EVALUATING MULTIVIEW 3D CONSISTENCY WHEN 3D FOUNDATION MODELS HALLUCINATE

Anonymous authors
Paper under double-blind review

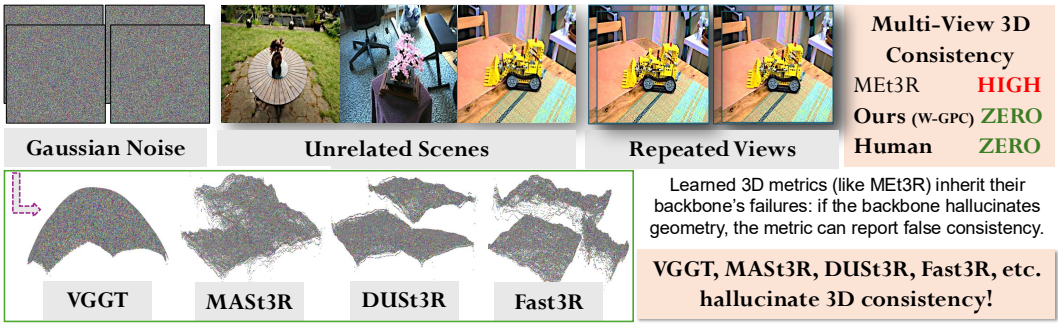


Figure 1: **Can these views be one scene?** Unrelated scenes, repeated views, and Gaussian noise should be scored as 3D-inconsistent because they do not define a single static scene. However, learned reconstruction backbones such as VGGT, MASt3R, DUSt3R, and Fast3R can still produce dense geometry on these inputs. Metrics built on these backbones, such as MEt3R, can therefore report spuriously high 3D consistency. We study this failure mode with SysCON3D, diagnose it with a backbone–residual–aggregation metric family, and introduce COLMAP-based metrics that score failed reconstruction as 3D inconsistency.

ABSTRACT

Multiview 3D consistency metrics assess whether generated images depict a single coherent, static 3D scene (eg. MEt3R scores random noise images to be near perfectly 3D consistent). We find that neural metrics, including MEt3R, can assign high consistency to inconsistent image sets and disagree substantially with human judgments on novel view synthesis (NVS) outputs. To diagnose these failures, we use SysCON3D, a controlled robustness benchmark that systematically isolates 3D inconsistency. We decompose reconstruction-based neural metrics into backbone, residual, and aggregation components, yielding a parametric family of neural metrics for controlled analysis and metric improvement. Despite our best neural variant being $\sim 3\times$ more robust than MEt3R on SysCON3D, crucial failure modes persisted. Our analysis identifies the cause of failure: feed-forward reconstruction backbones, including VGGT, MASt3R, DUSt3R, and Fast3R, hallucinate geometry and cross-view support for unrelated scenes and random noise. Motivated by these limitations, we introduce COLMAP-based metrics that use verified matches, camera registration, dense support, and reconstruction failure as explicit evaluation signals, and achieve the strongest human alignment of all metrics in our experiments.

1 INTRODUCTION

Multiview 3D consistency metrics exist to evaluate whether images represent observations of the same static 3D scene from different camera viewpoints. They are useful for scoring multiview outputs from an increasing number of sparse-view reconstruction models, generative models, and other methods, as well as evaluating multiview inputs.

For 3D reconstruction, these images can come from noisy sources, including images scraped from the internet or corrupted during capture and processing (Han et al., 2025). Recent Novel View Synthesis (NVS) and sparse-view reconstruction methods also operate on limited, noisy, or generated views (Xu

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.

et al., 2025; Yu et al., 2024; Wu et al., 2025a). Their inputs and outputs may contain geometric inconsistencies and appearance artifacts arising from data collection, model generalization failures, or deliberate corruption.

Having reliable multiview consistency metrics would reduce the extensive human labor needed to inspect hundreds of images for 3D inconsistencies. However, appearance metrics such as PSNR and SSIM require reference images and measure per-view fidelity, while measures such as Chamfer distance require reference geometry. These do not directly test whether a generated view set is mutually consistent with a single static scene without such references.

To address this gap, ground-truth-free metrics based on learned 3D reconstruction models have been proposed. MEt3R (Asim et al., 2025), for example, reconstructs image pairs using DUSt3R or MASt3R (Wang et al., 2024; Leroy et al., 2024), warps dense DINOv2/FeatUp features (Oquab et al., 2023; Fu et al., 2024), and averages feature disagreements. However, we found a serious misalignment: images from completely different scenes, or even random noise, can receive highly favorable consistency scores. Neural metric rankings also disagree substantially with human judgments on NVS outputs. This is concerning when downstream evaluations rely on these scores to compare models or select clean multiview data, especially in applications such as crime-scene reconstruction and autonomous driving.

