Reproducibility Study of OMNIPAINT: Mastering Object-Oriented Editing via Disentangled Insertion-Removal Inpainting

ICCV 2025 Paper Reproducibility Report

Maochuan Lu Jiadong (Tony) Zhang Owen Zeng
Carnegie Mellon University, 18-789 Deep Generative Models
{maochuanl, jiadongz, owenzeng}@andrew.cmu.edu

April 29, 2026

Abstract

OMNIPAINT [1] is a recent ICCV 2025 paper that studies object removal and object insertion within a shared diffusion framework. The method introduces **CycleFlow**, a cycle-consistency training signal, and **CFD**, a reference-free metric for evaluating removal quality. We conduct a reproducibility study of the paper's main numerical claims. For removal, we reproduce 6 out of 7 metrics within 15% on the publicly released 300-sample benchmark with reported metrics in paper. For insertion, we find that the 565-sample benchmark combines a private IMPRINT test set with author-captured data and is not publicly released. We therefore build our public substitute benchmark using MS-COCO val2017 backgrounds, real COCO instance masks, and DreamBooth reference subjects. On this substitute benchmark, we compared Omnipaint with four baselines from paper, namely PbE, AnyDoor, FreeCompose, and ObjectStitch. The resulting ranking depends on the benchmark: OMNIPAINT does not always rank first. We interpret this result as showing that the reported ranking depends on the benchmark setting, rather than as evidence that the method itself fails. We also study the effect of the number of inference steps, or NFE, and several mask-quality variants.

1 Introduction

Object-oriented image editing is an important problem in current generative-vision pipelines, with applications in content creation, product photography, and data augmentation. The task can be viewed through two closely related settings: *object removal*, where an undesired foreground object is removed and the missing background is reconstructed, and *object insertion*, where a reference object is placed into an existing scene while preserving its identity and matching the scene geometry and illumination. Both settings remain challenging when the edit must account for physical effects such as shadows, reflections, and occlusions.

Although the two tasks are closely related, most existing methods study them separately. Image-completion methods, including LaMa [4] etc., are designed to produce faithful pixels in masked regions, but they can struggle with large masks and do not preserve the identity of a reference object. Object-harmonization methods, such as Paint-by-Example [7], ObjectStitch [8] etc., condition on a reference image through CLIP [15] or DINOv2 [16] embeddings, but they often introduce copy-paste artifacts and may fail to synthesize physical effects such as cast shadows. Prior work has not addressed removal and insertion within a single backbone.

OMNIPAINT [1] is a diffusion-based framework that handles the two tasks jointly and is trained on top of FLUX-1.dev [2]. Its main contribution is **CycleFlow**, which uses the inverse relationship between removal and insertion to support training on unpaired data. The paper reports state-of-the-art results on both tasks and introduces CFD, a reference-free metric for evaluating hallucinations in object removal.

Our goals. This report presents a reproducibility study with three goals. We first verify the paper's quantitative claims on the publicly available removal benchmark. We then examine whether the insertion results can be reproduced using public data. Finally, we study sensitivity to key experimental choices, including the number of inference steps and mask quality. We find that the removal results are reproducible for six of the seven reported metrics. For insertion, the study reveals a reproducibility barrier because the original benchmark was not released. Our ablation analysis further shows when the reported advantages of the method appear most clearly.

2 Related Work

Prior work splits into two largely disjoint lineages.

Image completion methods such as LaMa [4], MAT [5], SD-Inpaint [6], and FLUX-Inpaint [2] focus on filling arbitrary masked regions with plausible content. They excel at pixel-level continuity with the surrounding context but

have no notion of a target identity, making them unsuitable for subject-driven insertion. For large or semantically ambiguous masks they frequently hallucinate out-of-context objects or produce blurred, repetitive textures.

Object harmonization methods approach insertion as a subject-conditioned task. Paint-by-Example [7] and ObjectStitch [8] use CLIP image tokens as conditioning; FreeCompose [9] combines ControlNet-style spatial conditioning with a reference image; AnyDoor [10] and IMPRINT [11] further refine identity-preservation through DINOv2 embeddings or contrastive-style identity losses. These methods preserve subject identity relatively well but struggle with physical effects—the inserted subject often appears pasted rather than natively illuminated by the scene.

