WEAVE: WARP EVERY ANGLE VIA 3D GEOMETRY FOR REFERENCE-GUIDED IMAGE COMPLETION

Anonymous authors
Paper under double-blind review

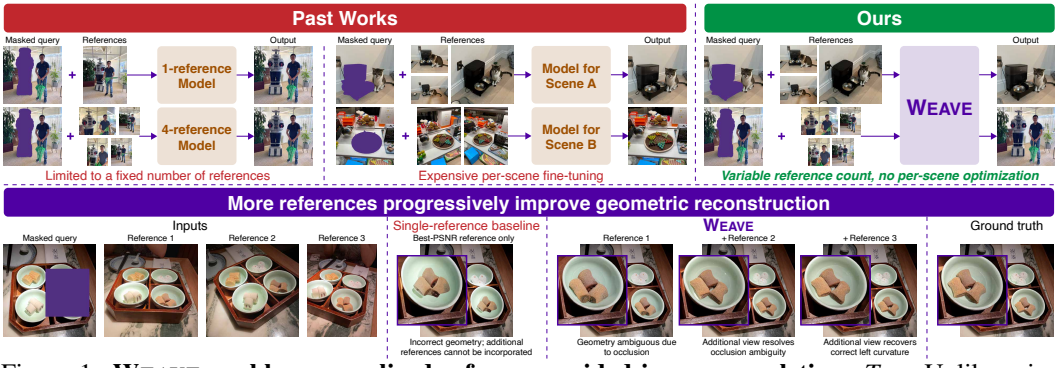

Figure 1: **WEAVE enables generalized reference-guided image completion.** *Top:* Unlike prior methods that use a fixed number of references or require costly per-scene optimization, WEAVE handles variable reference counts with a single feed-forward model, achieving stronger performance at low inference cost (Table 4). *Bottom:* Additional views provide complementary evidence, allowing WEAVE to resolve occlusions, improve local appearance, and better recover target geometry.

ABSTRACT

Reference-guided image completion uses auxiliary views of a scene to reconstruct missing image content faithfully. Existing approaches often require per-scene optimization, pairwise alignment, or a fixed number of reference images, limiting their flexibility and practical use. We introduce WEAVE, a feed-forward framework that supports variable numbers of references without per-scene fine-tuning. WEAVE estimates camera poses and depths in a single forward pass and uses them to reproject each reference into the target view. These projections form two complementary conditioning inputs: a merged query warp (Q-Warp) providing a target-aligned estimate, and separate reference warps (R-Warps) retaining view-specific details and disagreements that merging would discard. WEAVE combines these inputs with the original references to resolve cross-view discrepancies and synthesize the missing content. Experiments on MegaDepth and RealBench show consistent improvements with both MM-DiT and U-Net backbones. WEAVE achieves state-of-the-art reconstruction performance on both benchmarks, outperforming scene-finetuned methods on RealBench without test-time optimization.

1 INTRODUCTION

Conventional image completion methods produce visually convincing results, but may hallucinate scene content absent from the original image (Rombach et al., 2022; Suvorov et al., 2022; Podell et al., 2024; Gong et al., 2025). Reference-guided image completion addresses this limitation by providing the model with alternate views of the same scene, which is critical for applications requiring visual fidelity and geometric accuracy. However, recovering exact visual fidelity is inherently challenging. Reference images are rarely captured from the target's viewpoint, introducing parallax, scale variations, and non-linear perspective distortions across views. Furthermore, complex 3D scenes cause disocclusions and self-occlusions, leaving target regions visible in only a subset of views or hidden entirely. Handling these discrepancies requires accurate spatial correspondence and reasoning about the underlying 3D geometry across substantially different viewpoints.

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.

Existing approaches generally follow two paradigms. Explicit geometric methods first estimate cross-view geometry (*e.g.*, depth, camera pose) and then warp observed pixels from the reference images into the target view (Zhou et al., 2021; Zhao et al., 2023; Liu et al., 2025b). Warping preserves reference appearance where correspondence is accurate, but becomes unreliable under occlusions, disocclusions, or geometric misalignment, particularly when target regions are not clearly visible in the references. Generative methods instead condition a diffusion model on reference images and use its learned prior to synthesize missing target content (Tang et al., 2024; Cao et al., 2024; Liu et al., 2025a). These methods can generate plausible content despite incomplete visibility, but may misplace structures or alter scene details without spatial correspondence across viewpoints. Prior methods also introduce practical constraints. Some require per-scene optimization at test time, resulting in tens of minutes of computation for each scene (Tang et al., 2024; Lin et al., 2025). Others are designed for a fixed number of reference images, limiting flexibility when the number of available views varies across scenes (Cao et al., 2024; Liu et al., 2025a).

We propose WEAVE (**W**arp **E**very **A**ngle **V**ia 3D g**E**ometry), a feed-forward framework that grounds a generative denoising prior (Liu et al., 2023; Black Forest Labs, 2024; 2025) in geometry estimated jointly from the masked target and references. Unlike prior work, WEAVE supports a variable number of references without test-time fine-tuning. We condition the generative model on two representations derived from reference-to-target correspondences. A *Query Warp (Q-Warp)* merges reprojected references into a target-view estimate, providing consolidated context for the missing region. Because merging can discard source identity and conflicting observations, *Reference Warps (R-Warps)* retain each reprojected reference separately, preserving view-specific details. The model jointly attends to both, using the Q-Warp for target-view context and the R-Warps to resolve ambiguity from correspondence errors, occlusions, and appearance differences.

Our dual-path conditioning generalizes across generative architectures. We instantiate WEAVE with the FLUX.2 [klein] 4B MM-DiT backbone (Black Forest Labs, 2025), and WEAVE$_{[UN]}$ with Stable Diffusion 2 Inpainting (Rombach et al., 2022) for fair comparison with prior methods using the same backbone. Across the MegaDepth and RealBench benchmarks, WEAVE achieves state-of-the-art reconstruction fidelity. On RealBench, WEAVE surpasses the strongest scene-optimized baseline in PSNR$_{\text{mask}}$ without per-scene optimization. Our contributions are:

- We introduce **WEAVE**, a feed-forward framework for reference-guided image completion that supports variable reference counts without per-scene fine-tuning.
- Our proposed **dual-path geometric conditioning** combines a merged Q-Warp with per-reference R-Warps to preserve complementary and conflicting observations.
- Our framework generalizes across backbones. Both MM-DiT and U-Net instantiations improve reconstruction on MegaDepth and RealBench, with WEAVE achieving state-of-the-art masked-region PSNR on RealBench, surpassing scene-optimized methods.

2 RELATED WORK

Reference-Guided Image Completion. Early pixel transport methods geometrically warp reference pixels (Zhou et al., 2021; Zhao et al., 2023), while TransRef (Liu et al., 2025b) learns to align and fuse reference features. Among diffusion methods, RealFill (Tang et al., 2024) adapts the model per scene, while BlipFill (Wang et al., 2024a) uses averaged BLIP-2 subject embeddings to guide subject-specific fine-tuning. LeftRefill (Cao et al., 2024) avoids test-time optimization but requires a separately trained checkpoint for each fixed reference count; CorrFill (Liu et al., 2025a) keeps model weights fixed but optimizes latents during sampling. GeoComplete (Lin et al., 2025) combines geometry with diffusion through projected point-cloud conditioning, but retains per-scene fitting. In contrast, WEAVE eliminates per-scene fitting through a single model trained across scenes that supports variable reference counts at inference.

Geometry-Guided Image Generation. A broader line of research integrates 3D structure into generative models to improve cross-view consistency. Epipolar attention restricts cross-view attention to candidate correspondences along epipolar lines (Huang et al., 2024; Kant et al., 2024); these geometric constraints also enable training-free retrieval of reference features to guide novel-view synthesis (Ye et al., 2025). Other methods optimize 3D scenes using 2D inpainting priors (Mirzaei et al., 2023; Chen et al., 2024), with RefFusion (Mirzaei et al., 2024) and IMFine (Shi et al., 2025) additionally adapting diffusion models to each scene. GenWarp (Seo et al., 2024) uses warped co-

ordinate embeddings to guide attention over unwarped source features for single-image novel-view synthesis. However, epipolar-based approaches assume specified camera poses for view synthesis, while scene-based methods require per-scene optimization. WEAVE instead jointly estimates geometry from the masked target and references, directly reprojecting multiple references into the target view and enabling feed-forward completion without scene optimization.

Feed-Forward Geometry Estimation. Recent work makes cross-view geometry more accessible through feed-forward 3D prediction. DUSt3R (Wang et al., 2024b) predicts pairwise pointmaps followed by global alignment, while MASt3R (Leroy et al., 2024) adds dense matching and metric-scale predictions. Fast3R (Yang et al., 2025) processes multiple views jointly, and VGGT (Wang et al., 2025) jointly predicts cameras, depth, pointmaps, and tracks. VGGT-Ω (Wang et al., 2026) simplifies this architecture to predict cameras and depth more efficiently. These models recover geometry but do not address how incomplete or conflicting projections should guide image completion. WEAVE uses pretrained 3D information to align reference observations and a generative prior to correct unreliable projections and fill gaps in reference coverage.

3 METHOD

Our goal is to recover missing content in a partially observed target image from one or more reference views of the same scene. WEAVE estimates the geometry relating these views, reprojects their observations into the target frame, and uses the resulting aligned evidence to guide generation.

Formally, let $\mathbf{I}_0 \in \mathbb{R}^{3\times H_0\times W_0}$ denote the target image and $\mathbf{M} \in \{0,1\}^{H_0\times W_0}$ a binary mask, where $\mathbf{M}=1$ indicates the unknown region; the observed target is thus $\mathbf{I}_0^m = \mathbf{I}_0 \odot (\mathbf{1}-\mathbf{M}) \in \mathbb{R}^{3\times H_0\times W_0}$. We are also given N reference images $\{\mathbf{I}_j\}_{j=1}^N$ depicting the same scene from different viewpoints, where $\mathbf{I}_j \in \mathbb{R}^{3\times H_j\times W_j}$. The number of available references N may vary across examples and between training and inference. We train a completion model $\mathcal{F}_\theta$ that predicts

$$\hat{\mathbf{I}}_0 = \mathcal{F}_\theta\left(\mathbf{I}_0^m, \mathbf{M}, \{\mathbf{I}_j\}_{j=1}^N\right) \in \mathbb{R}^{3\times H_0\times W_0}. \tag{1}$$

The completed image should preserve the observed target pixels, $(\mathbf{1}-\mathbf{M}) \odot \hat{\mathbf{I}}_0 \approx (\mathbf{1}-\mathbf{M}) \odot \mathbf{I}_0$, while recovering the underlying scene appearance within the masked region $\mathbf{M} \odot \hat{\mathbf{I}}_0$.

3.1 WEAVE OVERVIEW

Geometry estimation. We jointly estimate the geometry of the masked target and all N reference views using a pretrained geometry foundation model. Given $\{\mathbf{I}_0^m, \mathbf{I}_1, \ldots, \mathbf{I}_N\}$, the geometry encoder Φ predicts an extrinsic matrix $\mathbf{E}_j = [\mathbf{R}_j \,|\, \mathbf{t}_j] \in \mathbb{R}^{3\times 4}$, an intrinsic matrix $\mathbf{K}_j \in \mathbb{R}^{3\times 3}$, and a dense metric depth map $\mathbf{D}_j \in \mathbb{R}^{H_j\times W_j}$ for every frame:

$$\{(\mathbf{E}_j, \mathbf{K}_j, \mathbf{D}_j)\}_{j=0}^N = \Phi\left(\mathbf{I}_0^m, \mathbf{I}_1, \ldots, \mathbf{I}_N\right). \tag{2}$$

The target remains masked, with reference-to-mask attention blocked in Φ to prevent leakage; the pretrained encoder estimates geometry for all views in one forward pass.

Dual-path conditioning. Using the estimated geometry, we construct two complementary target-aligned conditioning representations. Our *reference warps* (**R-Warps**) independently reproject each reference into the target frame while preserving its view-specific evidence. Keeping these projections separate allows the generator to compare observations across views when estimated correspondences disagree. Concretely, for each reference view j, we use its predicted depth to back-project every pixel through the source camera, transform the resulting 3D point into the target camera frame using the relative pose derived from $\mathbf{E}_j$ and $\mathbf{E}_0$, and project it through the target intrinsics $\mathbf{K}_0$. Forward splatting with a z-buffer retains the nearest point when multiple pixels from the same reference map to the same target location. We write this operation as

$$(\mathbf{W}_j, \mathbf{C}_j, \mathbf{Z}_j) = \mathrm{Warp}\left(\mathbf{I}_j \,|\, \mathbf{D}_j, \mathbf{K}_j, \mathbf{K}_0, \mathbf{R}_{0j}, \mathbf{t}_{0j}\right), \qquad j \in \{1,\ldots,N\}, \tag{3}$$

where $\mathbf{W}_j \in \mathbb{R}^{3\times H_0\times W_0}$ is the reprojected reference, $\mathbf{C}_j \in \{0,1\}^{H_0\times W_0}$ indicates valid coverage, and $\mathbf{Z}_j \in \mathbb{R}^{H_0\times W_0}$ stores the corresponding depth in the target camera frame. We apply a morphological fill to close small holes introduced by forward splatting.