We decided to study this misalignment systematically as a robustness problem for automatic evaluators. We introduce **SysCON3D**, a controlled benchmark that tests whether metrics distinguish clean multiview sets from simple, well-understood inconsistencies: single outlier views, cross-scene mixtures, patched noise, and full Gaussian noise. We also test repeated images as degenerate multiview evidence. The idea is simple: a reliable metric should score corrupted sets as less consistent than clean sets and respond to increasing inconsistency. These deliberately controlled tests expose poor robustness in the evaluated neural metrics.

To further diagnose the problem, we develop a parametric family of neural metrics by decomposing each evaluator into a reconstruction backbone, a residual function, and an aggregation function. Changing these components yields stronger metrics. On SysCON3D, our best variant, **MASt3R-W-IMQ**, assigns worse average scores to inconsistent sets than to consistent sets in roughly $3\times$ as many tests as MEt3R. However, it still rates Gaussian noise as more consistent than real multiview observations.

The analysis leads to an **important finding**: data-driven, feed-forward reconstruction models such as VGGT (Wang et al., 2025), MASt3R (Leroy et al., 2024), DUSt3R (Wang et al., 2024), and Fast3R (Yang et al., 2025) can hallucinate dense geometry and cross-view support for unrelated scenes and random noise. Neural metrics inherit these errors by evaluating feature agreement on unsupported view correspondences. Our initial analysis suggests poor robustness to unusual and noisy inputs, where the models produce geometry with no grounding in shared image evidence.

Confronted with these failures, we introduce **COLMAP-based 3D consistency metrics** (Schönberger & Frahm, 2016) that rely on geometric verification. They use verified matches, camera registration, dense support, and reconstruction failure as evaluation signals. They inherit the usual limitations of structure-from-motion and multiview stereo, but their verification and failure stages are explicit and interpretable. On NVS outputs, these metrics achieve the strongest alignment with human judgments in our experiments, with up to $4\times$ the Spearman rank correlation of MEt3R.

This result exposes a useful contrast between two traditions in vision: classical multiview geometry verifies reconstruction through explicit evidence but can be slow and can fail; learned reconstruction models are fast, flexible, and often pose-free, but their behavior on inconsistent inputs is less transparent. We find that the robustness and interpretability of classical geometric verification make COLMAP-based metrics the most reliable alternative among those evaluated.

Our contributions are: (i) SysCON3D, a controlled robustness benchmark for ground-truth-free multiview 3D consistency metrics; (ii) a parametric neural metric family separating backbone, residual, and aggregation components for systematic ablations and stronger metrics; (iii) a diagnostic study showing that learned 3D reconstruction backbones hallucinate geometric support on inconsistent inputs; (iv) failure-aware COLMAP metrics using classical geometric verification to produce interpretable consistency scores; and (v) a structured human study assessing alignment with composite judgments of 3D consistency, realism, and plausibility.

2 Related Work

NVS, sparse-view reconstruction, and learned 3D backbones. Recent NVS and sparse-view reconstruction methods include diffusion-based generators (Zhou et al., 2025; Yu et al., 2024; You et al., 2024; Cao et al., 2025; Wu et al., 2025a;b) and feed-forward 3D regression methods that predict Gaussian or scene representations from sparse views (Xu et al., 2025; Ziwen et al., 2025; Chen et al., 2024). Ground-truth-free 3D consistency metrics increasingly rely on learned reconstruction backbones. DUSt3R (Wang et al., 2024) and MASt3R (Leroy et al., 2024) predict dense pairwise geometry, while Fast3R (Yang et al., 2025) and VGGT (Wang et al., 2025) perform global multiview reconstruction in a single forward pass. These models are fast and pose-free, but their learned priors make failures difficult to interpret when they are used as evaluators.

Metrics for multiview generation. Reference-based NVS and reconstruction evaluation commonly use PSNR, SSIM, LPIPS, Chamfer distance, or benchmark-specific geometric criteria (Yao et al., 2018; Knapitsch et al., 2017). These metrics require reference renderings or ground-truth geometry and do not directly measure the mutual consistency of a generated view set. MEt3R (Asim et al., 2025) addresses this setting with a pose-free pairwise score from learned geometry and warped DINOv2 features (Oquab et al., 2023; Fu et al., 2024), but its mean aggregation can hide scene-level failures. SED/TSED (Yu et al., 2023) uses epipolar consistency of SIFT matches when poses are known, while PRISM (Stern et al., 2025) compares object-centric embedding distributions against a reference anchor set. Gen3DEval (Maiti et al., 2025) and Eval3D (Duggal et al., 2025) evaluate object-centric 3D generation with foundation-model probes. Our focus is scene-level, ground-truth-free multiview consistency under controlled corruptions.

