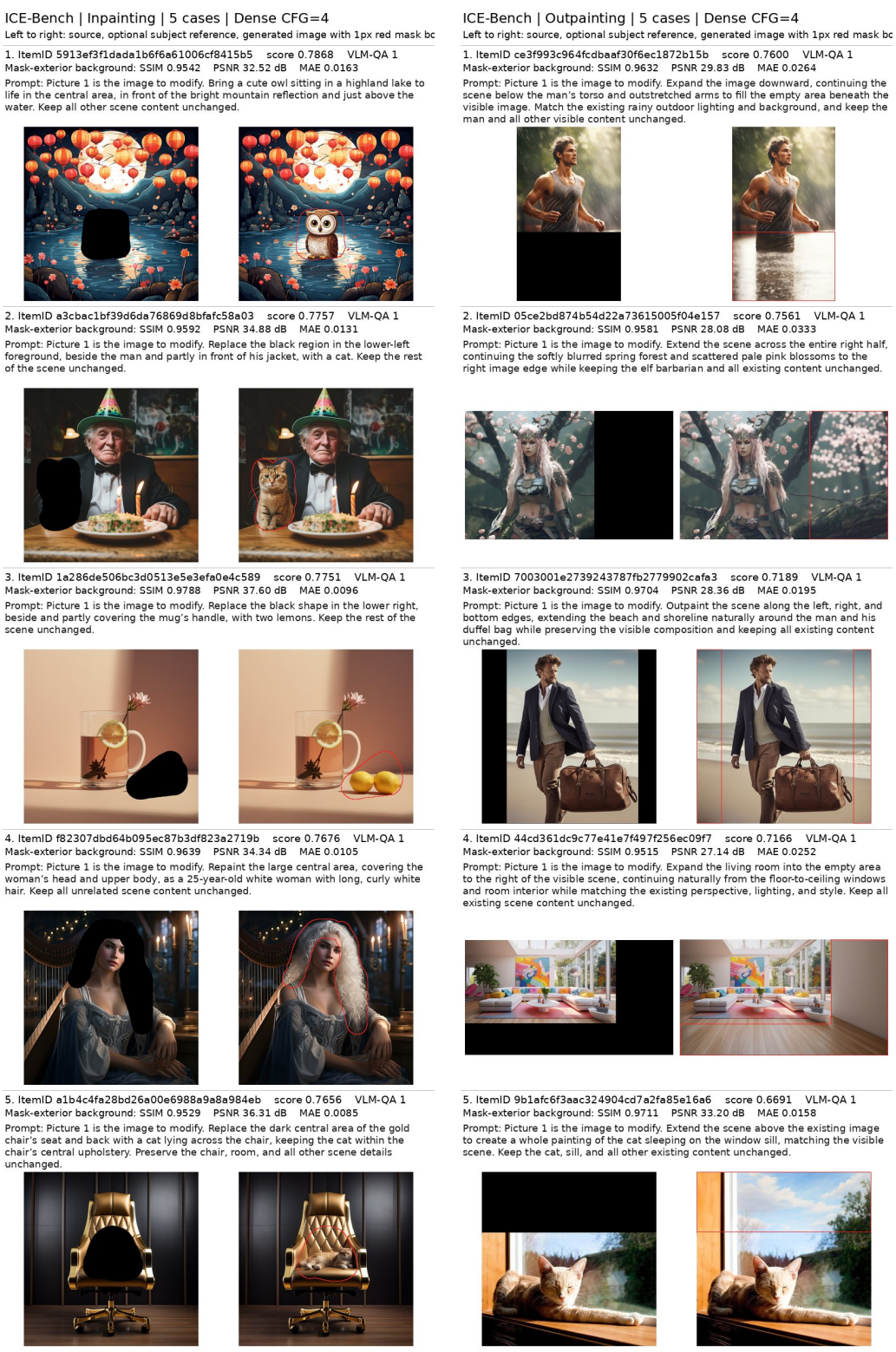


Figure 27: **ICE-Bench qualitative showcase (5/6).** Two diagnostic five-case panels from the same Dense/BBox run. These tasks were not used for task-specific training.