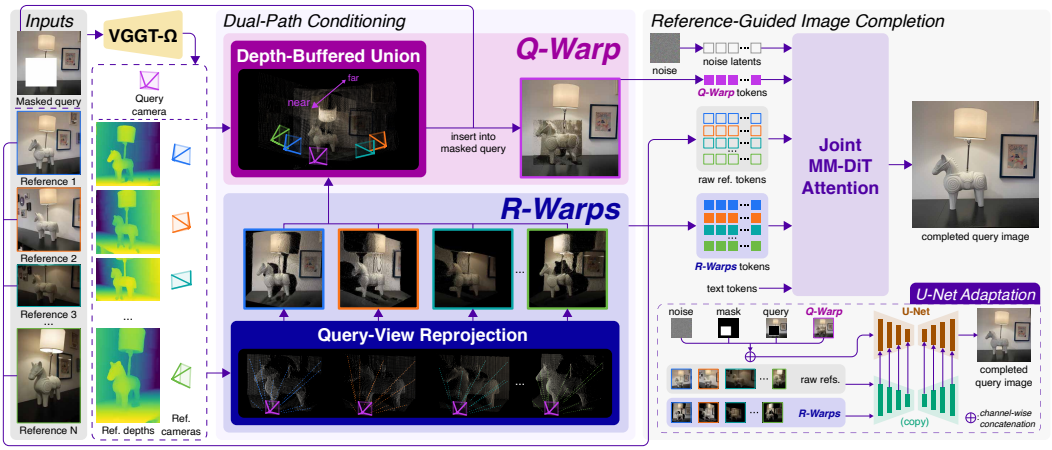

Figure 2: **WEAVE overview.** WEAVE completes a masked target from variable references by grounding a generative model in 3D geometry. Estimated camera poses and depths are used to reproject each reference into the target frame as a geometry-aligned conditioning stream. We propose a novel Dual-Path Conditioning approach comprising a query warp (Q-Warp) and spatially aligned reference warps (R-Warps) to unify multiple references for image completion.

We then construct the *query warp* (**Q-Warp**) by merging R-Warps into a single target-aligned estimate. For each target pixel $\mathbf{u}$, we select the frontmost projected observation:

$$j^\star(\mathbf{u}) = \arg \min_{j:\mathbf{C}_j(\mathbf{u})=1} \mathbf{Z}_j(\mathbf{u}). \tag{4}$$

The resulting depth-buffered union is

$$\widetilde{\mathbf{C}}^{\cup}(\mathbf{u}) = \mathbb{1}\left[\sum_j \mathbf{C}_j(\mathbf{u}) > 0\right], \qquad \widetilde{\mathbf{W}}^{\cup}(\mathbf{u}) = \begin{cases} \mathbf{W}_{j^\star(\mathbf{u})}(\mathbf{u}), & \widetilde{\mathbf{C}}^{\cup}(\mathbf{u}) = 1, \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{5}$$

We apply a morphological fill to close small holes, yielding $(\mathbf{W}^{\cup}, \mathbf{C}^{\cup})$. The union retains the frontmost reprojected observation at each target location and is inserted into the masked target:

$$\bar{\mathbf{W}} = \mathbf{I}_0^m + \mathbf{W}^{\cup} \odot (\mathbf{M} \odot \mathbf{C}^{\cup}) \in \mathbb{R}^{3\times H_0 \times W_0}, \tag{6}$$

where masks are broadcast across color channels. The resulting Q-Warp preserves the observed target, fills masked pixels supported by at least one reference, and leaves unsupported pixels empty.

Geometry-aligned conditioning. Our primary WEAVE model uses an MM-DiT backbone (Esser et al., 2024). The generator receives the Q-Warp $\bar{\mathbf{W}}$, the raw references $\{\mathbf{I}_j\}_{j=1}^N$, and the R-Warps $\{\mathbf{W}_j\}_{j=1}^N$. The raw references preserve appearance and context that may be lost through incomplete or inaccurate reprojection. Let $\mathcal{E}$ denote the VAE encoder and Tok the patchification and token-projection operation. We encode the three inputs as $\mathbf{s}^{\mathrm{q}} = \mathrm{Tok}\big(\mathcal{E}(\bar{\mathbf{W}})\big)$, $\mathbf{s}_j^{\mathrm{ref}} = \mathrm{Tok}(\mathcal{E}(\mathbf{I}_j))$, and $\mathbf{s}_j^{\mathrm{warp}} = \mathrm{Tok}(\mathcal{E}(\mathbf{W}_j))$ for $j \in \{1, \ldots, N\}$, where each block lies in $\mathbb{R}^{L\times d}$ for a block-specific token count L and hidden dimension d. We concatenate them with the patchified noisy target latent:

$$\left[\mathrm{Tok}(\mathbf{z}_\tau) \;\middle|\; \mathbf{s}^{\mathrm{q}} \;\middle|\; \mathbf{s}_1^{\mathrm{ref}}, \ldots, \mathbf{s}_N^{\mathrm{ref}} \;\middle|\; \mathbf{s}_1^{\mathrm{warp}}, \ldots, \mathbf{s}_N^{\mathrm{warp}}\right]. \tag{7}$$

Joint attention combines the merged Q-Warp with the per-view warped and raw reference evidence.

U-Net adaptation. To test WEAVE's transferability beyond MM-DiTs and enable comparison with prior U-Net-based methods, we instantiate it on a U-Net backbone, denoted WEAVE$_{[\mathrm{UN}]}$. We concatenate the four-channel VAE encoding of the Q-Warp with the standard nine-channel inpainting input, yielding a 13-channel U-Net input. The raw references and their individually query-aligned R-Warps are encoded by a trainable ReferenceNet (Hu, 2024). At each self-attention layer, the denoiser queries attend to keys and values from both its own tokens and the cached reference tokens. Appendix B formalizes our adaptation process in detail.

Table 1: **Reference-based inpainting on MegaDepth. Bold** and underlined numbers denote the best and second best scores, respectively.

Method	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓	CLIP↑	DINO↑
Single-image (1 reference)						
BAGEL (Deng et al., 2025)	16.12	0.813	0.203	0.225	0.863	0.841
Qwen Image Edit 2511 (Wu et al., 2025)	17.64	0.830	0.192	0.245	0.849	0.840
FLUX.2 [klein] 4B (Black Forest Labs, 2025)	19.11	0.831	0.140	0.101	0.944	0.911
FLUX.2 [klein] 9B (Black Forest Labs, 2025)	21.06	0.850	0.112	0.071	0.962	0.944
FLUX.2 [dev] (Black Forest Labs, 2025)	22.30	0.872	0.098	0.073	0.957	0.940
GPT Image 2 (OpenAI, 2026)	24.41	0.893	0.057	0.022	0.988	0.983
Nano Banana Pro (Gemini Team, 2023)	24.48	0.894	0.088	0.058	0.971	0.956
FLUX.1 Fill [dev] (Black Forest Labs, 2024)	24.99	0.901	0.062	0.034	0.978	0.973
CorrFill (Liu et al., 2025a)	24.72	0.897	0.052	0.021	**0.992**	0.981
LeftRefill (Cao et al., 2024)	24.75	0.897	0.051	0.021	**0.992**	0.981
WEAVE[UN]	26.01	0.907	0.049	0.019	0.987	0.982
WEAVE	**27.49**	**0.918**	**0.037**	**0.014**	0.990	**0.990**
Multi-image (4 references)						
BAGEL (Deng et al., 2025)	16.19	0.808	0.208	0.250	0.865	0.816
Qwen Image Edit 2511 (Wu et al., 2025)	15.74	0.803	0.231	0.362	0.759	0.754
FLUX.2 [klein] 4B (Black Forest Labs, 2025)	17.70	0.813	0.177	0.146	0.925	0.881
FLUX.2 [dev] (Black Forest Labs, 2025)	18.65	0.830	0.162	0.136	0.924	0.894
FLUX.2 [klein] 9B (Black Forest Labs, 2025)	19.50	0.832	0.141	0.112	0.945	0.913
GPT Image 2 (OpenAI, 2026)	24.17	0.891	0.060	0.022	0.988	0.982
Nano Banana Pro (Gemini Team, 2023)	25.09	0.901	0.078	0.047	0.976	0.964
CorrFill (Liu et al., 2025a)	23.71	0.890	0.063	0.034	0.988	0.970
LeftRefill (Cao et al., 2024)	23.80	0.890	0.062	0.034	0.988	0.970
WEAVE[UN]	26.48	0.911	0.046	0.016	**0.992**	0.984
WEAVE	**28.28**	**0.925**	**0.033**	**0.012**	0.991	**0.991**

3.2 Training Objective

We train both backbones with latent-space denoising objective $\mathcal{L} = \mathbb{E}_{\mathbf{z}_0,\boldsymbol{\epsilon},\tau}\left[\|f_\theta(\mathbf{z}_\tau, \mathcal{C}, \tau) - \mathbf{u}^\star\|_2^2\right]$, where $\mathbf{z}_0$ is the target latent, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathcal{C}$ denotes the masked query and reference-image conditioning. For WEAVE, we use rectified-flow training (Liu et al., 2023): $\mathbf{z}_\tau = (1-\tau)\mathbf{z}_0 + \tau\boldsymbol{\epsilon}$ and $\mathbf{u}^\star = \boldsymbol{\epsilon} - \mathbf{z}_0$. For WEAVE$_{[UN]}$, we use standard DDPM ϵ-prediction, with $\mathbf{u}^\star = \boldsymbol{\epsilon}$. At inference, both variants use fixed model weights and their native sampling procedures.

4 Experiments

4.1 Experimental Setup

Datasets. We train WEAVE on MegaDepth (Li & Snavely, 2018), sampling $N \in \{1, \ldots, 4\}$ references for each example. We evaluate in-domain performance on MegaDepth using one and four references, following past work (Cao et al., 2024; Liu et al., 2025a). Additionally, we test out-of-distribution generalization on RealBench (Tang et al., 2024), which contains diverse real-world multi-reference scenes. Appendices A and B provide additional dataset and implementation details.

Baselines. We compare WEAVE against baselines spanning four paradigms: reference-guided specialists (Cao et al., 2024; Liu et al., 2025a; Zhou et al., 2021), per-scene optimization approaches (Tang et al., 2024; Lin et al., 2025), generic single-view inpainters (Rombach et al., 2022; Adobe Inc., 2024; Yang et al., 2023), and zero-shot multimodal foundation models (Wu et al., 2025; Deng et al., 2025; Gemini Team, 2023; Black Forest Labs, 2024; 2025; OpenAI, 2026). Because LeftRefill is trained on MegaDepth and CorrFill builds on its pretrained model, neither is zero-shot on MegaDepth; the remaining baselines are evaluated without MegaDepth training. Since LeftRefill and CorrFill support only one reference with their available checkpoints, we follow RealFill (Tang et al., 2024) to sample one reference from the available set for multi-reference evaluation.

Metrics. For MegaDepth, we report PSNR and SSIM (Wang et al., 2004) for structural fidelity, LPIPS (Zhang et al., 2018) and DreamSim (Fu et al., 2023) for perceptual similarity, CLIP and DINO (Radford et al., 2021; Caron et al., 2021) feature similarity. For RealBench, following Real-

Table 2: **Zero-shot RealBench completion**. Gray rows show per-scene fine-tuning models; **bold** and underlined scores indicate the best and second best without per-scene fine-tuning.