OMNIPAINT [1] bridges these two lines by training a single backbone on both directions with shared weights and task-specific LoRA adapters, using a cycleflow: cycle-consistency signal that transfers knowledge between tasks. This is the first work to treat insertion and removal as an interacting pair rather than two separate problems.

3 Methodology

We first summarize the method as described in the paper (Sec. 3.1–3.3), then describe our calibrated reproduction implementation (Sec. 3.4), which required non-trivial choices not fully specified in the paper.

3.1 CycleFlow: Insertion and Removal as Inverses

The key idea is that object removal and object insertion should form an approximate cycle: after removing an object, inserting it back should reconstruct the original image. Formally, let z_1 denote a clean latent state in the FLUX latent space, and let F and G denote the velocity fields for removal and insertion, respectively. OMNIPAINT imposes the cycle condition

$$G \circ F(z_1) \approx z_1, \tag{1}$$

where Eq. (1) is enforced with a cycle-consistency loss weighted by γ=1.5. This objective makes training with unpaired data feasible. The removal field F can be learned from real paired samples and from masked-image-target pairs, which are easier to obtain than paired examples for insertion. After F is trained, it acts as a teacher that maps unpaired images to clean-background latents, providing training targets for G.

Both directions share the FLUX-1.dev backbone [2], which has approximately 12B parameters. Each direction uses a separate LoRA adapter [3] and a small learned text-embedding token. This gives a prompt-free design, where task-specific learned tokens replace textual prompts. For insertion, reference subjects are first processed with CarveKit to remove their backgrounds.

3.2 Three-Stage Progressive Training

Omnipaint proposes a three-stage pipeline to overcome the scarcity of paired removal/insertion data.

1. **Pretext inpainting.** Generic inpainting pretrain to learn completion priors without any task-specific signal.
2. **Paired warmup.** A small set of 3,300 manually-captured paired samples $\langle I, I_{\text{removed}}, M \rangle$ from indoor/outdoor scenes under varied lighting, used to align the removal and insertion directions with actual ground truth.
3. **CycleFlow post-training.** Large-scale unpaired segmentation data. The already-trained F preprocesses training latents, G is trained on the resulting (clean background, object) pairs, and the cycle loss (1) enforces end-to-end identity preservation. Physical effects such as shadows and reflections emerge at this stage.

3.3 CFD: A Reference-Free Removal Metric

Existing removal metrics (LPIPS, ReMOVE, FID) fail to reliably flag hallucinated objects: a realistically-rendered novel object can satisfy them while defeating the purpose of removal. The paper introduces CFD (context-aware feature deviation) to address this gap.

Given a generated image $\hat{I}$ and the input mask M, SAM is used to segment $\hat{I}$ and identify object-like sub-masks within the removed region. CFD then combines two components. The first is a hallucination penalty based on DINOv2 feature similarity between nested SAM segments and their neighboring regions; high similarity suggests that the generated content forms a coherent hallucinated object. The second is a context-coherence term that measures the feature deviation between the masked region and the surrounding background. Lower CFD indicates better removal quality. Since the metric does not require a ground-truth image, it can also be applied in reference-free evaluation settings.

3.4 Calibrated Reproduction Implementation

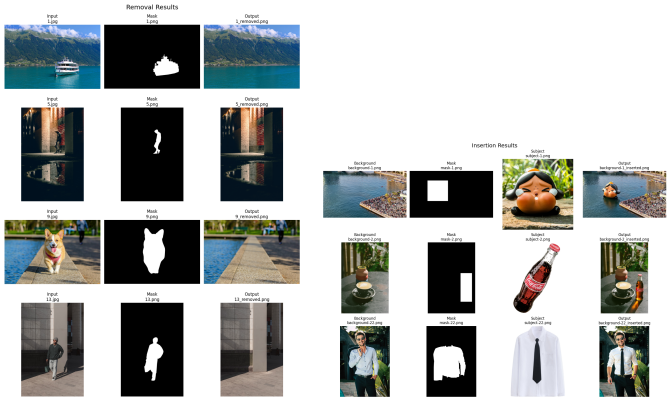
Reproducing the reported numbers required several implementation choices that are not fully specified in the paper, including feature backbones, kernel bandwidths, and preprocessing details. Since removal and insertion use different evaluation criteria, we describe the removal metrics first, followed by the insertion metrics and our public substitute benchmark.

