

LaRP: Efficient Multi-View Inpainting with Latent Reprojection Priors

Supplementary Material

A. Video Results

We compare the NVS results on the testing views of the SPIn-NeRF [44] dataset and include the video results in the file “video.mp4”.

B. Sample Data

We show sample two-view image pairs generated by applying our pipeline to the “chair”, “cereal_box”, “book”, and “shoe” categories of the Objectron dataset in Fig. 7.

C. Comparison with a ControlNet + Latent Reprojection Baseline?

We have discussed the core conceptual differences between the proposed LaRP architecture and a standard ControlNet formulation (Sec. 3.1 and Fig. 3), where one must *first* reproject the reference image before feeding it to the ControlNet. One might question the feasibility of applying our proposed latent reprojection scheme to a standard ControlNet. However, we note that the standard ControlNet is incompatible with such a deferred reprojection strategy. As illustrated in Fig. 3 (right), a ControlNet formulation incorporates the noisy latent of the target image into the ControlNet branch, inherently grounding the branch in the target’s spatial coordinate system. Reprojecting a UNet latent derived from this input would result in a geometrically incoherent control signal, which is unsuitable for precise spatial conditioning.

D. Architectural Details

In this section, we detail the alternative architectures for cross-view conditioned inpainting used in our ablation study. For the standard ControlNet formulation, please refer to Sec. 3.1 and Fig. 3 of the main paper. Qualitative inpainting results for these alternative architectures are shown in Fig. 10.

Cross Attention. Inspired by attention-based conditional image generation and editing [23, 77], one of our alternative architectures uses cross attention to inject the learned representations from the 3D foundation model into the diffusion UNet. Specifically, we adapt the decoupled cross-attention design used by IP-Adapter [77] by performing cross attention between the denoised image tokens and the reference view’s image tokens processed by VGGT’s alternating-attention [67] (*i.e.*, the tokens before being fed into VGGT’s DPT head). Additionally, we initialize the