Method	$\text{PSNR}_{\text{mask}}$ ↑	$\text{SSIM}_{\text{mask}}$ ↑	$\text{LPIPS}_{\text{mask}}$ ↓	DreamSim↓	CLIP↑	DINO↑
Paint-by-Example (Yang et al., 2023)	10.13	0.244	0.642	0.196	0.837	0.808
BAGEL (Deng et al., 2025)	10.33	0.250	0.628	0.179	0.873	0.857
Qwen Image Edit 2511 (Wu et al., 2025)	10.34	0.255	0.668	0.211	0.851	0.820
SD2-Inpaint (Rombach et al., 2022)	10.63	0.282	0.605	0.213	0.874	0.831
Adobe Generative Fill (Adobe Inc., 2024)	10.92	0.311	0.598	0.212	0.898	0.851
FLUX.2 [dev] (Black Forest Labs, 2025)	10.97	0.280	0.583	0.116	0.918	0.904
FLUX.2 [klein] 4B (Black Forest Labs, 2025)	11.65	0.288	0.549	0.104	0.925	0.911
CorrFill (Liu et al., 2025a)	12.10	0.360	0.451	0.095	0.919	0.916
GPT Image 2 (OpenAI, 2026)	12.14	0.339	0.478	0.090	0.940	0.923
FLUX.2 [klein] 9B (Black Forest Labs, 2025)	12.17	0.311	0.517	0.109	0.921	0.910
LeftRefill (Cao et al., 2024)	12.29	0.364	0.461	0.108	0.900	0.916
FLUX.1 Fill [dev] (Black Forest Labs, 2024)	12.81	0.334	0.534	0.168	0.847	0.828
TransFill (Zhou et al., 2021)	13.28	0.404	0.542	0.192	0.866	0.860
Nano Banana Pro (Gemini Team, 2023)	13.56	0.371	0.434	0.077	0.943	0.937
RealFill (Tang et al., 2024)	14.78	0.424	0.431	0.077	0.962	0.948
GeoComplete (Lin et al., 2025)	17.32	0.578	0.197	0.036	0.987	0.986
WEAVE[UN]	<u>15.58</u>	<u>0.450</u>	<u>0.342</u>	<u>0.077</u>	0.926	0.931
WEAVE	**18.35**	**0.576**	**0.166**	**0.027**	**0.976**	**0.976**

Fill (Tang et al., 2024), we compute $\text{PSNR}_{\text{mask}}$, $\text{SSIM}_{\text{mask}}$, and $\text{LPIPS}_{\text{mask}}$ only over the masked region, alongside full-image DreamSim, CLIP, and DINO.

4.2 Main Results

In-domain (MegaDepth). Table 1 reports results on MegaDepth with one and four references. WEAVE[UN], our Stable Diffusion 2 Inpainting (SD2I) variant, reaches 26.01 PSNR with one reference, outperforming the SD2I-based LeftRefill and CorrFill baselines by 1.26 and 1.29 dB, respectively. Our primary MM-DiT model, WEAVE, performs best overall, exceeding the strongest evaluated foundation model, FLUX.1 Fill [dev], by 2.50 dB. It also outperforms Nano Banana Pro and GPT-Image 2, with corresponding improvements across all metrics. With four references, both variants improve further: WEAVE[UN] reaches 26.48 PSNR, while WEAVE reaches 28.28 PSNR. WEAVE exceeds Nano Banana Pro and GPT-Image 2 by 3.19 and 4.11 dB, respectively, and achieves the best LPIPS, DreamSim, and DINO alongside comparable CLIP scores.

Generalization (RealBench). Table 2 reports zero-shot generalization to multi-reference RealBench scenes. Among SD2I-based methods without per-scene fine-tuning, WEAVE[UN] achieves the highest PSNRmask at 15.58 dB, outperforming CorrFill and LeftRefill by 3.48 and 3.29 dB. Unlike CorrFill, which performs test-time latent optimization, WEAVE[UN] is fully feed-forward, and it exceeds RealFill by 0.80 dB despite RealFill requiring roughly an hour of per-scene fine-tuning. Our primary WEAVE further improves performance, surpassing GeoComplete by 1.03 dB while remaining feed-forward. It also achieves the best LPIPSmask and DreamSim, with comparable $\text{SSIM}_{\text{mask}}$, and exceeds Nano Banana Pro by 4.79 dB. Overall, WEAVE achieves the strongest RealBench reconstruction performance among evaluated methods without per-scene optimization.

4.3 Ablations

Table 3 isolates the contributions of raw references, Q-Warp, and R-Warps on RealBench. Conditioning on raw references alone achieves 12.90 $\text{PSNR}_{\text{mask}}$ on WEAVE[UN] and 15.99 on WEAVE. Adding Q-Warp results in gains of 1.17 and 1.03 dB, respectively, while adding R-Warps produces larger improvements of 1.85 and 2.19 dB over raw references. Combining all three inputs yields gains over raw references by 2.68 and 2.36 dB. This progression shows that both forms of geometric conditioning improve reconstruction, with R-Warps providing the larger individual contribution.

The two warp representations are also complementary. Starting from raw references and Q-Warp, adding R-Warps improves $\text{PSNR}_{\text{mask}}$ by 1.51 dB on WEAVE[UN] and 1.33 dB on WEAVE. Conversely, adding Q-Warp to raw references and R-Warps yields gains of 0.83 and 0.17 dB. Raw references remain useful alongside both geometric paths: adding them to Q-Warp-plus-R-Warps improves $\text{PSNR}_{\text{mask}}$ by 0.13 and 0.42 dB, respectively. Therefore, R-Warps supply the strongest geometric cue, while Q-Warp and raw references provide beneficial complementary information.

Table 3: **Architecture ablation**. We compare combinations of raw references, Q-Warp, and R-Warps on RealBench across both the U-Net and MM-DiT backbones.

	raw refs.	Q-Warp	R-Warps	$\text{PSNR}_{\text{mask}}$ ↑	$\text{SSIM}_{\text{mask}}$ ↑	$\text{LPIPS}_{\text{mask}}$ ↓	DreamSim↓
WEAVE[UN]	✓	✗	✗	12.90	0.338	0.456	0.097
	✗	✓	✗	13.11	0.371	0.449	0.145
	✓	✓	✗	14.07	0.378	0.411	0.091
	✓	✗	✓	14.75	0.425	0.353	**0.075**
	✗	✗	✓	14.78	0.422	0.354	0.079
	✗	✓	✓	15.45	0.437	0.355	0.082
	✓	✓	✓	**15.58**	**0.450**	**0.342**	<u>0.077</u>
WEAVE	✓	✗	✗	15.99	0.497	0.250	0.035
	✗	✓	✗	16.08	0.469	0.264	0.048
	✓	✓	✗	17.02	0.511	0.197	0.030
	✗	✗	✓	17.71	0.560	0.189	0.032
	✗	✓	✓	17.93	0.562	0.186	0.032
	✓	✗	✓	18.18	0.575	0.171	**0.027**
	✓	✓	✓	**18.35**	**0.576**	**0.166**	**0.027**

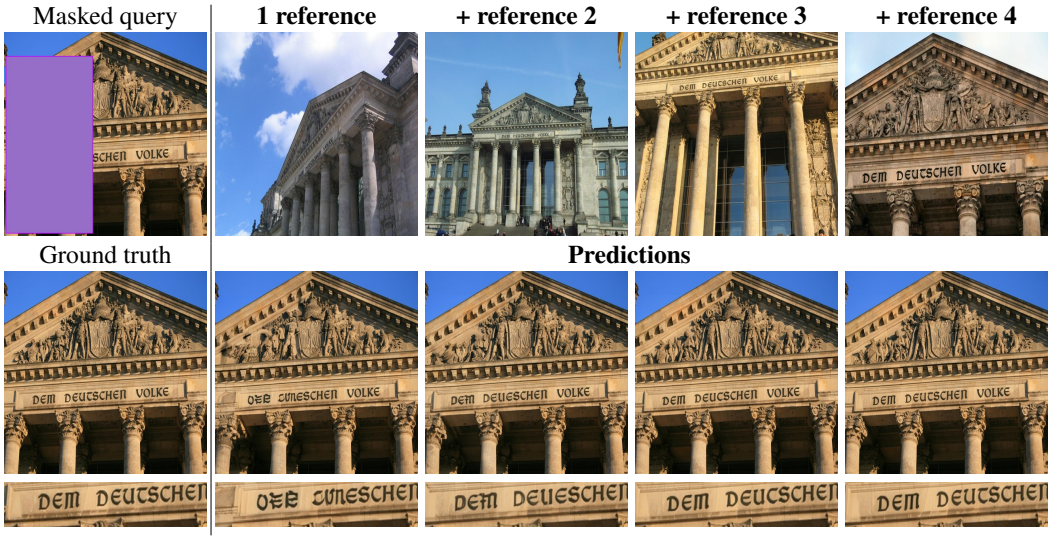

Figure 3: **Adding references incrementally refines WEAVE prediction.**

The full configuration achieves the best $\text{PSNR}_{\text{mask}}$, $\text{SSIM}_{\text{mask}}$, and $\text{LPIPS}_{\text{mask}}$ on both backbones. For DreamSim, the full WEAVE configuration ties the best result at 0.027. On $\text{WEAVE}_{\text{[UN]}}$, raw references plus R-Warps is marginally better (0.075 vs. 0.077). Overall, the ablation demonstrates that the largest gains come from geometry-aligned reference conditioning, while the merged Q-Warp and raw references provide complementary information that further improves the full model.

4.4 ANALYSIS

Reference count. We find that additional references can help, but only when models can reconcile their viewpoints. As shown in Table 1, increasing the reference count from one to four reduces PSNR for Qwen Image Edit 2511 (17.64 to 15.74) and FLUX.2 [klein] 4B (19.11 to 17.70), whereas $\text{WEAVE}_{\text{[UN]}}$ improves by 0.47 dB to 26.48 and WEAVE by 0.79 dB to 28.28. With four references, our models outperform specialists LeftRefill and CorrFill, as well as foundation models Nano Banana Pro and GPT-Image 2; WEAVE also achieves the best LPIPS (0.033) and DreamSim (0.012). Qualitatively, Figs. 1 (bottom) and 3 show the completion progressively approaching the ground truth as references are added. In the former, successive references resolve an occluded region of the dessert and recover the target curvature more faithfully. In Fig. 3, from a single oblique view, the inscription is illegible and the model generates plausible but incorrect lettering. As more frontal views become available, the prediction incrementally refines the engraving, demonstrating that each reference contributes complementary evidence. Analyses in Appendix C.1 support training WEAVE with variable reference counts of between one to four per training example.



Figure 4: **Warp conditioning aligns reference attention with the target.** *Left*: attention from the marked target location to raw references (top) and R-Warps (bottom), with each column showing the same reference view. *Right*: the masked target and ground-truth crop, alongside the baseline and WEAVE predictions. WEAVE more closely recovers the teddy bear's facial details and proportions. In the masked target, ☆ marks the query pixel, ■ indicates the masked region whose original pixels are hidden from the model, and □ identifies the zoomed region and corresponding crops.

Table 4: **Per-scene computational cost on RealBench.** † Paired times indicate measured single-reference runtime / estimated total runtime over all references.

Method	Test-time optimization	Time / scene	Hardware	$\text{PSNR}_{\text{mask}}$ ↑
Per-scene optimized				
GeoComplete (Lin et al., 2025)	LoRA, 2000 steps	∼72 min	24 GB	17.32
RealFill (Tang et al., 2024)	LoRA, 2000 steps	∼1 hour	A100	14.78
CorrFill (Liu et al., 2025a)	per-image latent optimization	∼39.1 s / ∼126.6 s†	A100	12.10
No test-time optimization				
LeftRefill (Cao et al., 2024)	none	∼3.5 s / ∼11.2 s†	A100	12.29
WEAVE (ours)	none	∼16.1 s	A100	**18.35**

Attention. Fig. 4 shows how R-Warps align reference content with the target, allowing masked target tokens to attend near the same grid position in each warped view. Quantitatively, 49.9% of R-Warp attention falls within a two-token radius of the co-located target position, *vs.* 1.2% under uniform attention. Moreover, diffusing attention over the combined raw-reference and R-Warp tokens at temperature $T = 2$, while preserving their total attention mass, reduces $\text{PSNR}_{\text{mask}}$ by 4.94 dB; applying the same operation to text attention reduces it by only 0.26 dB. These results indicate that the model uses the aligned reference content and relies on selective attention over reference tokens for accurate completion. Appendix C.2 provides additional details.

Inference efficiency. Table 4 compares reconstruction quality and per-scene runtime on RealBench. WEAVE achieves the highest $\text{PSNR}_{\text{mask}}$ (18.35 dB) in 16.1 s on average on an A100 GPU without per-scene tuning. RealFill (Tang et al., 2024) and GeoComplete (Lin et al., 2025) require approximately one hour on an A100 GPU and 72 minutes on a 24 GB GPU per scene,[1] respectively, while trailing WEAVE by 3.57 and 1.03 dB $\text{PSNR}_{\text{mask}}$. For the single-reference baselines, processing all references independently yields estimated total runtimes of 126.6 s for CorrFill (Liu et al., 2025a) and 11.2 s for LeftRefill (Cao et al., 2024). WEAVE outperforms CorrFill and LeftRefill by 6.25 and 6.06 dB $\text{PSNR}_{\text{mask}}$, respectively, despite CorrFill's use of test-time latent optimization.

Alternative conditioning mechanisms. We compare alternative conditioning strategies to determine how the form of geometric guidance affects completion quality. Under the same training protocol, warped images supplied as separate token blocks outperform VGGT feature conditioning, epipolar attention bias, and channel concatenation, supporting WEAVE's design for combining spatially aligned reference observations through joint attention. Appendix D provides detailed analyses.