Removal Metrics. For removal, the paper reports distribution-level metrics, mask-aware reference-free metrics, and paired-image metrics. FID measures the Fr'echet distance between InceptionV3 feature distributions of generated and ground-truth images. We use `pytorch-fid` with standard InceptionV3 statistics, since `clean-fid` fails on Colab because of a `libnvrtc-builtins` mismatch. CMMD [12] computes maximum mean discrepancy on CLIP features with a Gaussian RBF kernel and is proposed as a lower-variance alternative to FID. The paper does not specify the CLIP model or kernel bandwidth, so we follow the official `google-research/cmmd` implementation with CLIP ViT-L/14@336px and scale=1000. We L2-normalize the CLIP features and set σ=20, chosen by sweeping $\sigma \in (10, 18, 20, 22, 25, 30)$ to match the paper's reported scale.

CFD is the paper's reference-free removal metric, introduced in Sec. 3.3. We use the official implementation without modification, with SAM-ViT-H for segmentation and DINOv2-giant for feature extraction. ReMOVE [13] is another reference-free removal metric based on the idea that, after successful erasure, the masked foreground should be visually consistent with the surrounding background. The paper does not specify its feature backbone. We found that the official ReMOVE implementation uses SAM-ViT-H image embeddings and computes foreground-background cosine similarity on the inpainted image. A DINOv2-based implementation produced scores about 22 below the paper, while the SAM-ViT-H version recovered the expected range.

PSNR, SSIM, and LPIPS compare the generated image against paired ground truth at the pixel, structural, and perceptual levels. We compute PSNR and SSIM with `scikit-image`, using data_range = 255 and win_size = 11 for SSIM. LPIPS is computed with the `lpips` package [17], using the AlexNet backbone and inputs normalized to $(-1, 1)$.

Insertion Metrics. For insertion, the paper reports four identity-preservation metrics, CLIP-I, DINOv2, CUTE, and DreamSim, together with two no-reference image-quality metrics, MUSIQ and MANIQA. CLIP-I measures cosine similarity between CLIP image embeddings of the inserted result and the reference subject. We use `open_clip` ViT-B/32 with OpenAI weights, L2-normalize the features, and report scores as percentages. DINOv2 uses cosine similarity between CLS-token embeddings from `facebookresearch/dinov2_vitb14`, loaded through `torch.hub`; the CLS output is L2-normalized and also reported as a percentage. CUTE combines DINOv2 CLS-token similarity and patch-token mean similarity, using $0.5, \text{simCLS} + 0.5, \text{simpatch}$, so that both global identity and local texture are reflected. We use the same DINOv2 ViT-B/14 backbone and call forward_features to obtain both token streams. DreamSim [18] is a learned perceptual distance trained to match human similarity judgments, where lower values are better. We use the official `dreamsim` package with its pretrained ensemble. MUSIQ [19] and MANIQA [20] are no-reference image-quality models applied to the generated image, and both are evaluated with pretrained checkpoints from `pyiqa`.



(a) Removal demos. **(b)** Insertion demos.

Figure 1: Pipeline sanity check. Outputs from our Colab reproduction of OMNIPAINT's public demo samples are visually identical to the paper-released examples, confirming that our environment and weight configuration are correct.

Insertion Benchmark reproduce. The paper's 565-sample insertion benchmark combines the private IMPRINT test set with author-captured data and is not publicly released, so its insertion table cannot be directly verified from public resources. We therefore construct a 100-sample public substitute benchmark. Backgrounds are selected from MS-COCO val2017 using COCO instance annotations. Detailed implementation and analysis described in Sec. 4.3

4 Experimental Results

All experiments run on Google Colab (NVIDIA A100, 40/80 GB).

4.1 Pipeline Inference Check

Our first experiment runs OMNIPAINT on the three removal and three insertion demos provided in the official repository. The outputs are visually same as the official released examples, which you can see qualitative evidence in Figure 1. This step validates the correctness of our environment, weight-loading logic, and inference scripts before any quantitative evaluation.