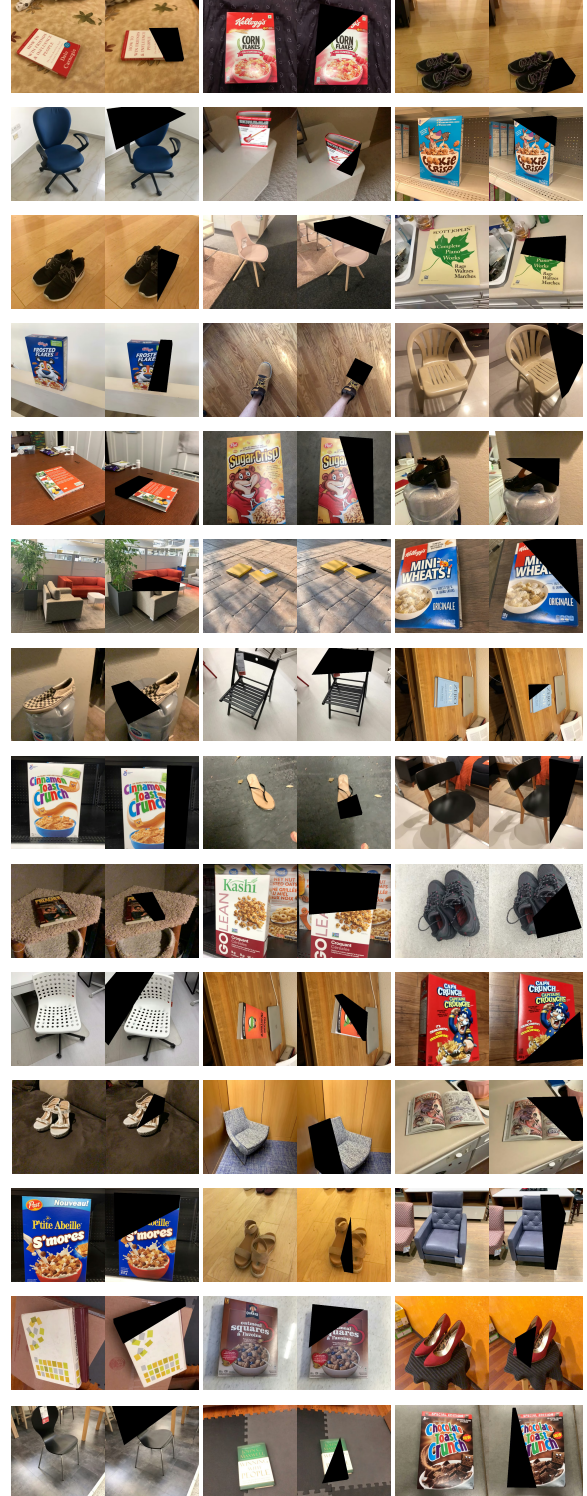


Figure 7. Sample image pairs generated by our pipeline.

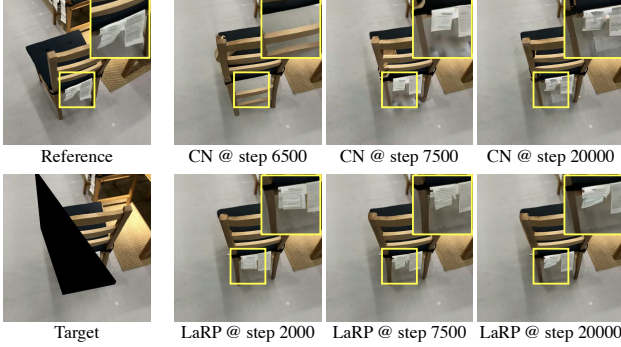


Figure 8. **Sudden convergence comparison.** LaRP exhibits the “sudden convergence” phenomenon much earlier in training than the ControlNet (CN) baseline.

output linear layer of each attention module to zero to ensure stable training. As shown in Fig. 10 and Tab. 5 (b), this attempt failed to yield effective cross-view conditioning even after extended training. We attribute this failure to the significant modality gap between the VGGT features (explicit 3D representations) and the CLIP text embeddings that the diffusion UNet is originally optimized to interpret. Furthermore, while IP-Adapter initializes its new attention weights from pre-trained layers to ease optimization, our module must be initialized from scratch, as there is no existing adapter mapping VGGT’s internal representations to the diffusion UNet’s latent space.

Unlocking UNet’s Decoder. We explored a variant where we unfreeze the parameters of the main UNet decoder, which receives the reprojected reference latents. However, we observed that this approach distorts the generative priors of the denoising UNet (Fig. 10 and Tab. 5 (c)). We attribute this to the significant discrepancy in data diversity between the pretraining stage and our proposed cross-view training stage. The diffusion UNet possesses robust generative priors derived from the extremely diverse data used during its pretraining. In contrast, our training dataset, constructed to facilitate cross-view conditioning, is significantly more limited in semantic diversity. Unfreezing the decoder allows this limited diversity to overfit the weights, thereby compromising the model’s pretrained knowledge and generative quality. Therefore, in our proposed LaRP architecture, we freeze the original UNet and only train the copied UNet encoder.

E. Sudden Convergence

The *sudden convergence* phenomenon is documented in [82], which states that “*the model does not gradually learn the control conditions, but abruptly succeeds in following the input conditioning image*”. We observe that the proposed LaRP architecture demonstrates significantly faster sudden convergence during training compared to a standard

Table 6. **Percentile of FPS view pair selection.** Values of the MET3R metric on the inpaintings are reported. Lower is better.

FPS percentile	5%–10%	10%–30% (Ours)	30%–50%
MET3R _M ↓	0.1325	0.1109	0.1144
MET3R _R ↓	0.1438	0.1293	0.1328