[1] GeoComplete's runtime is taken from Lin et al. (2025), as no public checkpoint was available for direct measurement. They report 72 minutes for GeoComplete and 48 minutes for RealFill on the same 24 GB GPU. For context, A100 measurements are ∼ 16.1 s for WEAVE and ∼ 1 hour for RealFill (Tang et al., 2024).

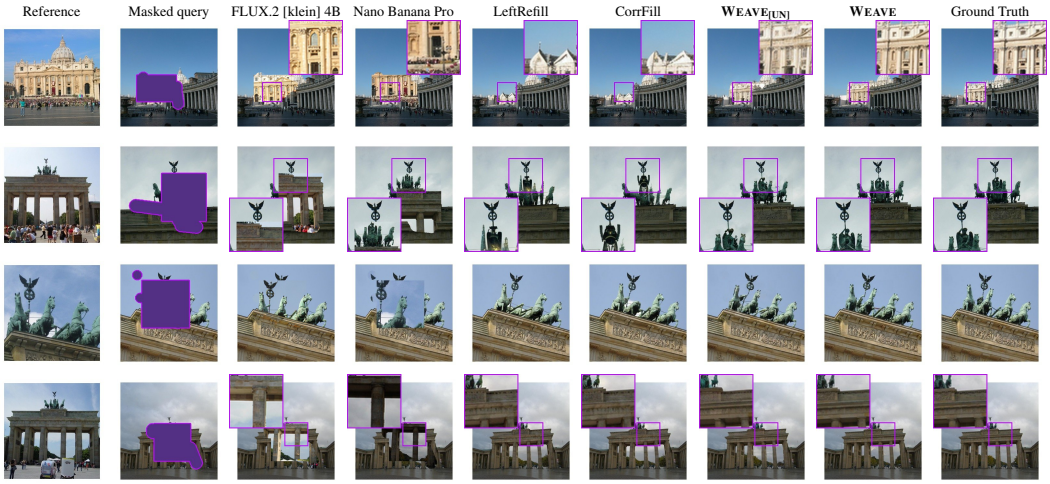

Figure 5: **Single-reference image completion on MegaDepth.**

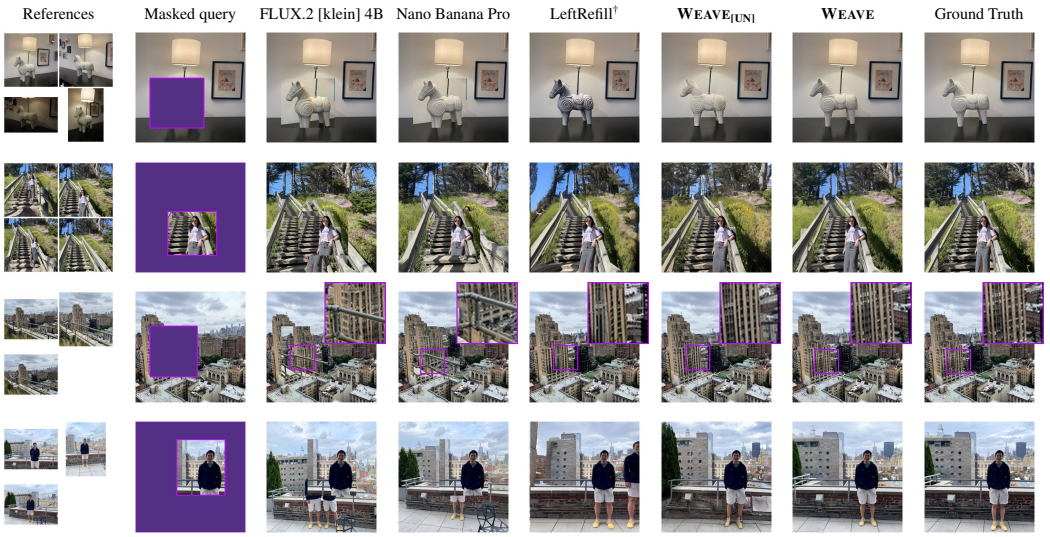

Figure 6: **Zero-shot multi-reference image completion on RealBench.** †: The best prediction is shown for one-reference LeftRefill against all references.

Qualitative analysis. Fig. 5 shows single-reference completion on MegaDepth, where both WEAVE variants better preserve landmark structure while several baselines alter architectural details or omit visible elements. Fig. 6 shows multi-reference completion on RealBench: our models faithfully recover the figurine's shape and texture (row 1), stairway and vegetation (row 2), building window pattern (row 3), and rooftop railing and cityscape (row 4), while baselines introduce texture, structural, and content errors. Both backbones recover the overall scene well, with WEAVE retaining finer details. Appendices E, F, and G provide additional qualitative examples.

5 CONCLUSION

We present WEAVE, a method for reference-guided image completion which uses feed-forward geometry estimates to align reference views. Our dual-path conditioning includes Q-Warp, which provides a merged target-aligned estimate, and R-Warps, which preserve individual observations for the generator to compare and reconcile. Our primary MM-DiT-based WEAVE model and its U-Net adaptation WEAVE$_{[UN]}$ both support variable reference counts without test-time optimization. WEAVE achieves state-of-the-art reconstruction performance on both MegaDepth and RealBench, surpassing scene-optimized baselines on RealBench with fully feed-forward inference.

AI USE STATEMENT

In this work, we did not use generative AI for any tasks requiring disclosure, as these were either performed entirely by the authors or were not applicable to this work. We used generative AI for tasks for which disclosure is recommended: editing and optimizing code to improve experimental efficiency, obtaining feedback on the presentation of experimental results, and refining individual sentences to improve readability and correct grammatical errors.

All AI-assisted code changes were reviewed and tested by the authors, and all AI-assisted suggestions concerning the presentation of results and manuscript text were reviewed before inclusion. Generative AI was not used to determine the scientific conclusions of the work. We take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

To support ethical and responsible use, our data are drawn from publicly available sources and prior works in accordance with applicable ethical standards. This framework is intended for research and should not be used for harmful or malicious purposes. We encourage downstream users to apply the technology responsibly and in accordance with relevant ethical and legal guidelines. The authors declare no conflicts of interest related to this work.

REPRODUCIBILITY STATEMENT

We support reproducibility by providing complete training and evaluation hyperparameters, together with detailed implementation information, in Appendix B. Where applicable, we also include pseudocode and experiment-specific implementation details. These materials are intended to enable replication of our experimental results. We will release the source code and model weights publicly.

REFERENCES

Adobe Inc. *Adobe Photoshop*. Adobe Inc., San Jose, CA, 2024. URL https://www.adobe.com/products/photoshop.html. Version 25.x, featuring Generative Fill capability powered by the Adobe Firefly generative AI model.

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

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Chenjie Cao, Yunuo Cai, Qiaole Dong, Yikai Wang, and Yanwei Fu. LeftRefill: Filling Right Canvas Based on Left Reference through Generalized Text-to-Image Diffusion Model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7705–7715, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Honghua Chen, Chen Change Loy, and Xingang Pan. MVIP-NeRF: Multi-view 3D Inpainting on NeRF Scenes via Diffusion Prior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5344–5353, 2024.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging Properties in Unified Multimodal Pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning New Dimensions of Human Visual Similarity Using Synthetic Data. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:50742–50768, 2023.

Gemini Team. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*, 2023.

Yuan Gong, Xionghui Wang, Jie Wu, Shiyin Wang, Yitong Wang, and Xinglong Wu. OneReward: Unified Mask-Guided Image Generation via Multi-Task Human Preference Learning. *arXiv preprint arXiv:2508.21066*, 2025.

Li Hu. Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8153–8163, 2024.

Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. EpiDiff: Enhancing Multi-View Synthesis via Localized Epipolar-Constrained Diffusion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9784–9794, 2024.

Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. SPAD: Spatially Aware Multi-View Diffusers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10026–10038, 2024.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding Image Matching in 3D with MASt3R. In *European Conference on Computer Vision (ECCV)*, pp. 71–91. Springer, 2024.

Zhengqi Li and Noah Snavely. MegaDepth: Learning Single-View Depth Prediction from Internet Photos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2041–2050, 2018.

Beibei Lin, Tingting Chen, and Robby Tan. GeoComplete: Geometry-Aware Diffusion for Reference-Driven Image Completion. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Kuan-Hung Liu, Cheng-Kun Yang, Min-Hung Chen, Yu-Lun Liu, and Yen-Yu Lin. CorrFill: Enhancing Faithfulness in Reference-Based Inpainting with Correspondence Guidance in Diffusion Models. In *Winter Conference on Applications of Computer Vision (WACV)*, pp. 1618–1627. IEEE, 2025a.

Taorong Liu, Liang Liao, Delin Chen, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin'ichi Satoh. TransRef: Multi-Scale Reference Embedding Transformer for Reference-Guided Image Inpainting. *Neurocomputing*, 632:129749, 2025b.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinshtein. SPIn-NeRF: Multiview Segmentation and Perceptual Inpainting with Neural Radiance Fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20669–20679, 2023.

Ashkan Mirzaei, Riccardo De Lutio, Seung Wook Kim, David Acuna, Jonathan Kelly, Sanja Fidler, Igor Gilitschenski, and Zan Gojcic. RefFusion: Reference Adapted Diffusion Models for 3D Scene Inpainting. *arXiv preprint arXiv:2404.10765*, 2024.

Seoung Wug Oh, Sungho Lee, Joon-Young Lee, and Seon Joo Kim. Onion-Peel Networks for Deep Video Completion. In *International Conference on Computer Vision (ICCV)*, pp. 4403–4412, 2019.

OpenAI. GPT Image 2 Model. https://developers.openai.com/api/docs/models/gpt-image-2, 2026.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. GenWarp: Single Image to Novel Views with Semantic-Preserving Generative Warping. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:80220–80243, 2024.

Zhihao Shi, Dong Huo, Yuhongze Zhou, Yan Min, Juwei Lu, and Xinxin Zuo. IMFine: 3D Inpainting via Geometry-Guided Multi-View Refinement. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26694–26703, 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 *Winter Conference on Applications of Computer Vision (WACV)*, pp. 2149–2159, 2022.

Luming Tang, Nataniel Ruiz, Qinghao Chu, Yuanzhen Li, Aleksander Holynski, David E Jacobs, Bharath Hariharan, Yael Pritch, Neal Wadhwa, Kfir Aberman, et al. RealFill: Reference-Driven Generation for Authentic Image Completion. *ACM Transactions on Graphics (ToG)*, 43(4):1–12, 2024.

Heqing Wang, Tianwei Cao, Ling Jin, Weidong Liu, Rui Wang, and Zhanyu Ma. BlipFill: Completing Authentic Images with BLIP-Enhanced Subject Representation. In *International Conference on Computer Science and Artificial Intelligence*, pp. 112–118, 2024a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing (TIP)*, 13 (4):600–612, 2004.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image Technical Report, 2025. URL https://arxiv.org/abs/2508.02324.

Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep Flow-Guided Video Inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3723–3732, 2019.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by Example: Exemplar-based Image Editing with Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18381–18391, 2023.

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D Reconstruction of 1000+ Images in One Forward Pass. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935, 2025.

Botao Ye, Sifei Liu, Xueting Li, Marc Pollefeys, and Ming-Hsuan Yang. Synthesizing Consistent Novel Views via 3D Epipolar Attention without Re-Training. In *International Conference on 3D Vision (3DV)*, pp. 337–346. IEEE, 2025.

Julio Zaragoza, Tat-Jun Chin, Michael S Brown, and David Suter. As-Projective-As-Possible Image Stitching with Moving DLT. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2339–2346, 2013.

Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-Resolution Image Inpainting with Iterative Confidence Feedback and Guided Upsampling. In *European Conference on Computer Vision (ECCV)*, pp. 1–17. Springer, 2020.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Yunhan Zhao, Connelly Barnes, Yuqian Zhou, Eli Shechtman, Sohrab Amirghodsi, and Charless Fowlkes. GeoFill: Reference-Based Image Inpainting with Better Geometric Understanding. In *Winter Conference on Applications of Computer Vision (WACV)*, pp. 1776–1786, 2023.

Yuqian Zhou, Connelly Barnes, Eli Shechtman, and Sohrab Amirghodsi. TransFill: Reference-guided Image Inpainting by Merging Multiple Color and Spatial Transformations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2266–2276, 2021.