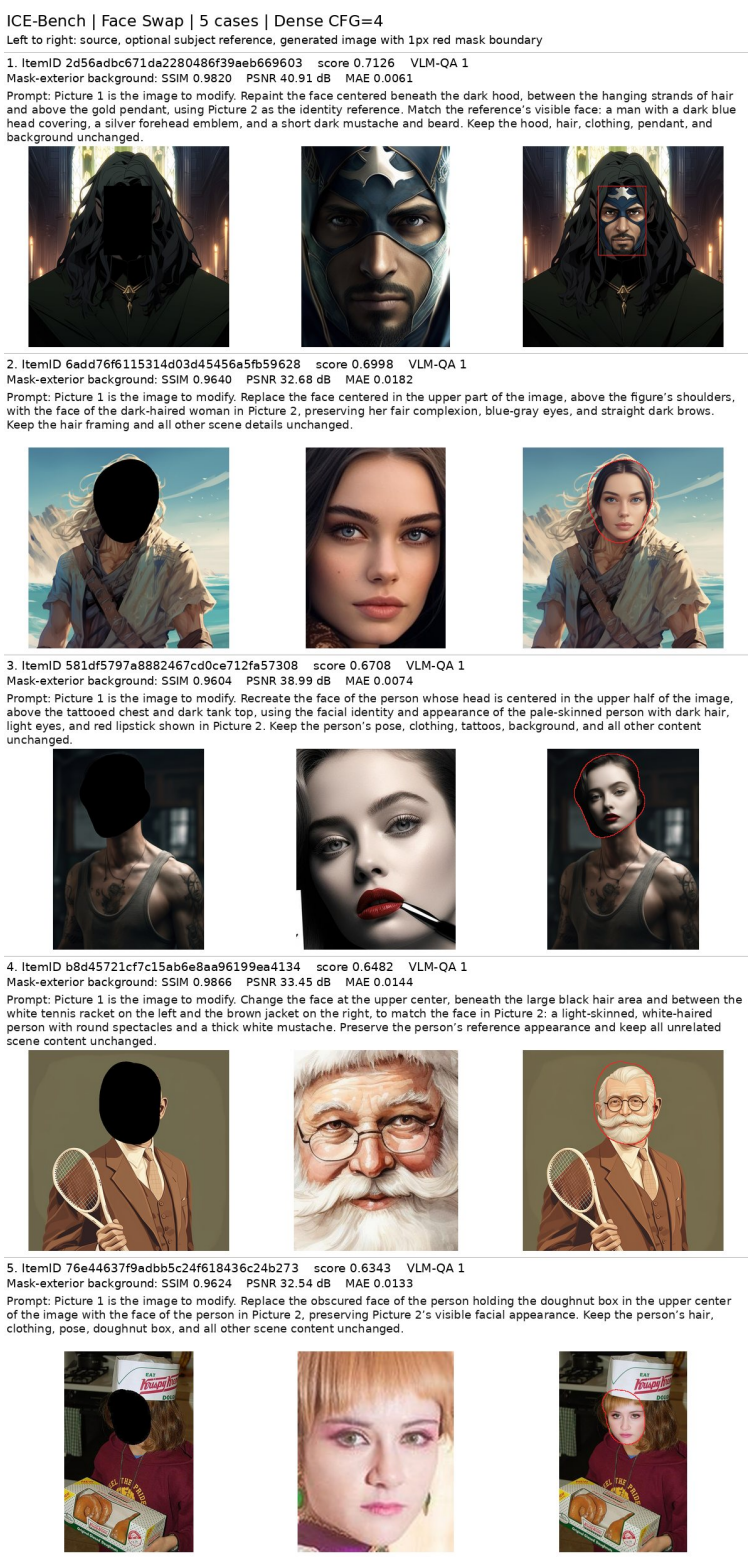


Figure 28: **ICE-Bench qualitative showcase (6/6).** One diagnostic five-case panel from the same Dense/BBox run. Face Swap was not used for task-specific training.