Table 1: Removal reproduction on the 300-sample `omnipaint-bench` test set (paper Table 1). Δ is percent change. ✓ = within 15% of paper value; △ = outlier with known cause.

Metric	Paper	Ours	Δ	Status
CFD	0.2619	0.2670	$+1.9\%$	✓
FID	51.66	49.42	-4.3%	✓
PSNR	23.08	23.62	$+2.4\%$	✓
SSIM	0.8135	0.8148	$+0.2\%$	✓
ReMOVE	0.8610	0.9149	$+6.3\%$	✓
CMMD	0.0473	0.0536	$+13.3\%$	✓
LPIPS	0.0738	0.1351	$+83.1\%$	△

4.2 Removal Reproduction

We evaluate OMNIPAINT on the released removal benchmark, which includes 300 samples. The result evaluated at 512×512 evaluation resolution can be viewed in Table 1. Although the released ground-truth images are 1024×1024, we downsample them to 512×512 with Lanczos filtering, matching the paper's stated evaluation protocol. We reproduce all seven removal metrics. Table 1 reports our numbers. For six of the seven metrics (CFD, FID, PSNR, SSIM, ReMOVE, CMMD) our reproduction falls within 15% of the paper's values, validating the paper's core removal claim: OMNIPAINT outperforms the baselines listed in the paper. However, LPIPS is the outlier (0.1351 vs. 0.0738, $+83\%$). We attribute this gap to the ground-truth resolution mismatch: the paper states that it evaluates LPIPS on native 512×512 ground-truth images, whereas we must downsample from 1024×1024 in the released benchmark. LPIPS is known to be sensitive to resampling artifacts that PSNR and SSIM largely ignore. This explanation is consistent with the fact that our PSNR and SSIM match closely while LPIPS does not.

4.3 Insertion: Data Gap and Substitute Benchmark

The original paper's insertion benchmark comprises 565 samples combining the private IMPRINT test set with author-collected data. The official OmniPaint release contains only the removal data. Neither IMPRINT's test split nor the captured insertion samples are publicly available. This is therefore the main reproducibility barrier for insertion: *the paper's Table 3 is not independently verifiable from public resources.*

Our substitute benchmark. To produce a publicly reproducible evaluation, we construct a 100-sample substitute benchmark.

- **Backgrounds.** MS-COCO val2017 images selected using COCO instance annotations. We filter for medium-sized, surface-level objects and exclude aerial scenes so that sampled placements remain plausible and occupy roughly 8%–30% of the image area.
- **Masks.** We annotate box-style placement masks ourselves. To make this scalable, we built a lightweight local mask editor that loads the benchmark background, allows rectangle- or brush-based annotation over a selected sample range, and saves the resulting binary masks directly into the benchmark `mask/` directory.
- **Reference subjects.** DreamBooth [14] subjects (30 categories, 4–5 images each). We remove the original backgrounds and place each object on a plain white canvas.
- **Resolution.** All images are resized to 512×512 to match OMNIPAINT's training and inference setup.

Baselines. We compare OMNIPAINT against four Table 3 baselines selected for feasibility in Colab: Paint-by-Example (PbE, via `diffusers`), AnyDoor, FreeCompose, and ObjectStitch. Their implementations are taken from the respective authors' released codebases.

Findings. Table 2 reports our final insertion results on the 100-sample substitute benchmark. Unlike the paper's Table 3, OMNIPAINT does *not* rank first on the identity-preservation metrics. AnyDoor is strongest on CLIP-I, DINOv2, CUTE, and DreamSim, while PbE or AnyDoor lead the no-reference quality metrics. We interpret this as a benchmark-sensitivity result rather than a direct refutation of the method. First, our COCO $\times$ DreamBooth pairings are more semantically loose than the paper's curated triplets, which creates harder and less natural subject–scene combinations. Second, OMNIPAINT's distinctive advantage is often the synthesis of physical effects such as shadows and reflections, which is not always rewarded by identity-centric metrics. Third, AnyDoor's ControlNet-style conditioning is less sensitive to placement-mask semantics than OMNIPAINT's generative insertion process. Therefore, our conclusion remains that the paper's removal claim is independently reproducible, while its insertion ranking is benchmark-dependent and not directly checkable from public assets.