Table 5: **Training and inference settings.**

	WEAVE	WEAVE[UN]
Backbone	FLUX.2-klein-base-4B	SD2 inpainting
Parameterization	Rectified flow	ϵ-prediction
Fine-tuning	Full	Full
Precision	bfloat16	float32
Resolution	Variable, area $\approx 512^2$	512×512
Optimizer	AdamW	AdamW
Learning rate	3×10^{-5}	3×10^{-5}
LR schedule	Cosine	Cosine to 0
Warmup steps	200	120
Gradient clipping	1.0	1.0
Micro-batch / GPU	1	1
Grad. accumulation	1	8
GPUs	8×H200	2×H200
Effective batch	8	16
Sharding	DeepSpeed ZeRO-2	DeepSpeed ZeRO-2
Peak memory / GPU	∼30 GiB	∼31 GiB
Grad. checkpointing	Yes	Yes
Training steps	24,000	8,000
Seed	0	0
Sampling	Rectified flow	DPM++ 2M Karras
Sampling steps	28	25
Guidance scale	1.0	7.5

A DATASET DETAILS

We use MegaDepth for training and in-domain evaluation, and RealBench to evaluate out-of-distribution generalization. Each completion example consists of a masked target image and reference views of the same scene, with the original target serving as ground truth.

MegaDepth. MegaDepth (Li & Snavely, 2018) contains Internet photographs of landmarks and was originally introduced for single-image depth estimation. Its public release covers 196 locations, with camera parameters and depth maps reconstructed using COLMAP structure-from-motion and multi-view stereo. The photographs capture the same structures under different viewpoints, lighting conditions, and occlusions, making them useful for learning to complete images from reference observations. We construct completion examples by masking a target photograph and selecting other views of the same scene as references. During training, we sample between one and four references per example; in-domain evaluation uses one-reference and four-reference configurations. Our model estimates its conditioning geometry directly from the masked target and references using the frozen geometry encoder Φ, instead of using ground truth scene information. We obtain train and test data for MegaDepth following LeftRefill (Cao et al., 2024).

RealBench. RealBench, introduced by RealFill (Tang et al., 2024), contains 33 scenes: 23 for outpainting and 10 for inpainting. Each scene provides a target image, a binary mask, and one to five reference images. The benchmark includes indoor and outdoor scenes with substantial differences in viewpoint, lighting, focus, and subject pose across images. These conditions test whether a model trained on landmark photographs can recover scene-specific content from casually captured references. We evaluate on all 33 scenes using the provided masks and available reference views, without per-scene fine-tuning or test-time optimization; no scenes are used for training.

B IMPLEMENTATION DETAILS

Both WEAVE and WEAVE[UN] use a frozen VGGT-Ω-1B geometry encoder[2] with 512 px inputs, a confidence threshold of 0.5, and mask dilation of 2. The predicted depth and camera parameters are

[2] https://hf.co/facebook/VGGT-Omega/blob/main/vggt_omega_1b_512.pt.

used to construct Q-Warp and R-Warps, which condition the generator alongside the raw references. Both models are trained on MegaDepth with one to four references per example. We use co-visible masks derived from LoFTR correspondences for 25% of training examples and random synthetic masks otherwise. Each backbone's frozen text encoder processes the same instruction: *"Fill in the blacked-out region of the first image, keeping it faithful and consistent with the content, lighting, and details of the reference images."* Table 5 summarizes the training and inference settings.

WEAVE. We fine-tune the undistilled FLUX.2-klein-base-4B transformer[3] in bfloat16 using its rectified-flow objective with logit-normal noise sampling. Images are resized to approximately 512^2 pixels while preserving their aspect ratio, with both dimensions rounded to multiples of 32. Q-Warp, raw references, and R-Warps are encoded as separate token blocks and concatenated with the target tokens. We use a micro-batch size of one to accommodate varying image dimensions and reference counts. Training runs on eight H200 GPUs on a single node, without optimizer offload.

WEAVE[UN]. We fine-tune Stable Diffusion 2 Inpainting[4] and its ReferenceNet copy in float32. Images are resized with the longer side set to 512 and edge-padded to 512×512. The Q-Warp latent is concatenated with the denoiser input, increasing it from nine to 13 channels. A trainable ReferenceNet processes the raw references and R-Warps, supplying keys and values to the denoiser's corresponding self-attention layers. After inference, we composite the generated region with the observed target pixels following LeftRefill (Cao et al., 2024).

We provide our U-Net adaptation process in detail. Formally, let $\downarrow$ denote nearest-neighbor downsampling to the latent resolution. WEAVE[UN]'s denoiser input is

$$\mathbf{x}_t = \left[\mathbf{z}_t \;\middle|\; \downarrow \mathbf{M} \;\middle|\; \mathcal{E}(\mathbf{I}_0^m) \;\middle|\; \mathcal{E}(\bar{\mathbf{W}})\right], \tag{8}$$

where the four blocks have 4, 1, 4, and 4 channels, respectively. Thus, the Q-Warp latent augments the standard nine-channel inpainting input to 13 channels. We initialize the weights connected to the four added channels to zero, preserving pretrained behavior at initialization. The ReferenceNet is initialized as a deep copy of the pretrained Stable Diffusion 2 Inpainting U-Net (*i.e.*, *before* the denoiser's input projection is expanded with the extra channels).

The raw references and R-Warps are processed in a separate reference write pass. For each $\mathbf{R} \in \{\mathbf{I}_1, \ldots, \mathbf{I}_N, \mathbf{W}_1, \ldots, \mathbf{W}_N\}$, the ReferenceNet receives the standard no-hole inpainting input

$$\mathbf{x}^{\mathrm{ref}}(\mathbf{R}) = [\mathcal{E}(\mathbf{R}) \mid \mathbf{0} \mid \mathcal{E}(\mathbf{R})] \tag{9}$$

at timestep zero. Let $\mathbf{H}_\ell$ denote the normalized denoiser tokens entering self-attention layer ℓ, and let $\mathbf{H}^{\mathrm{ref}}_{\ell,j}$ and $\widetilde{\mathbf{H}}^{\mathrm{ref}}_{\ell,j}$ denote the corresponding cached ReferenceNet tokens for $\mathbf{I}_j$ and $\mathbf{W}_j$, respectively. We concatenate all cached reference tokens,

$$\mathbf{H}^{\mathrm{cache}}_\ell = \left[\mathbf{H}^{\mathrm{ref}}_{\ell,1} \;\middle|\; \cdots \;\middle|\; \mathbf{H}^{\mathrm{ref}}_{\ell,N} \;\middle|\; \widetilde{\mathbf{H}}^{\mathrm{ref}}_{\ell,1} \;\middle|\; \cdots \;\middle|\; \widetilde{\mathbf{H}}^{\mathrm{ref}}_{\ell,N}\right], \tag{10}$$

and compute self-attention as

$$\mathbf{Q}_\ell = \mathbf{H}_\ell \mathbf{W}^Q_\ell, \tag{11}$$

$$\mathbf{K}_\ell = \left[\mathbf{H}_\ell \;\middle|\; \mathbf{H}^{\mathrm{cache}}_\ell\right] \mathbf{W}^K_\ell, \tag{12}$$

$$\mathbf{V}_\ell = \left[\mathbf{H}_\ell \;\middle|\; \mathbf{H}^{\mathrm{cache}}_\ell\right] \mathbf{W}^V_\ell, \tag{13}$$

$$\mathrm{Attn}_\ell = \mathrm{softmax}\left(\frac{\mathbf{Q}_\ell \mathbf{K}_\ell^\top}{\sqrt{d}}\right) \mathbf{V}_\ell. \tag{14}$$

Thus, denoiser queries attend jointly to their own tokens and to cached tokens from every raw reference and R-Warp. Text cross-attention is unchanged. Because the reference tokens are concatenated along the sequence dimension, the same weights support a variable number of references N. After inference, we composite the generated content within $\mathbf{M}$ with the observed target pixels.

[3] `https://hf.co/black-forest-labs/FLUX.2-klein-base-4B`.

[4] `https://hf.co/nlightcho/stable-diffusion-2-inpainting` (community upload).

C Additional Experiments

C.1 Reference-Count Generalization

Table 6 compares the number of references used during training with those available at inference. The columns indicate the maximum number of references supplied at inference: for evaluation budget K, each scene s uses the first $\min(K, n_s)$ references from its fixed canonical ordering, where n_s is the number available for that scene. Every entry averages over the same 33 RealBench scenes.

Table 6: **Reference-count generalization for WEAVE.** Colors indicate performance: low → mid → high.

Train ↓ / Max. eval refs. →	1	2	3	4	5	Avg.
N=1	16.22	16.13	16.15	16.52	16.53	16.31
N=2	16.18	17.72	17.82	17.90	17.90	17.50
N=3	15.53	17.15	17.68	18.00	18.08	17.29
N=4	15.21	17.05	17.48	17.96	17.91	17.12
N=5	15.83	17.38	17.82	18.15	18.12	17.46
$N \in \{1, 2\}$	16.30	17.73	17.51	17.73	17.67	17.39
$N \in \{1, 2, 3\}$	16.19	17.63	18.01	18.06	18.03	17.58
$\mathbf{N \in \{1, 2, 3, 4\}}$ (WEAVE)	**16.91**	**17.84**	**18.17**	**18.29**	**18.35**	**17.91**
$N \in \{1, 2, 3, 4, 5\}$	15.88	17.05	17.26	17.71	17.93	17.17

Sampling $N \in \{1, \ldots, 4\}$ during training achieves the highest average $\text{PSNR}_{\text{mask}}$ (17.91 dB) across one to five inference references. In contrast, fixed-count training is less robust when the number of available references changes: the strongest fixed-count setting, $N = 2$, reaches an average of 17.50 dB, 0.41 dB below our variable-count training scheme. The $N \in \{1, \ldots, 4\}$ model also outperforms every alternative training configuration at all five evaluated reference counts. Under this training scheme, performance increases from 16.91 dB with one reference to 17.84, 18.17, and 18.29 dB with two, three, and four references, respectively. Notably, the same model reaches 18.35 dB with five references, despite never observing five-reference examples during training. Extending the training range to $N \in \{1, \ldots, 5\}$ does not improve this generalization and instead lowers the average to 17.17 dB. These results support variable-count training with $N \in \{1, \ldots, 4\}$ as a robust choice across different numbers of available reference views.

Additionally, to understand the similar performance between four- and five-reference evaluation settings, we conduct a paired inference analysis on the six RealBench scenes with five references to assess the contribution of the fifth view. We hold the WEAVE model, target, first four ordered references, mask, and sampling seed fixed, then add only the fifth reference. Table 7 shows that the fifth reference increases warp-union coverage by 1.86 percentage points (0.9795 → 0.9981), but changes mean $\text{PSNR}_{\text{mask}}$ by only +0.002 dB. This average obscures heterogeneous scene-level effects, ranging from −1.09 to +0.69 dB. Thus, the fifth view is not uniformly redundant: it can provide additional geometric support without improving the *average* reconstruction quality. This pattern is consistent with additional views sometimes providing complementary evidence and sometimes introducing conflicting or low-quality evidence.

Table 7: **Paired fifth-reference diagnostic.**

	$\text{PSNR}_{\text{mask}}$ ↑		Warp coverage ↑	
Scene	N=4	Δ_{5-4}	N=4	Δ_{5-4}
0	14.54	+0.34	1.0000	+0.0000
4	16.75	+0.69	1.0000	+0.0000
6	14.10	+0.02	1.0000	+0.0000
9	17.61	-1.09	0.9704	+0.0217
10	20.43	-0.22	0.9064	+0.0898
17	21.37	+0.26	1.0000	+0.0000
Mean	17.47	**+0.002**	0.9795	**+0.0186**

Together, these results support variable-reference completion as a setting in which complementary views can improve reconstruction, while also showing that additional references are not uniformly beneficial. Training with variable reference counts allows WEAVE to exploit this evidence without assuming a fixed number of views or requiring retraining when additional references are available.

C.2 Attention Analysis

We examine how WEAVE attends to reference content and whether the spatial distribution of this attention affects completion quality.

Attention as co-located lookup. To examine where the model reads within R-Warps, we capture all 25 MM-DiT blocks over all 28 denoising steps on every RealBench scene. For each block, masked query token, and R-Warp, we normalize attention over that warp's tokens, then measure its cumulative mass inside a Euclidean-radius disk around the co-located token in the scene's native $H_{\text{tok}} \times W_{\text{tok}}$ MM-DiT token grid. We compute locality directly on this grid, without

resampling attention maps. For RealBench, all scenes yield 32×32 token grids under the standard preprocessing, corresponding to approximately 16×16 image pixels per token. We average the resulting CDFs across blocks and scenes.