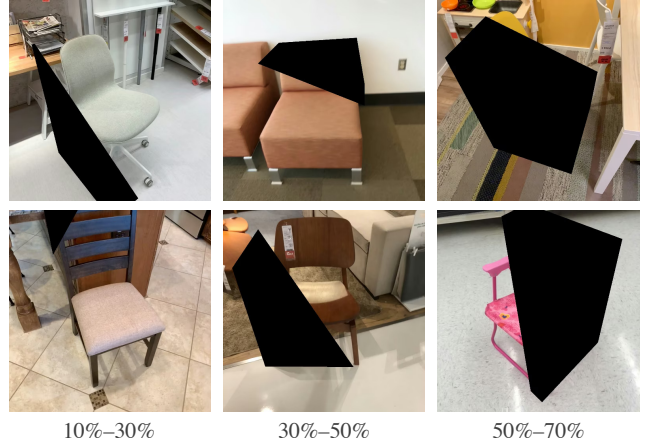


Figure 9. **Area of the 3D-aware masks.** We limit the sliced mask to approximately 30%–50% of the projected object bounding box area to ensure a balance between inpainting difficulty and available visual context.

ControlNet formulation on the task of cross-view conditioned inpainting, as shown in Fig. 8.

F. More Ablation Results

F.1. Hyperparameters of the Data Pipeline.

We design our data pipeline to generate high-quality two-view image pairs that facilitate stable training. This involves carefully tuning two key hyperparameters: the camera baseline for view selection and the mask area for occlusion simulation. These design choices ensure that the model receives a control signal that is sufficiently diverse and representative of real-world inference scenarios, while maintaining a balance between task difficulty and available visual context.

FPS Percentile. After performing FPS to select a subset of frames, we generate all possible pairs and sort them by camera distance. We then sample pairs from the 10th to 30th percentile of this sorted pool. Tab. 6 shows that reducing the pairwise distance (5%–10%) of the generated training pairs hinders the final performance of the cross-view conditioned inpainting model. This observation is consistent with our single-view dataset ablation in Tab. 5 (d), where the effective baseline is zero and performance is significantly degraded. On the other hand, increasing the pairwise distance (30%–50%) also reduces performance, as larger viewpoint differences lead to reduced overlap and sparser reprojected

