HAD: Hallucination-Aware Diffusion Priors for 3D Reconstruction

Supplementary Material

Overview. This supplementary material provides more analysis and experimental validation of our proposed HAD (Hallucination-Aware Diffusion) framework. In Sec. 7, we analyze hallucination patterns in NVS tasks across two diffusion paradigms – i.e., diffusion-assisted NVS with explicit 3DGS model and NVS directly via diffusion. In Sec. 8, we detail the architecture, training dataset curation, and implementation of our hallucination scoring network. In Sec. 9, to demonstrate generalization, we show that our scoring network effectively identifies hallucinations in video diffusion (GenFusion [43]) and multi-view diffusion (SVC [51]) without fine-tuning. We further show the quantitative improvement of GenFusion in NVS via diffusion-assisted 3DGS training. In Sec. 10, we provide additional qualitative comparisons on DL3DV and MipNeRF360 datasets, along with ablation studies validating our design choices, particularly the importance of pretrained initialization for reliable hallucination detection.

7. Hallucination in NVS via diffusion

We provide additional analysis to deepen understanding the hallucination issue introduced by diffusion models in NVS task. Specifically, we evaluate recent state-of-the-art methods from two representative paradigms: *Diffusion-assisted NVS with explicit 3DGS model* (e.g., Difix3D [41], GenFusion [43] and 3DGS-enhancer [28]), and *NVS directly via diffusion* such as SVC [51]. We consistently observe hallucinations across both paradigms, regardless of whether image diffusion, video diffusion, or direct diffusion-based synthesis is employed.

Diffusion-assisted NVS with explicit 3DGS model. As shown in Fig. 4, hallucinations mainly arise from imperfect geometry in the underlying 3D representation. Specifically, in regions with sparse observations or unseen viewpoints, 3DGS often produces floaters or severely distorted structures. When applied to such renderings, diffusion models tend to "correct" these artifacts by introducing semantically plausible but incorrect content, often borrowing from reference views, thereby amplifying inconsistencies and causing hallucinated geometry and appearance.

NVS directly via diffusion. As shown in Fig. 5, SVC [51], while achieving the photorealistic rendering, still suffers from a different mode of hallucination. As they directly rely on diffusion models without explicit 3D representations, the lack of geometric constraints often leads to structural distortions and inconsistent geometry across views, leading to the geometrically inconsistent structures across viewpoints.

8. Details of Hallucination Scoring Network

Overview of hallucination score network. We provide a detailed model architecture Fig. 6.

Training dataset curation. We provide additional details on the constructing training dataset for the hallucination score network. For all training scenes, we follow the Difix3D [41] pipeline under the 9-view setting to first reconstruct a 3DGS model from sparse input views. We then render all remaining views that are not included in the 9 input views, obtaining the corresponding 3DGS renderings. For each such view, we further apply the diffusion-based refinement module in Difix3D to generate enhanced images. This process results in a triplet of aligned images $(I_{\text{GT}}, I_{\text{difix}}, I_{\text{3DGS}})$ for each viewpoint:

- $I_{\textbf{GT}}$: the ground-truth image from the captured view.
- $I_{\textbf{3DGS}}$: the rendering from the reconstructed 3DGS model, which often contains artifacts such as floaters or distorted structures due to incomplete geometry.
- $I_{\textbf{difix}}$: the diffusion-enhanced result, where diffusion models attempt to refine the 3DGS rendering but may introduce hallucinated content.

These triplets $(I_{\text{GT}}, I_{\text{difix}}, I_{\text{3DGS}})$ and corresponding camper pose $\mathbf{c}$ serve as supervision for training the hallucination score network, enabling it to learn to identify hallucination introduced by diffusion model.

Training details. Given the constructed triplets, we train the hallucination score network based on the pretrained LVSM encoder. For each triplet, we use its camera pose as the target view, and select the three nearest input views from the training views used in Difix3D+ (i.e., the 9 input views for 3DGS reconstruction) based on camera pose proximity. The selected input views are encoded using the multiview encoder to extract multi-view features $\mathbf{F}$. The diffusion-enhanced image (I_{difix}) of the triplet is treated as the hallucinated target view. We concatenate the multi-view features with the target-view features extracted from I_{difix}, and feed them into a learnable U-Net to predict the hallucination score map. Optionally, I_{3DGS} can also be included as an additional input. The network is trained to estimate the MAE between the hallucinated view and the corresponding ground-truth image (I_{GT}), enabling it to identify hallucination regions introduced by diffusion models.