More precisely, for a given scene, block, and denoising step, let a_{ijk} denote attention from masked query token i to token k in R-Warp $\mathbf{W}_j$, whose token-index set is $\mathcal{T}_j^{\text{warp}}$. We suppress scene, block, and timestep indices for clarity. We normalize attention within each R-Warp:

$$p_{ijk} = \frac{a_{ijk}}{\sum_{k' \in \mathcal{T}_j^{\text{warp}}} a_{ijk'}}, \qquad k \in \mathcal{T}_j^{\text{warp}}.$$

Let $g(\cdot)$ map query and R-Warp token indices to their two-dimensional spatial coordinates on the scene's native $H_{\text{tok}} \times W_{\text{tok}}$ target-view grid. We compute the locality CDF:

$$L(\rho) = \mathbb{E}_{i,j}\left[\sum_{k \in \mathcal{T}_j^{\text{warp}}} p_{ijk}\, \mathbb{1}\{\|g(k) - g(i)\|_2 \le \rho\}\right],$$

where the expectation averages uniformly over masked query tokens and reference views. We average these CDFs over denoising steps, blocks, and scenes. For comparison, the uniform baseline uses the same grids and query-token sets, replacing p_{ijk} with $1/|\mathcal{T}_j^{\text{warp}}|$.

As shown in Fig. 7, hole queries place 19.5% of their within-R-Warp attention on the exact co-located token, 38.8% within one token, and 49.9% within two. This is not explained by the larger neighborhood: uniform attention would place only 1.2% of its mass within radius two. The same pattern holds over all target queries (41.6% within two), but is stronger in the masked region. Thus, when synthesizing missing content, the model uses each R-Warp as a spatially grounded lookup table, concentrating its reference reads near the geometrically corresponding location rather than searching diffusely across the view.

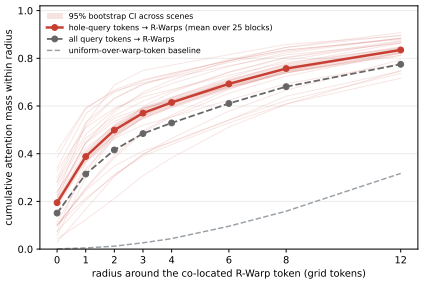

Figure 7: **R-Warp attention is spatially concentrated around the aligned token.**

Concentrated reference attention supports completion. We also test whether completion depends on where the model reads within its reference inputs, rather than simply on how much attention those inputs receive. For each hole token at inference, we apply a temperature transform to attention over the combined reference-token set $\mathcal{R}$ (raw references and R-Warps), while preserving the total attention mass assigned to that set. Let a_{ij} denote the original attention weight from hole-query token i to reference token $j \in \mathcal{R}$, and let $m_i = \sum_{j \in \mathcal{R}} a_{ij}$ be the query's total reference-attention mass. We replace only the within-reference distribution by

$$\tilde{a}_{ij}^{(T)} = m_i \frac{a_{ij}^{1/T}}{\sum_{k \in \mathcal{R}} a_{ik}^{1/T}}, \qquad j \in \mathcal{R},$$

leaving attention to all $j \notin \mathcal{R}$ unchanged. Thus, the intervention changes concentration without changing how much total attention the query assigns to references. $T = 1$ leaves the learned distribution unchanged; $T < 1$ sharpens it and $T > 1$ diffuses it.

As shown in Fig. 8, on all RealBench scenes, diffusing reference attention with $T = 2$ reduces hole PSNR by 4.94 dB relative to the learned distribution. Applying the same $T = 2$ transform to text tokens costs only 0.26 dB. Thus, completion quality depends strongly on the learned concentration of attention over reference tokens, rather than reflecting a generic sensitivity to attention perturbation.

Together, these results support WEAVE's use of R-Warps: they place reference content in target coordinates, and the generator uses spatially concentrated attention to access this aligned evidence.

C.3 REFERENCE-ORDER SENSITIVITY

We evaluate whether reconstruction quality remains stable when the reference images are reordered, comparing both backbones with and without warp conditioning. For each RealBench scene with

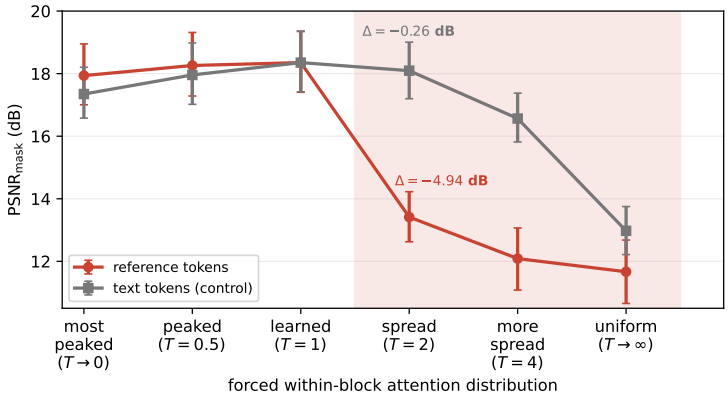

Figure 8: **Effect of reshaping within-block attention.** Relative to learned attention, $T = 2$ softening reduces hole PSNR by 4.94 dB for reference tokens vs. 0.26 dB for the matched text-token control. Error bars are 95% paired bootstrap CIs over RealBench scenes.

Table 9: **Backbone scaling from 4B to 9B.** The Δ row reports the paired 9B−4B difference. Green indicates an interval excluding zero and red indicates an interval overlapping zero.

Model	Parameters	MegaDepth PSNR↑	RealBench PSNR$_{\text{mask}}$ ↑
WEAVE	4B	27.49 [27.18, 27.80]	18.35 [17.42, 19.37]
	9B	**27.94** [27.63, 28.25]	**18.53** [17.51, 19.61]
Δ (9B–4B; 95% CI)		+0.45 [+0.39, +0.51]	+0.18 [−0.21, +0.52]

at least three references, we generate completions under $K = 8$ random reference orderings while holding the reference set and initial noise fixed. Within the masked region, we measure *output agreement*, the mean pairwise PSNR between completions from different orderings, and *quality variation*, the standard deviation of PSNR against ground truth across orderings. Higher output agreement indicates more similar predictions, while lower quality variation indicates more consistent reconstruction accuracy. Both statistics are averaged across scenes.

Table 8: **Sensitivity to reference order.** Output agreement measures pairwise PSNR$_{\text{mask}}$ across eight orderings; quality variation is the standard deviation of PSNR$_{\text{mask}}$ against ground truth.

Model	Output agreement ↑	Quality variation ↓
FLUX, no warps	24.5	0.201
WEAVE	28.6	0.163
SD2 Inpaint, no warps	64.3	0.001
WEAVE$_{\text{[UN]}}$	32.1	0.037

Table 8 shows that both variants maintain consistent reconstruction quality across the tested orderings, with average quality variation of 0.163 dB for WEAVE and 0.037 dB for WEAVE$_{\text{[UN]}}$. For WEAVE, warp conditioning also increases output agreement from 24.5 to 28.6 dB and reduces quality variation from 0.201 to 0.163 dB. For WEAVE$_{\text{[UN]}}$, the no-warp control produces nearly identical outputs across orderings; adding warps decreases output agreement from 64.3 to 32.1 dB and increases quality variation from 0.001 to 0.037 dB. Although reordering can therefore change the generated content, its effect on reconstruction quality remains small on average for both variants, supporting their robustness to reference order under the tested conditions.

C.4 BACKBONE SCALING

We examine whether increasing generator capacity improves WEAVE by adapting it to the 9B FLUX.2 [klein] backbone, which has $2.25\times$ as many parameters as the primary 4B model. Both variants use the same Q-Warp and R-Warps, MegaDepth training data, 24,000 training steps, cosine schedule, effective batch size of 8, DeepSpeed ZeRO-2, and training seed. For the 9B model, we reduce the learning rate from 3×10^{-5} to 2×10^{-5} for training stability. As shown in Table 9, scaling to 9B improves MegaDepth PSNR from 27.49 to 27.94 dB, a paired gain of 0.45 dB (95% CI:

Algorithm 1 Alternative geometry-conditioning implementations.
d_{epi} is the pose/intrinsics-derived distance from a target token to a reference token's epipolar line.
g_{epi} and g_b are learnable scalar gates initialized to 0; ZeroProj is a zero-initialized output projection.

Require: $\mathbf{H}_{\text{img}}$, VGGT features $\mathbf{F}$, and, where applicable, warps $\{\mathbf{W}_j, \mathbf{C}_j\}_{j=1}^{N}$

1: **Epipolar bias**
For each target–reference token pair (i, k), compute $B_{ik} \leftarrow g_{\text{epi}} \max\{-d_{\text{epi}}(i,k)^2/(2\sigma^2), b_{\min}\}$.
Add $\mathbf{B}$ to the joint-attention logits: $\mathbf{A} \leftarrow \mathrm{softmax}(\mathbf{Q}\mathbf{K}^\top/\sqrt{d} + \mathbf{B})$.

2: **Feature injection**
$\widetilde{\mathbf{F}} \leftarrow \mathrm{ResizeToFLUXGrids}(\mathbf{F})$; set query-hole rows of $\widetilde{\mathbf{F}}$ to zero to avoid target leakage.
$\mathbf{H}_{\text{img}} \leftarrow \mathbf{H}_{\text{img}} + \mathrm{ZeroProj}(\widetilde{\mathbf{F}})$.

3: **Dedicated cross-attention**
$\Delta \leftarrow \mathrm{MHA}(\mathbf{W}_Q\mathbf{H}_{\text{img}}, \mathbf{W}_K\widetilde{\mathbf{F}}, \mathbf{W}_V\widetilde{\mathbf{F}}; g_b\mathbf{B}')$, where $g_b\mathbf{B}'$ is an optional logit bias.
$\mathbf{H}_{\text{img}} \leftarrow \mathbf{H}_{\text{img}} + \mathrm{ZeroProj}(\Delta)$.

4: **Warp channel concatenation**
VAE-encode every warp and concatenate channels via the VAE $\mathcal{E}$, with zero-padding for $N < 5$:
$\mathbf{U} \leftarrow [\mathcal{E}(\bar{\mathbf{W}}); \mathcal{E}(\mathbf{W}_1); \ldots; \mathcal{E}(\mathbf{W}_N)]$.
Pointwise: $\mathrm{Tok}(\mathbf{z}_\tau) \leftarrow \mathrm{Tok}(\mathbf{z}_\tau) + \mathrm{ZeroProj}(\mathbf{U})$.
Spatial conv: $\mathrm{Tok}(\mathbf{z}_\tau) \leftarrow \mathrm{Tok}(\mathbf{z}_\tau) + \mathrm{Zero}_{1\times1}(\mathrm{SiLU}(\mathrm{PW}_{1\times1}(\mathrm{DW}_{7\times7}(\mathbf{U}))))$.

5: **Separate warp token slots (WEAVE)**
$\mathbf{H} \leftarrow \left[\mathrm{Tok}(\mathbf{z}_\tau) \;\middle|\; \mathbf{s}^{\mathrm{q}} \;\middle|\; \mathbf{s}_1^{\mathrm{ref}}, \ldots, \mathbf{s}_N^{\mathrm{ref}} \;\middle|\; \mathbf{s}_1^{\mathrm{warp}}, \ldots, \mathbf{s}_N^{\mathrm{warp}}\right]$; apply MM-DiT joint attention to $\mathbf{H}$.

$[+0.39, +0.51]$). On RealBench, however, the improvement is only 0.18 dB, from 18.35 to 18.53 dB, with a 95% confidence interval of $[-0.21, +0.52]$ that overlaps zero. Thus, increasing backbone capacity yields a consistent in-domain gain but no clear improvement in out-of-distribution RealBench performance. We therefore use the 4B backbone as our primary model, as it achieves comparable RealBench reconstruction quality with less than half the generator parameters.

D ALTERNATIVE GEOMETRY CONDITIONING

We compare several ways of incorporating VGGT-Ω geometry into the FLUX.2 [klein] backbone. All configurations follow the same training protocol as WEAVE: fine-tuning for 24,000 steps at a learning rate of 3×10^{-5}, with identical evaluation settings. These experiments examine whether geometry features or attention constraints can provide the same benefits as reference warping.

Conditioning mechanisms. We evaluate three alternatives to warping: *epipolar bias*, which adds a geometry-derived bias to target–reference attention logits; *feature injection*, which projects VGGT-Ω's 2048-dimensional patch features and adds them to the backbone's image tokens; and *cross-attention*, which introduces a dedicated block using these features as keys and values. The added branches use zero-initialized gates or output projections to preserve the base model's initial predictions. We also compare two ways of supplying warped-image latents: concatenating them along the input channels, using either a pointwise projection or a 7×7 depthwise-separable convolution, and retaining them as separate token blocks as in WEAVE.