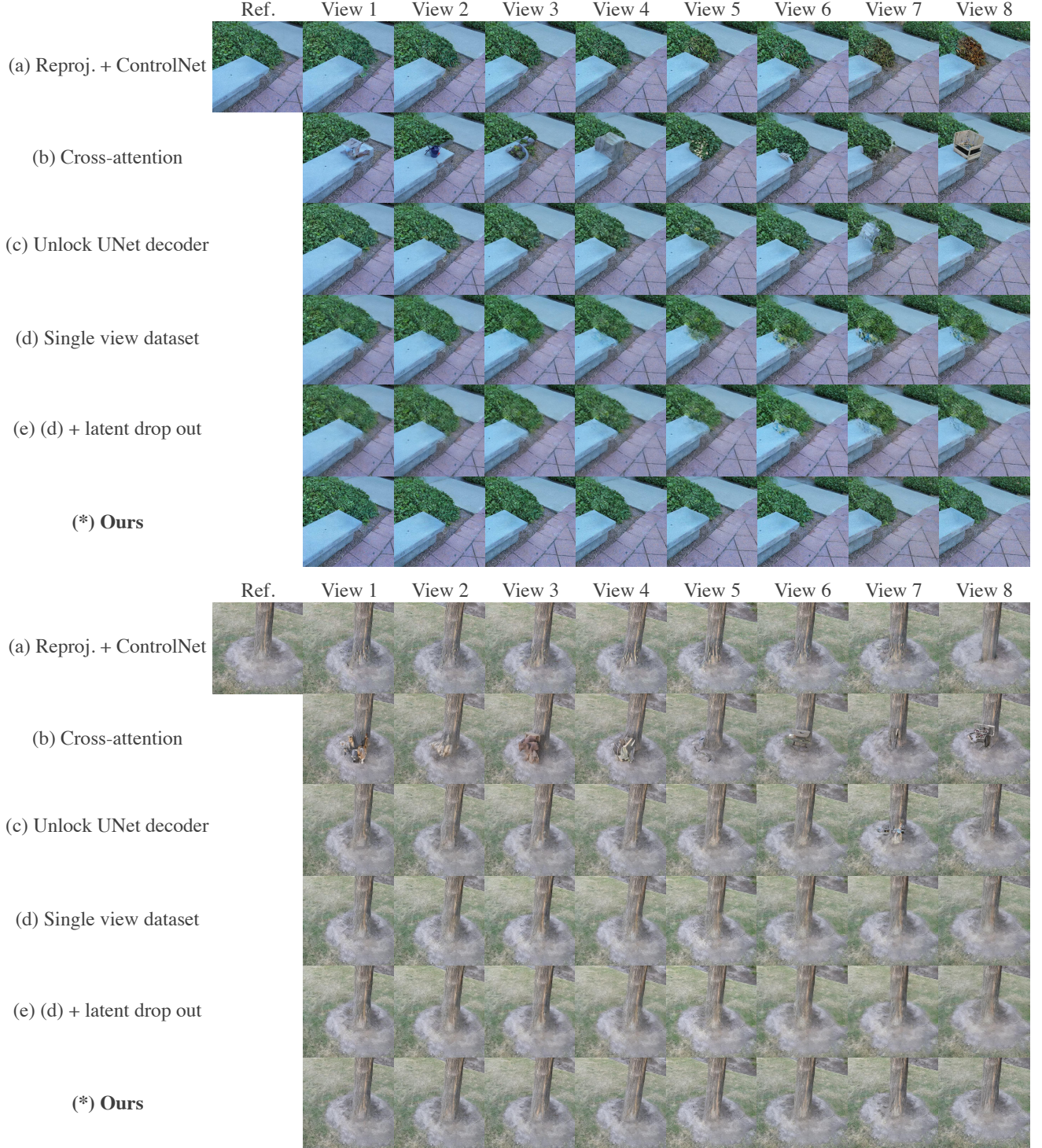


Figure 10. **Qualitative results for variants ablated in Sec. 4.3 and Tab. 5.** Please refer to Sec. F.2 for more discussion.

conditioning signals, which in turn results in less effective training.

3D-Aware Mask Area. We constrain the area of the generated 3D-aware mask to be approximately 30%–50% of

the projected object bounding box area. We determined this range through empirical inspection, as illustrated in Fig. 9. A smaller mask area (10%–30%) yields a trivial inpainting task with a weak training signal, while a larger mask area

(50%–70%) leaves insufficient contextual information for both the 3D foundation model and the pretrained inpainting model to operate effectively.

F.2. Qualitative Ablation Results

In this section, we provide a detailed qualitative analysis of the ablation studies reported in Sec. 4.3 and Tab. 5 of the main paper. We visually compare our full LaRP model against two categories of baselines: architectural alternatives (a–c) and data pipeline variations (d–e). As shown in Fig. 10, while some variants achieve partial success, they exhibit distinct failure modes that highlight the necessity of our specific design choices. The standard ControlNet (a) demonstrates a certain level of conditional inpainting ability, but its color control remains suboptimal. The cross-attention variant (b) completely fails to condition on the input reference view, generating inpaintings irrelevant to the reference content. The variant with an unlocked UNet decoder (c) learns cross-view conditioning to some extent. However, the results are unstable and prone to artifacts, which we attribute to the skewing of generative priors caused by training on data that is less diverse than that of the pretraining stage. The model trained on a single-view dataset (d) struggles to generate plausible inpaintings when the viewpoint difference becomes large, as it has never encountered the real “hole” patterns produced by latent reprojection during training. Simulating these holes by manually dropping out the control signal (e) improves upon (d), but still underperforms our full method (*), which leverages real reprojection during training. These results highlight the efficacy of our proposed LaRP architecture for cross-view conditioning and validate the necessity of the two-view dataset produced by our data pipeline.

G. Training Details

During training, we use a single masked target image for supervision. In addition to the reference and target image pair, we provide a text prompt “A <obj_label>” using the object label from the Objectron sequence processed by our pipeline. The masked region is filled with Gaussian noise rather than a solid color (*e.g.*, black) to prevent the 3D foundation model from confusing the masked region with meaningful geometric context.