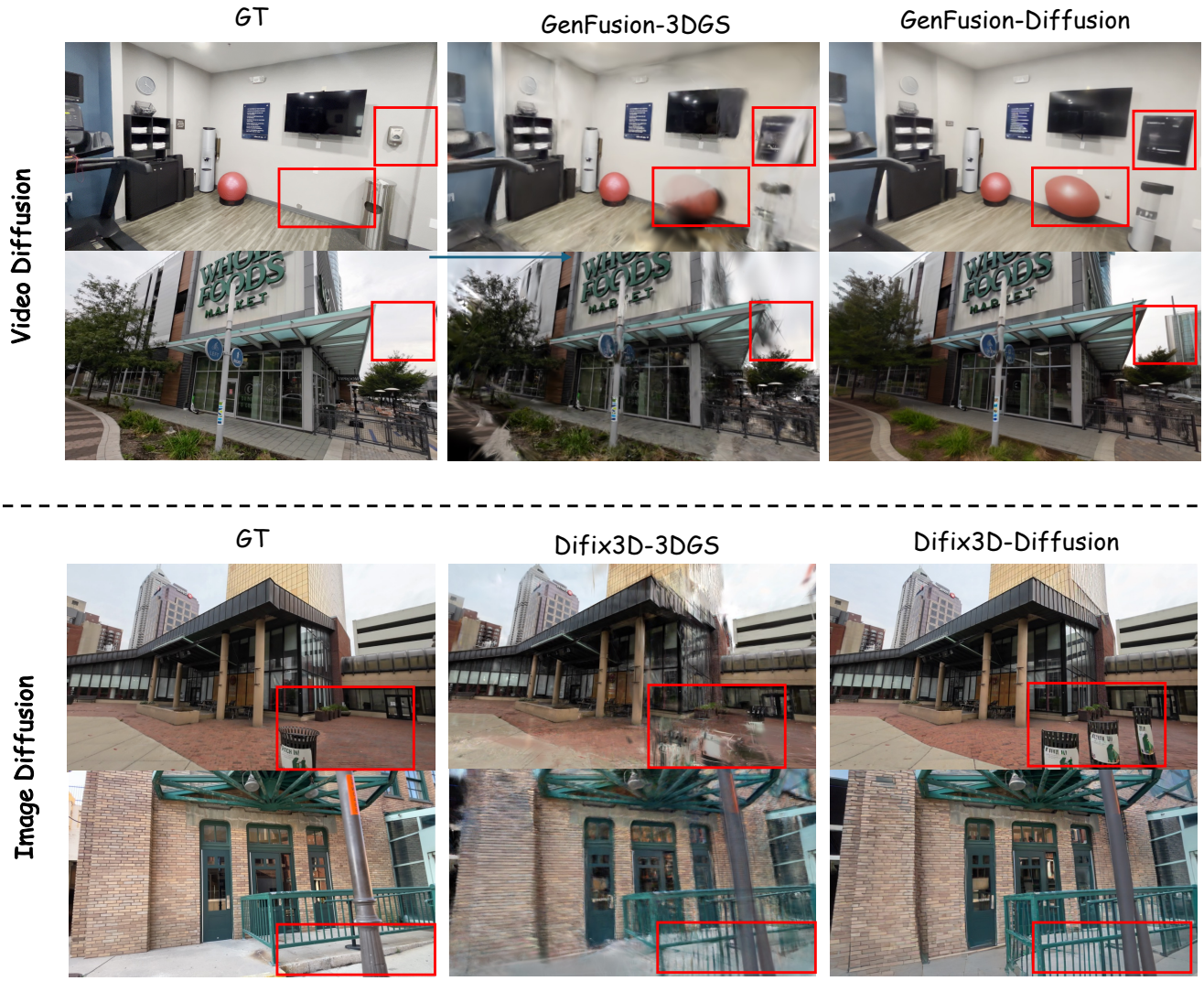

Figure 4. Qualitative demonstration of hallucination pattens in diffusion-assisted 3DGS pipelines. Both Difix3D (image diffusion) and GenFusion (video diffusion) iteratively enhance the 3DGS rendering. As a result, even mild artifacts in the initial 3DGS output, such as small floaters or subtle geometric distortions, can be progressively amplified by diffusion priors and eventually evolve into clearly visible hallucinations. This demonstrates that hallucination is not introduced abruptly, but can accumulate and become more pronounced during iterative refinement. **Note:** *method-3DGS* denotes direct rendering from 3DGS, while *method-diffusion* denotes diffusion-enhanced results.