Table 10: **Alternative geometry conditioning.**

Conditioning	MegaDepth PSNR ↑	RealBench $\text{PSNR}_{\text{mask}}$ ↑
Raw references only	26.74	15.99
VGGT feature injection	26.77	16.00
VGGT cross-attention	26.75	16.08
Epipolar attention bias	26.89	16.80
Warps: channel concatenation	26.96	17.04
Warps: channel concatenation + conv.	26.92	16.63
Warps: separate tokens (WEAVE)	**27.49**	**18.35**

Algorithm 1 elucidates these mechanisms. Let $\mathbf{H}_{\text{img}} = \left[\mathrm{Tok}(\mathbf{z}_\tau) \;\middle|\; \mathbf{s}^{\mathrm{m}} \;\middle|\; \mathbf{s}_1^{\mathrm{ref}}, \ldots, \mathbf{s}_N^{\mathrm{ref}}\right]$ denote the MM-DiT's image-token rows (noisy target, blacked target, and raw references) where we assign $\mathbf{s}^m = \mathrm{Tok}\left(\mathcal{E}(\mathbf{I}_0^m)\right)$, and let $\mathbf{F}$ be VGGT-Ω's frozen 2048-D patch features, bilinearly resampled to the MM-DiT token grids. Thus, the variants that ingest VGGT features never receive warped pixels: they inject only $\mathbf{F}$. Conversely, channel variants receive warped image *latents*, but collapse the references into a fixed channel stack before the MM-DiT. Our implementation keeps each warped reference as its own sequence block, which permits subsequent joint attention to choose content from a local neighborhood when the warp is imperfect.

Table 12: **Zero-shot performance on the RealEstate10K Small Set** (Zhou et al., 2021).

Method	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓	CLIP↑	DINO↑
Qwen Image Edit 2511 Wu et al. (2025)	22.08	0.913	0.134	0.234	0.845	0.880
FLUX.2 [klein] 9B (Black Forest Labs, 2025)	27.92	0.943	0.051	0.042	0.960	0.973
FLUX.2 [klein] 4B (Black Forest Labs, 2025)	28.73	0.947	0.041	0.034	0.968	0.976
FLUX.1 Fill [dev] 12B (Black Forest Labs, 2024)	29.48	0.956	0.042	0.022	0.967	0.982
FLUX.2 [dev] 32B (Black Forest Labs, 2025)	30.45	0.956	0.042	0.055	0.963	0.964
Nano Banana Pro (Gemini Team, 2023)	32.18	0.959	0.040	0.033	0.972	0.979
GPT Image 2 (OpenAI, 2026)	33.79	0.971	0.026	0.021	0.972	0.990
ProFill (Zeng et al., 2020)	30.95	0.969	0.042	—	—	—
APAP (Zaragoza et al., 2013)	31.94	0.974	0.025	—	—	—
OPN (Oh et al., 2019)	33.40	0.977	0.021	—	—	—
DFG (Xu et al., 2019)	36.17	0.987	0.016	—	—	—
TransFill (Zhou et al., 2021)	38.83	0.988	0.013	0.004	**0.994**	**0.997**
WEAVE	**38.94**	**0.989**	**0.007**	**0.003**	0.993	**0.997**

Comparison. Table 10 shows that direct VGGT feature conditioning changes RealBench $\text{PSNR}_{\text{mask}}$ by < 0.1 dB relative to the raw-reference base, while epipolar bias improves it by 0.81 dB. Warped-image tokens (WEAVE) provide largest gains of 2.36 dB on RealBench and 0.75 dB on MegaDepth. For settings that ingest warps, channel concatenation improves RealBench $\text{PSNR}_{\text{mask}}$ by 1.05 dB vs. 2.36 dB for WEAVE; adding a convolution does not close this gap.

Inference-time component diagnostics. To test whether the trained models use the auxiliary conditioning components at inference, we ablate each component while holding all learned backbone weights fixed. For additive components, we additionally measure the relative residual magnitude $\rho = \|\Delta h\|_2 / \|h\|_2$, where Δh is the component output added to hidden state h. We average this quantity over the 33 RealBench scenes and 28 denoising steps. Removing either VGGT feature injection or VGGT cross-attention changes PSNR by at most 0.006 dB, indicating limited inference-time reliance on these components under this training setup. Their mean relative residual magnitudes are likewise small: 2.0% and 0.2%, respectively. Although training with epipolar bias improves PSNR by 0.81 dB relative to the base model, disabling that bias at inference decreases PSNR by only 0.037 dB. Coupled with the fact that this setting underperforms warped tokens (WEAVE), we omit it from the final pipeline without materially affecting performance.

Table 11: **Inference-time reliance on alternative conditioning.** RealBench $\text{PSNR}_{\text{mask}}$ with each learned mechanism enabled or disabled after training, together with its mean relative residual norm ρ. Epipolar bias modifies attention logits directly and therefore has no residual norm.

Mechanism	Enabled	Disabled	Δ(Disabled − Enabled)	Mean ρ
Feature injection	15.998	15.995	−0.004	2.00%
Cross-attention	16.083	16.077	−0.006	0.20%
Epipolar bias	16.804	16.766	−0.037	—

Together, these results motivate WEAVE's use of warped reference images as separate token blocks: warping provides spatial guidance through the generator's native image representation, while attention combines evidence across views. Under the same training protocol, our design outperforms the tested feature-injection, attention-bias, and channel-concatenation alternatives.

E ADDITIONAL BENCHMARK RESULTS

We evaluate on the RealEstate10K Small Set (Zhou et al., 2021), a single-reference benchmark of 300 test pairs at 256×256 resolution, with masks covering 11.5% of the target image. Reference and target frames are sampled only ten video frames apart, yielding much smaller viewpoint changes than those in MegaDepth and RealBench. This near-aligned setting favors earlier pixel-transfer methods, which can recover missing content by warping observed pixels from a nearby reference.

As shown in Table 12, WEAVE achieves the highest PSNR (38.94 dB) and SSIM (0.989), outperforming TransFill (38.83 dB and 0.988). It also achieves the lowest reported perceptual distances, reducing LPIPS from TransFill's 0.013 to 0.007 and DreamSim from 0.004 to 0.003. Among the general-purpose models, GPT Image 2 obtains the highest PSNR at 33.79 dB, compared with WEAVE's 38.94 dB. These results demonstrate that WEAVE performs well in narrow-baseline completion, including against specialized reference-alignment methods.

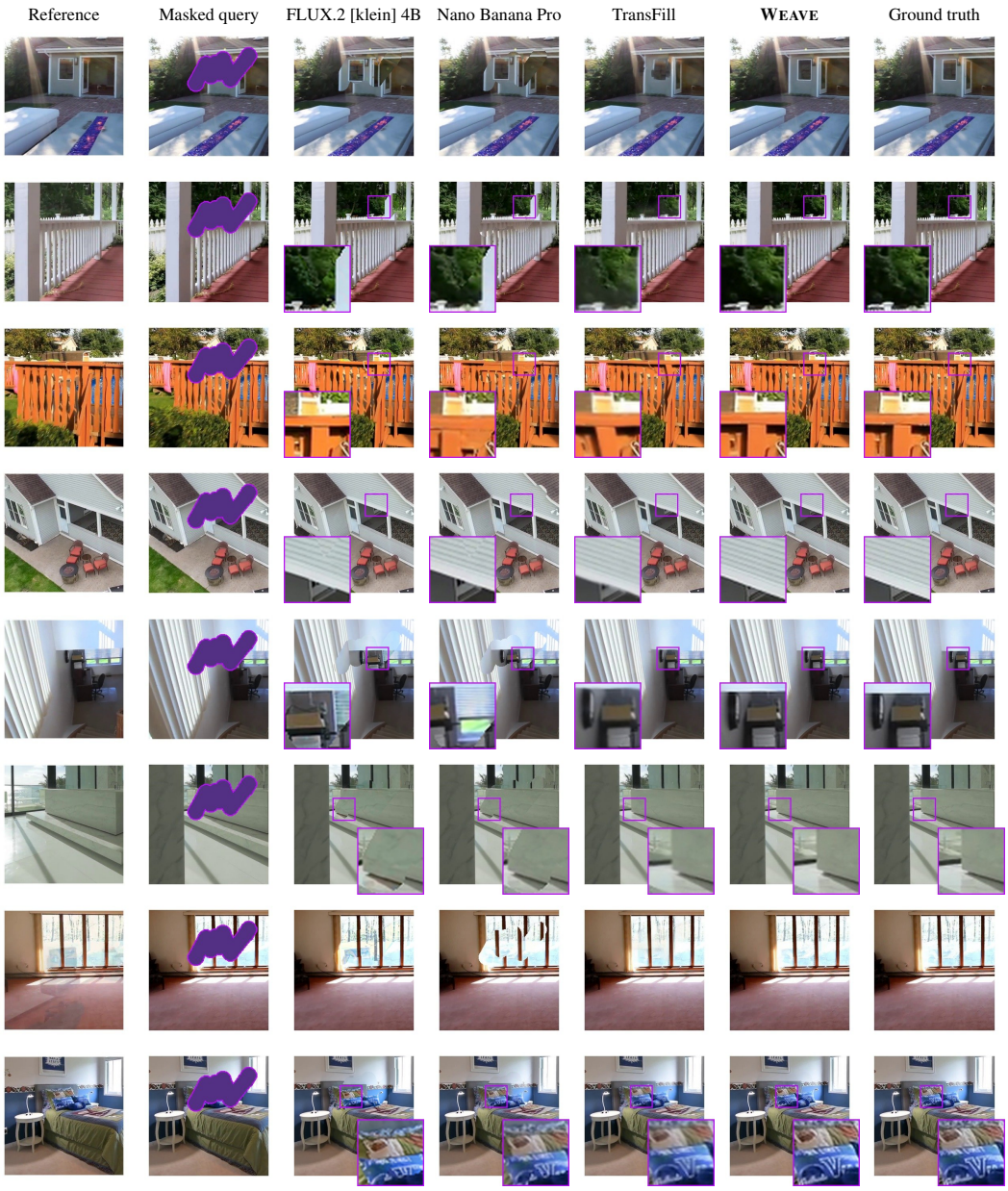

Figure 9: **Qualitative single-reference image completion on the RealEstate10K Small Set (Zhou et al., 2021).** Given the reference and masked query, our method reconstructs the masked region faithfully; TransFill blurs structured surfaces and Nano Banana Pro hallucinates.

Fig. 9 illustrates these differences qualitatively. FLUX.2 [klein] 4B and Nano Banana Pro often introduce incorrect content, including altered furniture and window details (rows 5 and 7). TransFill generally preserves the scene layout, but misaligns the house facade in the first example and blurs fine details in several others. WEAVE more closely matches the ground truth, preserving structural alignment while recovering sharper foliage, siding, and bedding textures.

F ADDITIONAL QUALITATIVE EXAMPLES

Figs. 10, 11, and 12 compare Nano Banana Pro, WEAVE$_{[UN]}$, and WEAVE on additional examples. Each row contains four references, a masked query, and enlarged crops (purple boxes) of the synthesized region. Nano Banana Pro frequently appears to copy visually similar content from a reference

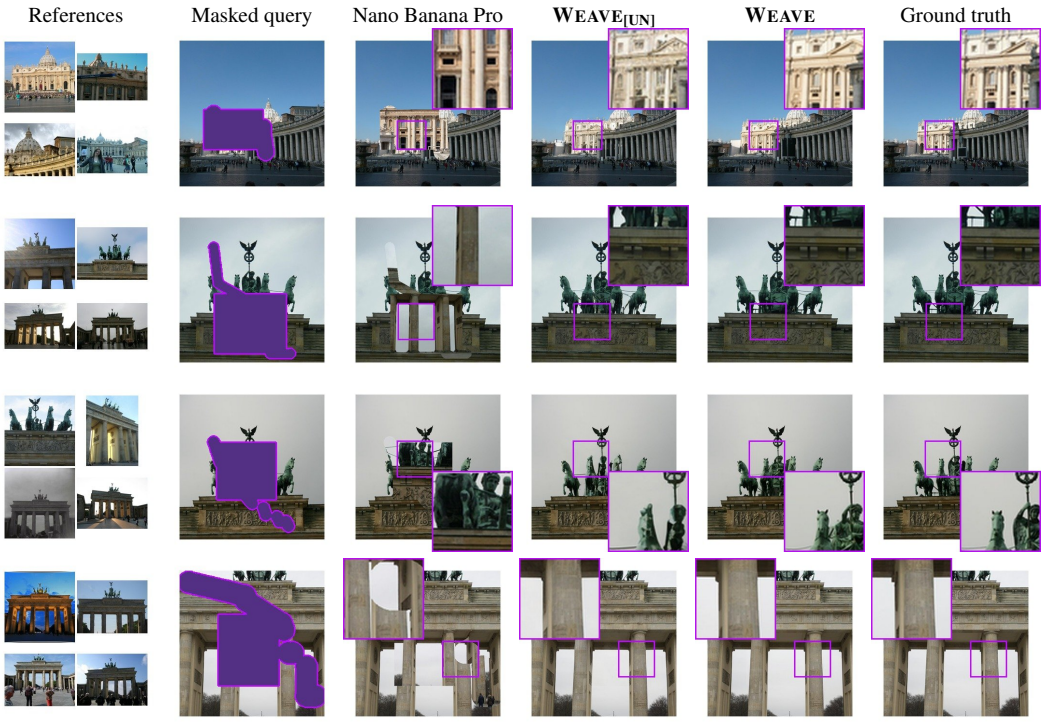

Figure 10: **Qualitative multi-reference completion on MegaDepth (4 references).**

view without fully transforming it into the query view. Consequently, its completions often have incorrect perspective, scale, or placement and can be visibly inconsistent with the surrounding target image. In contrast, WEAVE$_{[UN]}$ uses the multi-view evidence to produce target-frame-consistent structure: it recovers the appropriate architectural layout, including arches, columns, façade reliefs, and repeated elements, across the diverse landmark scenes. WEAVE retains these geometric benefits while producing the sharpest and most detailed completions, particularly for fine repeated texture and thin architectural boundaries. Thus, WEAVE$_{[UN]}$ demonstrates that warp conditioning improves multi-view completion even with the SD2-Inpainting backbone, while the stronger WEAVE backbone further improves the fidelity of the reconstructed detail.