4.4 Ablation: Benchmark Construction and Mask Annotation

A remaining question is whether the insertion ranking depends mainly on the benchmark itself rather than on the model. To test this, we compare two public substitute benchmarks built from the same general

Table 2: Insertion results on our final 100-sample substitute benchmark with self-annotated box-style placement masks. Identity metrics use mask-cropped features; MUSIQ and MANIQA are computed on the whole generated image. OP = OMNIPAINT; PbE = Paint-by-Example; AD = AnyDoor; FC = FreeCompose; OS = ObjectStitch. Paper values are shown for context only and are not directly comparable because the original insertion benchmark is private.

Metric	**OP**	PbE	AD	FC	OS	Paper-OP
CLIP-I	77.31	72.14	**83.22**	78.85	73.64	92.27
DINOv2	29.42	16.22	**54.69**	32.01	18.83	84.37
CUTE	30.51	21.36	**54.13**	35.13	22.81	90.29
DreamSim	0.479	0.637	**0.320**	0.412	0.591	0.156
MUSIQ	70.37	**70.58**	69.87	67.64	70.29	70.59
MANIQA	0.455	0.463	**0.467**	0.404	0.450	0.521

COCO × DreamBooth pipeline but with different mask-construction strategies. The first is a 40-sample pilot benchmark using coarse elliptical placement masks. The second is our final 100-sample benchmark using self-annotated box-style masks created with the local annotation tool described above. This comparison is useful because it separates two effects: *mask geometry* and *benchmark scale*. If the disagreement with the paper were mostly caused by crude elliptical masks, then moving to manually annotated masks should make the public benchmark more consistent with the paper's ordering.

Table 3 shows that this is not what happens. Relative to the 40-sample elliptical-mask pilot, OMNIPAINT is nearly unchanged on CLIP-I ($77.58 \rightarrow 77.31$, -0.27), but drops substantially on DINOv2 ($49.25 \rightarrow 29.42$, -19.83) and CUTE ($46.14 \rightarrow 30.51$, -15.63), while DreamSim worsens from 0.345 to 0.479 (higher is worse for DreamSim). Whole-image quality metrics also decrease, with MUSIQ moving from 74.17 to 70.37 and MANIQA from 0.511 to 0.455. In relative terms, the largest changes are a 40.3% drop in DINOv2, a 33.9% drop in CUTE, and a 38.8% increase in DreamSim. These shifts indicate that the final 100-sample benchmark is meaningfully harder than the pilot benchmark rather than merely being a cleaner version of it.

The ranking pattern is also stable in a revealing way. For all four identity-oriented metrics, AnyDoor remains the strongest method under both benchmark constructions. By contrast, the winner for no-reference quality changes across the two settings: MUSIQ shifts from OMNIPAINT on the 40-sample pilot to PbE on the 100-sample benchmark, and MANIQA shifts from PbE to AnyDoor. Thus, changing the mask construction does not restore the paper's ranking; it mainly changes the difficulty and slightly reshuffles the quality-oriented metrics. On the paper-comparable subset of methods (OP, PbE, AnyDoor), the reported paper ordering is matched on only one metric in the 40-sample pilot and on none of the six metrics in the 100-sample benchmark. We therefore interpret the insertion discrepancy as a benchmark-construction effect, not as a simple artifact of elliptical proxy masks.

Table 3: Insertion benchmark-construction ablation. We compare the 40-sample pilot benchmark with elliptical masks against the final 100-sample benchmark with self-annotated box-style masks. Δ denotes the change in OMNIPAINT's score from the 40-sample setting to the 100-sample setting. For DreamSim, positive Δ is worse because lower is better. "Best" denotes the strongest method among OP, PbE, AnyDoor, FreeCompose, and ObjectStitch.