Table 8. **Improving GenFusion [43] via HAD**. We demonstrate that our hallucination scoring network generalizes to video diffusion models without fine-tuning, effectively masking hallucinated pixels and improving reconstruction quality. We test it on DL3DV [25]

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Genfusion	20.57	0.7396	0.2845
Genfusion + HAD	**20.80**	**0.7415**	**0.2817**

9. Generalizing to other diffusion models

9.1. Improving GenFusion [43]

We integrate our hallucination scoring network into GenFusion [43] – the state-of-the-art video-diffusion-assisted 3DGS training pipeline. Importantly, we apply the same HAD model as in the main paper without any additional fine-tuning on video diffusion data. To ensure a fair comparison, we keep all original GenFusion settings and only incorporate our hallucination scoring module, testing on the DL3DV dataset [25]. Specifically, for each generated novel view, HAD predicts a hallucination score map and mask out pixels with low confidence, following the same strategy as Difix3D + HAD. As shown in Tab. 8, this simple integration leads to a PSNR improvement of +0.23. The qualitative results in Fig. 7 shows that our method effectively reduces hallucinated content and improves reconstruction fidelity.

Figure 5. Hallucination analysis for multi-view diffusion (SVC). Hallucination maps are computed against ground-truth images to highlight inconsistencies. We observe that hallucinations mainly arise from inaccurate 3D structure, where objects exhibit misaligned positions and distorted geometry across views due to the absence of explicit 3D constraints.

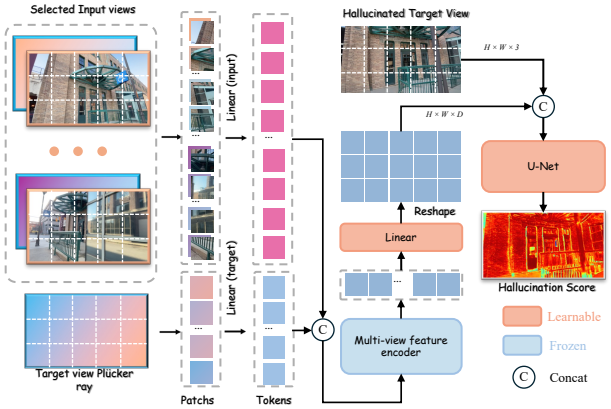

Figure 6. **Overview of hallucination scoring network –** The network predicts a pixel-wise hallucination score map $\mathbf{s}$ for a hallucinated novel view $\tilde{\mathbf{i}}_{\mathcal{G}}$. It consists of a multi-view feature encoder $\mathcal{V}$ (the frozen feature backbone of a pre-trained LVSM) and a three-layer U-Net score branch $\mathcal{S}$, which estimates hallucination scores using both multi-view features and the novel view image. The model is trained on curated multi-view and hallucinated novel-view pairs.

9.2. Hallucination Detection in SVC [51] and GenFusion [51]

To demonstrate the generalization beyond Difix3D, we apply the hallucination scoring network other diffusion-based pipelines, including video diffusion (GenFusion), multi-view diffuison (SVC). As shown in Fig. 8 and Fig. 9, our hallucination scoring network is able to effectively identify hallucination regions introduced by different diffusion models, despite not being trained on these data. This demonstrates the strong generalization capability across diverse diffusion paradigms.

Table 9. We compare different variants of hallucination score network including our model without 3DGS rendering input, our model without pretrained weights, and our full model in score estimation accuracy.

Method	Ours (w/o 3dgs input)	Ours (w/o pretrained)	Ours (full)
MAE ↓	0.044	0.054	**0.043**

10. More Results

10.1. Additional Qualitative Comparisons

We provide additional qualitative results on both the DL3DV – as shown in Fig. 11 and Fig. 10, and MipNeRF360 datasets – see Fig. 12. Note we also include the corresponding rendered videos in project website, providing a clearer comparison across viewpoints. Both the qualitative results and videos show that our method produces more geometrically consistent renderings with fewer hallucinated structures compared to the baselines.

10.2. More Ablation Studies

We study how two factors affect the performance of our scoring network: using pre-trained weights for the multi-view encoder $\mathcal{V}$ and adding the 3DGS-rendered image $\mathcal{R}_{\Phi}$ as an extra input. To evaluate the effect of pre-training, we train the multi-view encoder from scratch on the same dataset. To examine the role of the 3DGS-rendered image, we remove this input and only use the multi-view features $\mathbf{F}$ and diffusion-generated images $\tilde{\mathbf{i}}_{\mathcal{G}}$.