H. Inference Details

To obtain the reference image for each scene, we select the input view with the largest mask and use the `stable-diffusion-2-inpainting` model to inpaint it using a descriptive prompt (Tab. 7). In line with our design principle of maximizing pretrained generative capability, we crop a square area centered on the masked region from the reference image for inpainting, and then

Table 7. **Per-scene text prompt used for inpainting the initial reference view.** All reference-based methods reported in the main paper use the same inpainting results for fair comparisons.

Scene	Inpainting Prompt
1	a blue bench, with a bush in the background
2	a trunk on a lawn
3	a red plastic fence
4	an empty concrete staircase
7	a manhole on a lawn
9	corner of a wall on the floor
10	a wooden bench in front of a white fence
12	a garden
book	a table in front of a wall
trash	corner of a wall on the floor

Table 8. **No-Reference IQA Evaluation.** Evaluated on the SPIn-NeRF dataset. ■ Best results. ■ Second-best results.

Method	SPIn-NeRF	3DGIC	MVI	MALD-NeRF	Ours
MUSIQ \uparrow	56.94	68.44	68.53	69.80	68.60
NIQE \downarrow	3.37	2.42	2.33	3.25	2.36

softly blend the result back into the original view. Specifically, after replacing the masked pixels with the inpainted content, we dilate the mask by 29 pixels, apply a Gaussian blur to the dilated mask, and perform alpha blending using the blurred mask as weights. To perform cross-view conditioned inpainting efficiently, LaRP supports using a single reference view to inpaint multiple target images in parallel, provided that the target images share sufficient visual overlap with the reference view. In this parallel setting, we feed the batch of target images along with the reference view to the 3D foundation model in a single pass to estimate the 3D attributes required for latent reprojection.

I. No-Reference IQA Analysis

Following the suggestions of our reviewers, we evaluate the absolute perceptual quality of the generated novel views using No-Reference Image Quality Assessment (NRIQA), specifically MUSIQ [30] and NIQE [45], shown in Tab. 8.

MUSIQ, a modern deep-learning-based metric aligned with human perception, scores MALD-NeRF and ours (LaRP+NeRF) the highest, which corroborates our qualitative observations (Fig. 5). Conversely, the traditional statistical metric NIQE ranks MALD-NeRF poorly. This highlights a known limitation of purely 2D statistical metrics: they often penalize the high-frequency details generated by adversarial optimization (MALD-NeRF) while favorably scoring smoother, less realistic outputs (3DGIC).

Furthermore, as noted during the review process, NRIQA metrics evaluate frames independently and cannot

Table 9. **Timing Breakdown.** Computational time required for each stage of our pipeline. The generalizable training of the LaRP model is a one-time cost (~ 14 hours). In contrast, the per-scene deployment (generating 60 inpainted views and optimizing the NeRF) is highly efficient, totaling only ~ 23 minutes. All timings are measured on a single NVIDIA RTX 4090 GPU.

Stage	One-Time Training	Per-Scene Deployment (60 images)			
	LaRP	VGGT	Inpainting	NeRF	Total
Time	~ 14 h	< 1 min	2 min	20 min	~ 23 min

penalize multi-view inconsistencies or temporal flickering. Therefore, while LaRP consistently achieves top-tier 2D visual quality scores, its true advantage lies in combining this visual fidelity with better 3D consistency, as demonstrated by Tab. 1.

J. Detailed Timing Breakdown

To provide a comprehensive understanding of the computational requirements of our framework, we detail the timing breakdown for each stage of our pipeline in Tab. 9.

The proposed multi-view inpainting method consists of two distinct phases: **1) One-Time Training:** Training the generalizable, pose-aware LaRP 2D diffusion model. **2) Per-Scene Deployment:** Per-scene inference (using VGGT and our inpainter to generate 60 multi-view images) followed by the 3D optimization of a NeRF on these generated results.

Since the LaRP diffusion model generalizes across different scenes without requiring per-scene adversarial fine-tuning, its ~ 14 -hour training is a strict one-time cost. Consequently, the actual per-scene processing time is highly efficient, requiring roughly 23 minutes in total to achieve state-of-the-art 3D-consistent inpaintings.