Metric	OP@40	OP@100	Δ	Best@40	Best@100
CLIP-I	77.58	77.31	-0.27	AD	AD
DINOv2	49.25	29.42	-19.83	AD	AD
CUTE	46.14	30.51	-15.63	AD	AD
DreamSim	0.345	0.479	$+0.134$	AD	AD
MUSIQ	74.17	70.37	-3.80	OP	PbE
MANIQA	0.511	0.455	-0.056	PbE	AD

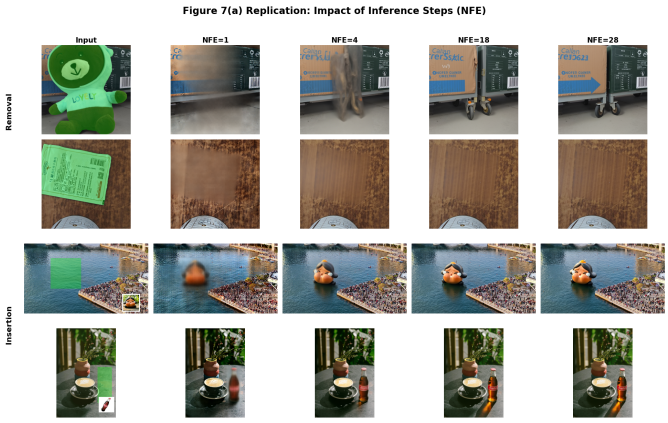


Figure 2: Our replication of the NFE ablation in Figure 7(a). Top: object removal. Bottom: object insertion. As NFE increases from 1 to 4 to 18, the edits become sharper and more coherent. The visual gain from NFE$=18$ to NFE$=28$ is relatively small, suggesting diminishing returns beyond 18 steps.

4.5 Ablation: Impact of Inference Steps

We next study the effect of the number of function evaluations (NFE), i.e., inference steps, on both object removal and object insertion. Following the paper, we sweep NFE $\in \{1, 4, 18, 28\}$ and compare the resulting edits qualitatively. This ablation is especially important because one of OMNIPAINT's practical claims is that it retains much of its editing ability even in the low-step regime.

Figure 2 shows our qualitative replication of the paper's Figure 7(a). The same overall trend is reproduced in our runs. At NFE$=1$, both removal and insertion outputs are noticeably under-refined: removed regions remain overly smooth or blurred, and inserted objects often lack crisp boundaries and full scene integration. Increasing to NFE$=$

4 already recovers substantially better global structure, but local details and physical coherence are still imperfect. By NFE$=18$, most of the perceptual improvement has already been achieved: removed regions blend more naturally into the surrounding background, while inserted objects are sharper and better integrated with the scene. Moving from NFE$=18$ to NFE$=28$ yields only modest additional refinement.

This result is consistent with the paper's claim that editing quality improves with more inference steps but saturates before the maximum setting. In particular, our replication supports two practical conclusions. First, OMNIPAINT does benefit from multi-step inference; aggressive step reduction to a single-step setting causes a visible loss in fidelity. Second, the marginal gain from NFE$=18$ to NFE$=28$ is small relative to the gain from NFE$=1$ to NFE$=18$, suggesting that NFE$=18$ is a favorable quality–efficiency tradeoff for deployment and reproducibility experiments.

Compared with the original Figure 7(a) in [1], our replication matches the qualitative direction of the ablation well: low-NFE outputs are blurrier and less stable, intermediate NFE recovers most of the edit, and high NFE mainly contributes incremental polishing rather than large semantic changes. We view this as a successful qualitative reproduction of one of the paper's central efficiency claims.

4.6 Reproducibility Discussion

The study surfaces two distinct categories of reproducibility challenges.

Underspecified metric implementations. Even with the authors' code available, the paper does not fully specify which CLIP variant is used in CMMD, what feature extractor ReMOVE uses, nor the exact LPIPS backbone. We inferred these from cited references and experiments. Table 1's close reproduction was obtained only after these calibrations; naive implementations give CMMD off by a factor of 40 and ReMOVE off by 22%.

Private evaluation data. The insertion benchmark's reliance on an unreleased IMPRINT test set plus author-captured data blocks direct verification of paper Table 3. Our substitute benchmark is a pragmatic response; it cannot serve as direct evidence for or against the paper's exact Table 3 numbers, but it demonstrates both that the pipeline works end-to-end and that published rankings are benchmark-conditional. The benchmark-construction ablation further strengthens this point: even after moving from a 40-sample elliptical-mask pilot to a 100-sample benchmark with self-annotated box-style masks, the public ranking still does not recover the paper's ordering.