As shown in Tab. 10 and Tab. 9 where we study the impact on score estimation accuracy and the final 3D reconstruction, removing the pre-trained initialization leads to a clear performance drop, showing that the strong 3D-awareness is essential for predicting reliable hallucination

Figure 7. **Hallucination Scoring for GenFusion.** Our hallucination scoring network can also mitigate hallucinations in video diffusion without fine-tuning.

Table 10. The impact of different design choices in hallucination score network on the 3D reconstruction.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	**22.134**	**0.757**	**0.190**
W/o pretrain	21.600	0.748	0.1974
W/o input of 3dgs	21.960	0.755	0.1891

scores. In contrast, excluding the 3DGS-rendered input results in only a minor change, indicating that this cue is helpful but not crucial for our scoring network.

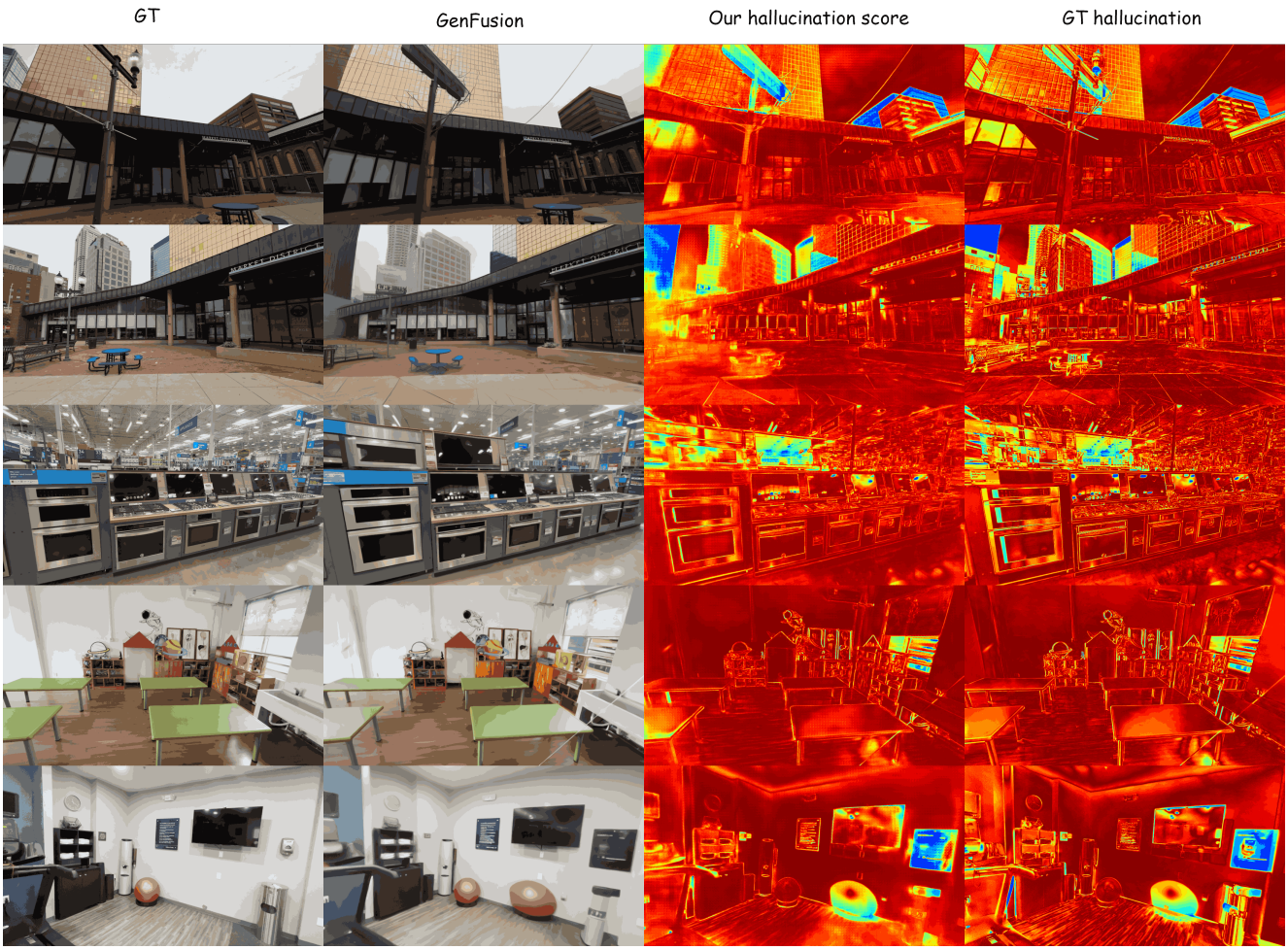

Figure 8. Generalization of our hallucination scoring network to video diffusion (GenFusion). Our model is **not fine-tuned** on the target diffusion model, demonstrating strong generalization across diffusion paradigms.