G LIMITATIONS

WEAVE can produce visually convincing completions while missing precise spatial alignment or fine detail. The displacement examples in Fig. 13 preserve the main structures and appearance but shift some elements relative to the ground truth, reducing pixelwise accuracy. Thin structures can also be simplified or reconstructed incorrectly. Large missing regions present a further challenge: the model may generate plausible content that does not match the true scene layout. The outpainting example in Fig. 13(d) illustrates this distinction. The completion preserves the target's monochrome appearance despite receiving color references, but reconstructs some scene elements incorrectly. Thus, visual coherence does not always imply pixel-perfect recovery of the underlying scene, especially when the ground truth contains elements that cannot be deduced from the given context and references (*e.g.*, the shape and position of the clouds).

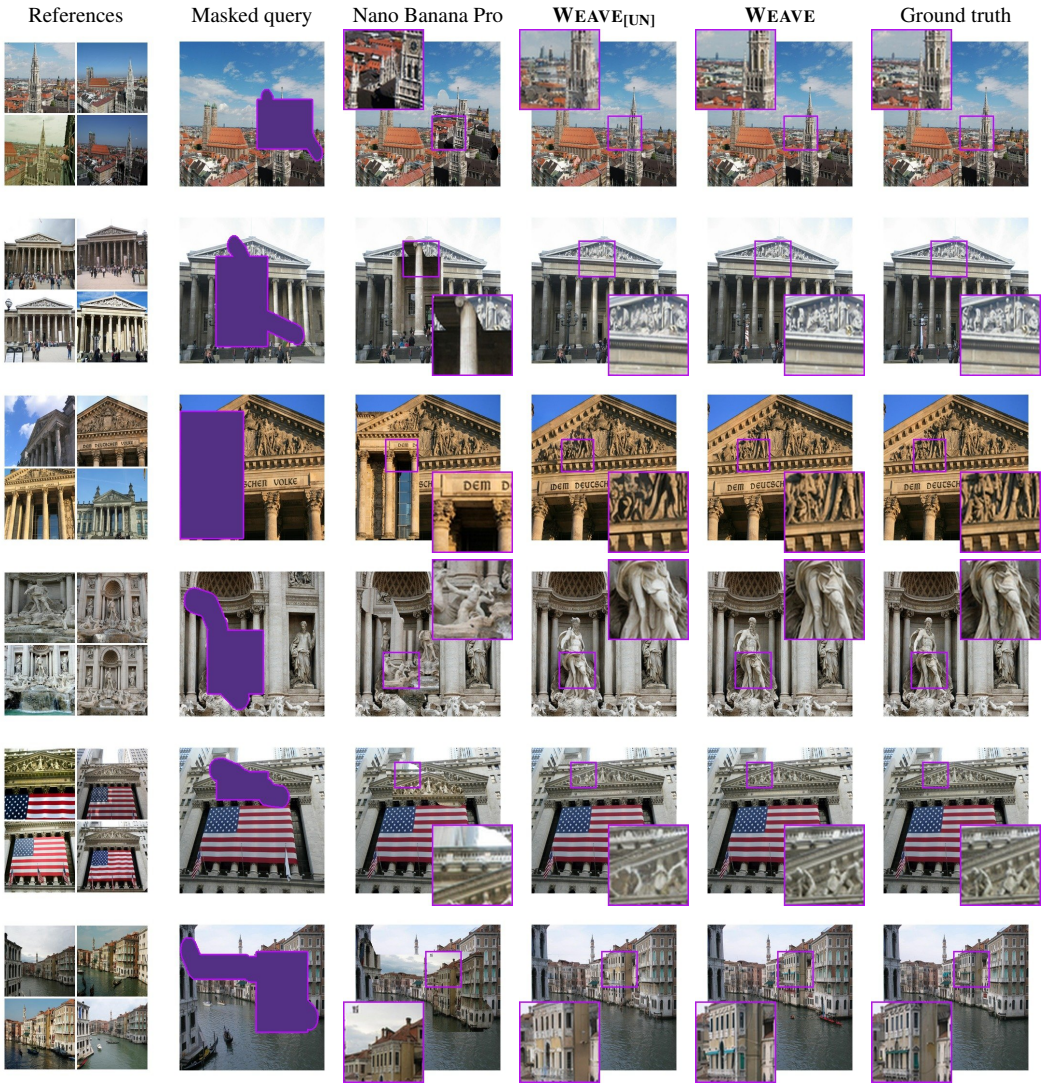

Figure 11: **Qualitative multi-reference completion on diverse MegaDepth landmarks**, part 1.

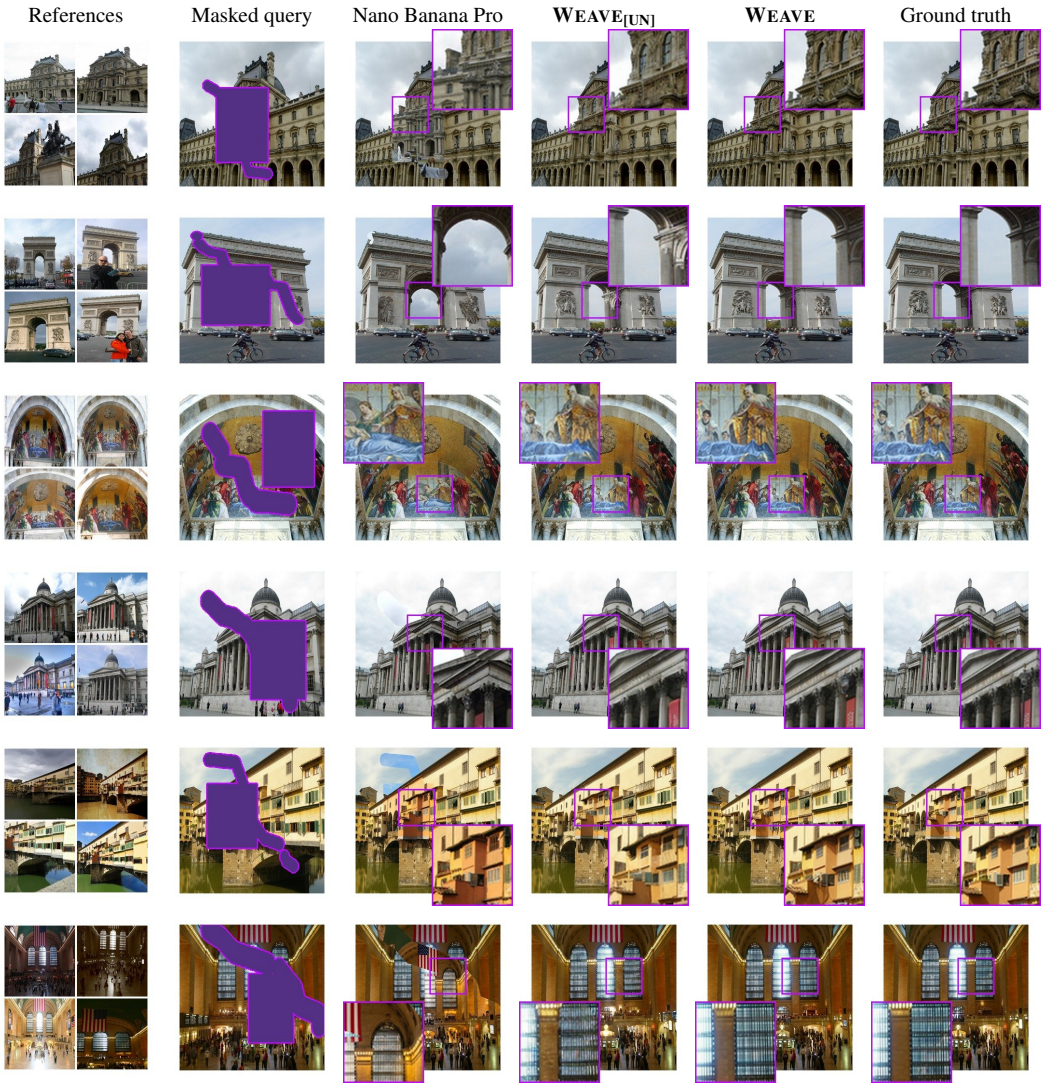

Figure 12: **Qualitative multi-reference completion on diverse MegaDepth landmarks**, part 2.



Figure 13: **Qualitative limitations. (a, b)** The main structures and appearance are preserved, but small spatial offsets cause misalignment with ground truth. **(c)** The extended façade is coherent, but its upper storey has incorrect lighting. **(d)** The completion retains the target's monochrome appearance despite color references, but reconstructs parts of the background incorrectly.

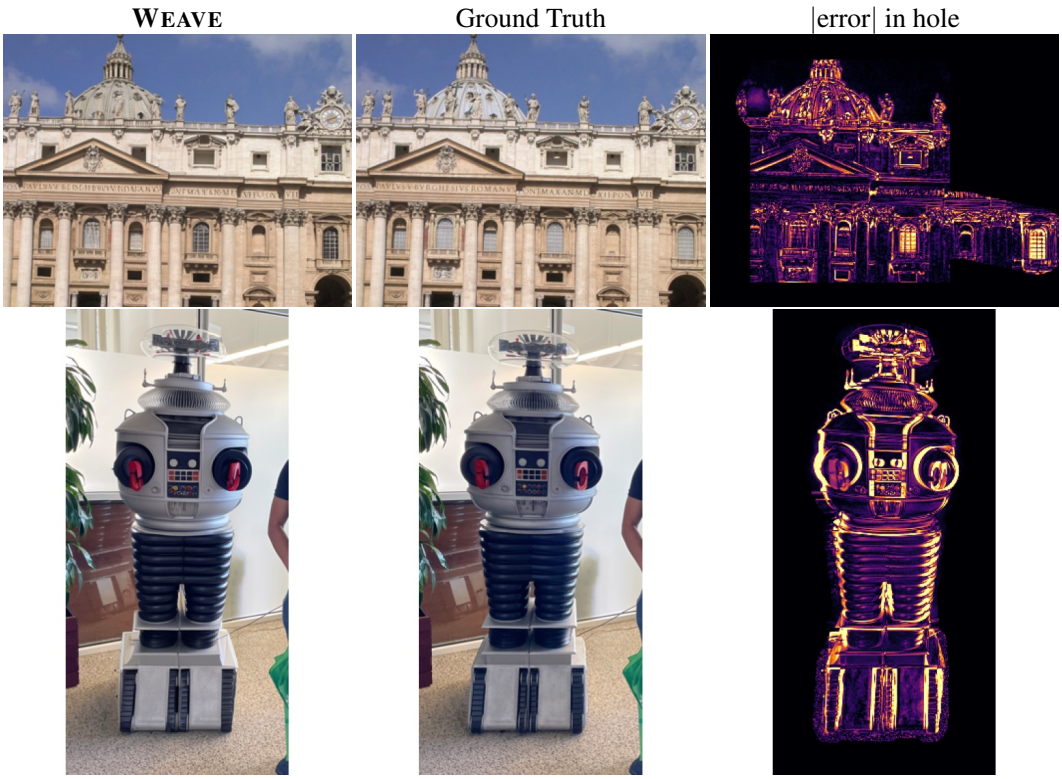

Figure 14: **Small spatial offsets reduce pixelwise accuracy.** Magnified crops of the façade (top) and figurine (bottom) from Fig. 13(a,b). The completions closely resemble the ground truth, but small misalignments produce errors along architectural edges, the figurine's silhouette, and panel boundaries. The right column shows absolute pixelwise error within the masked region.