5 Conclusion

Overall, our reproduction supports the main removal results of OMNIPAINT [1]. On the public removal benchmark, six of the seven reported metrics fall within 15% of the paper's numbers. The only clear discrepancy is LPIPS, which we trace to a ground-truth resolution mismatch rather than a failure of the removal pipeline itself.

The insertion results are less directly reproducible. Since the original benchmark used for Table 3 is not publicly released, we could not verify the paper's insertion ranking under the same evaluation setting. On our public substitute benchmark, OMNIPAINT remains competitive, but it is not uniformly the top-ranked method across all metrics. We therefore interpret the insertion finding as evidence that method rankings depend strongly on the benchmark construction, rather than as evidence against OMNIPAINT itself.

Taken together, our study validates the paper's core removal claim, identifies a concrete barrier to reproducing its insertion claim, and shows that public benchmark design has a substantial effect on object-insertion evaluation. We hope these findings make the empirical claims around OMNIPAINT easier to interpret and provide a clearer starting point for future reproducibility studies on object-oriented image editing.

Individual Contributions

Owen Zeng led the removal reproduction track: benchmark download and preprocessing, metric implementation and calibration (CMMD bandwidth sweep, ReMOVE backbone diagnosis, FID fallback), and the seven-metric reproduction table.

Jiadong (Tony) Zhang led the insertion track: substitute-benchmark construction, pipeline integration for four insertion baselines, metric-cropping fix, and the ranking-sensitivity analysis.

Maochuan Lu led infrastructure and systems: Colab environment, dependency pinning and conflict resolution (numpy, carvekit, Pillow), weight downloads and HuggingFace gated-repo setup, report writing, and overall project coordination.

All three members jointly read the paper and contributed to the Methodology and Experimental Results sections.

References

[1] Yu, Y., Liu, Z., Wu, X., Yang, H., and Shan, Y. OMNIPAINT: Mastering Object-Oriented Editing via Disentangled Insertion-Removal Inpainting. *ICCV*, 2025. arXiv:2503.08677.

[2] Black Forest Labs. FLUX.1-dev: a 12B rectified-flow transformer for text-to-image synthesis, 2024. https://huggingface.co/black-forest-labs/FLUX.1-dev.

[3] Hu, E. et al. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*, 2022.

[4] Suvorov, R. et al. Resolution-Robust Large Mask Inpainting with Fourier Convolutions. *WACV*, 2022.

[5] Li, W. et al. MAT: Mask-Aware Transformer for Large-Hole Image Inpainting. *CVPR*, 2022.

[6] Rombach, R. et al. High-Resolution Image Synthesis with Latent Diffusion Models. *CVPR*, 2022.

[7] Yang, B. et al. Paint by Example: Exemplar-based Image Editing with Diffusion Models. *CVPR*, 2023.

[8] Song, Y. et al. ObjectStitch: Object Compositing with Diffusion Model. *CVPR*, 2023.

[9] Zhang, L. et al. FreeCompose: Generic Zero-Shot Image Composition with Diffusion Prior. *ECCV*, 2024.

[10] Chen, X. et al. AnyDoor: Zero-shot Object-level Image Customization. *CVPR*, 2024.

[11] Song, Y. et al. IMPRINT: Generative Object Compositing by Learning Identity-Preserving Representation. *CVPR*, 2024.

[12] Jayasumana, S. et al. Rethinking FID: Towards a Better Evaluation Metric for Image Generation. *CVPR*, 2024.

[13] Chandrasekar, A. et al. ReMOVE: A Reference-Free Metric for Object Erasure. *CVPR Workshops*, 2024.

[14] Ruiz, N. et al. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. *CVPR*, 2023.

[15] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. *ICML*, 2021.

[16] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. DINOv2: Learning Robust Visual Features without Supervision. *TMLR*, 2024.

[17] Zhang, R. et al. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *CVPR*, 2018.

[18] Fu, S. et al. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. *NeurIPS*, 2023.

[19] Ke, J. et al. MUSIQ: Multi-Scale Image Quality Transformer. *ICCV*, 2021.

[20] Yang, S. et al. MANIQA: Multi-Dimension Attention Network for No-Reference Image Quality Assessment. *CVPR Workshops*, 2022.