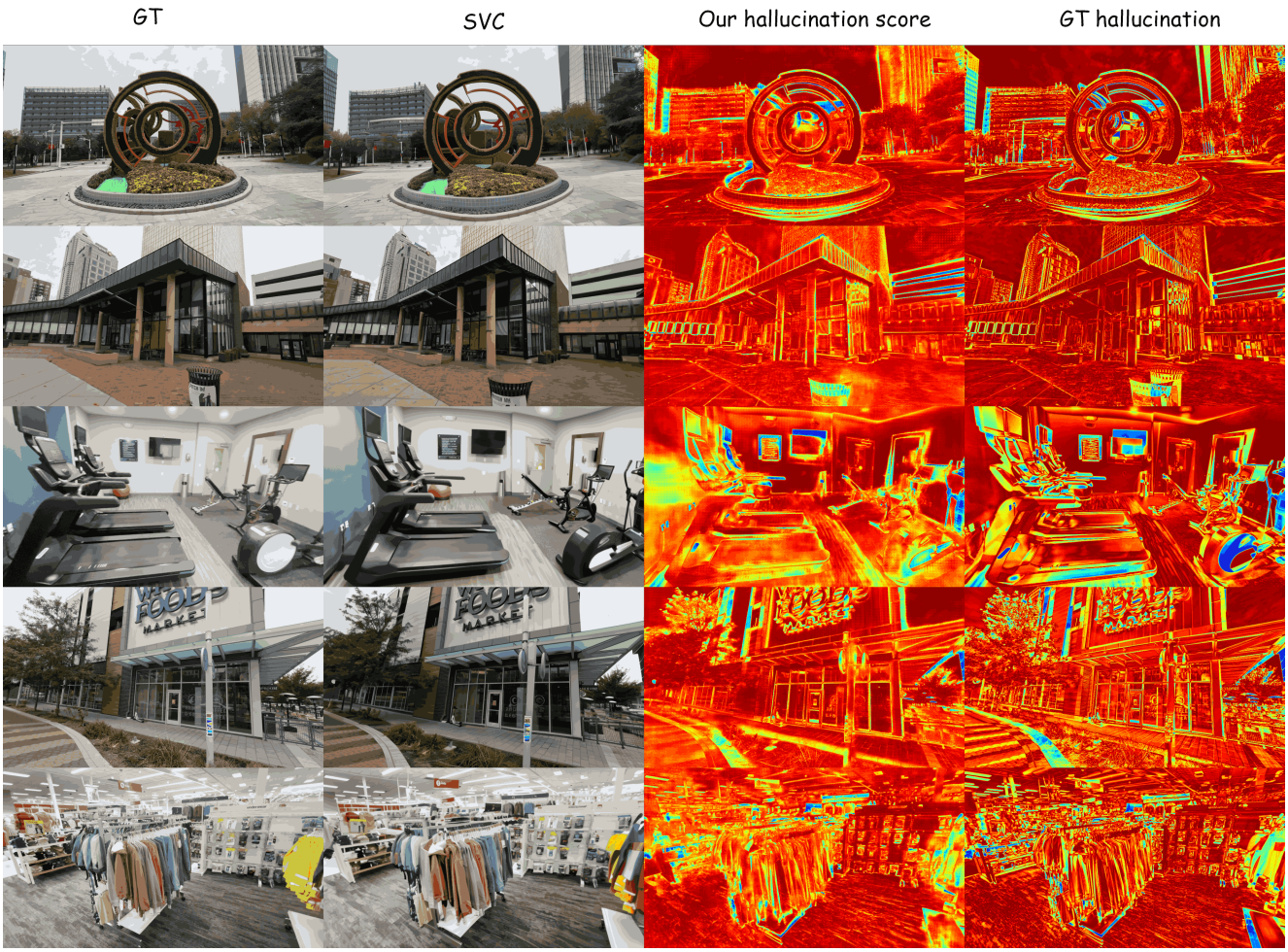

Figure 9. Generalization of our hallucination scoring network to multi-view diffusion (SVC). Our model is **not fine-tuned** on the target diffusion model, demonstrating strong generalization across diffusion paradigms.

Figure 10. **More Qualitative Results on DL3DV [25].**

Figure 11. **More Qualitative Results on DL3DV [25].**

Figure 12. **More Qualitative Results on MipNeRF-360 [3].**