Robustness and geometric verification. Robustness benchmarks and robust statistics study how estimators behave under corruptions, outliers, and contamination (Hendrycks & Dietterich, 2019; Huber, 1964; Hampel, 1974). We bring this view to 3D consistency metrics through SysCON3D, where corruptions are multiview inconsistencies such as outlier views, cross-scene mixtures, repeated inputs, and noise. Our neural metric family uses distributional aggregation, related to kernel two-sample distances such as MMD (Gretton et al., 2012), to preserve residual structure lost by pairwise means. Our classical metrics instead use geometric verification: RANSAC-style inlier support (Fischler & Bolles, 1981) and COLMAP SfM/MVS (Schönberger & Frahm, 2016) expose registration, dense support, and reconstruction failure as measurable signals of 3D consistency.

3 Systematic Corruption Analysis for 3D Consistency Metrics

The aforementioned examples (like in Figure 1) show that neural metrics such as MEt3R and PRISM can score inconsistent image sets as highly 3D-consistent. To determine whether these failures are systematic, we introduce **SysCON3D**, a controlled benchmark spanning clean multiview observations, cross-scene mixtures, and noise. The idea is simple: a reliable metric should distinguish corrupted sets from consistent ones and reflect increasing inconsistency. We vary the corruption type and number of views to study where this behavior breaks down.

Benchmark Construction We construct three groups of multi-view image sets from Mip-NeRF360 (Barron et al., 2022) scenes at each viewpoint count $K \in \{3, 6, 9, 12, 15, 18, 21\}$. The groups comprise real 3D-consistent image sets and controlled inconsistencies of increasing severity, and each level isolates a distinct failure mode that an evaluator should detect.

(i) Consistent (L_0). K training views from a single scene (9 samples per K, one per scene). Any reliable metric should score such scenes as most 3D-consistent.

(ii) SysCON3D-M*: Cross-scene mixtures.* Foreign views are introduced at increasing density (8 samples per K each):

L_1 – **Single-outlier**: $K-1$ views from one scene plus 1 foreign view ($\sim 2/K$ cross-scene pairs).

L_2 – **Controlled mixture**: $\sim 30\%$ foreign views, $\sim 50-60\%$ cross-scene pairs at $K \geq 6$.

L_3 – **Random mixture**: each view sampled i.i.d. from the 9 scenes ($\sim 89\%$ cross-scene pairs).

(iii) SysCON3D-N*: Noise corruptions.*

- **Patched Gaussian**: consistent views with 4 $\mathcal{N}(0.5, 0.2^2)$ noise patches (4 samples/K).
- **Gaussian**: every pixel sampled i.i.d. from $\mathcal{N}(0.5, 0.2^2)$, clipped to $[0, 1]$ (4 samples per K).

In total, SysCON3D contains 9 consistent and 32 inconsistent samples per K across 7 view counts, all with fixed random seeds.

Robustness criteria. We evaluate two complementary properties.

Win rate measures how often a corrupted group receives a worse mean score than the consistent reference L_0. We report it per corruption across seven view counts, and overall across all $5 \times 7 = 35$ corruption–view-count comparisons. We also report Cohen's d, the mean score gap divided by the pooled standard deviation, with positive d indicating correct separation. These comparisons evaluate group means, not individual-set classification accuracy.

Rank agreement additionally tests whether the group means follow the prescribed corruption-severity ordering, using Kendall's τ averaged across view counts (Section J). This distinguishes metrics that merely separate corruptions from L_0 from those that also rank corruption levels correctly.

Table 1: **Robustness of existing metrics on SysCON3D.** Entries show mean Cohen's d (win rate) across seven view counts $K \in \{3, 6, \ldots, 21\}$. A win ($d > 0$) means the corrupted group receives a worse mean score than consistent sets (L_0). Overall pools all 35 corruption–view-count comparisons.

Metric	Single outlier (L_1)	Controlled mixture (L_2)	Random mixture (L_3)	Patched Gaussian	Gaussian noise	Overall win rate
MEt3R	-0.48 (14%)	-1.03 (0%)	-1.57 (0%)	$+1.13$ (100%)	-3.96 (0%)	23%
PRISM-MMD	-0.08 (43%)	-0.16 (43%)	-0.09 (29%)	$+1.61$ (100%)	$+4.71$ (100%)	63%

As shown in Table 1, both MEt3R and PRISM-MMD struggle to distinguish cross-scene mixtures from consistent sets. This is particularly concerning for MEt3R, which targets ground-truth-free, scene-general evaluation, whereas PRISM uses an object-centric reference distribution. These failures motivate us to investigate why the metrics fail and whether we can design more robust alternatives.

4 Diagnosing and Improving Neural Metrics

To diagnose these failures and design better metrics, we decompose reconstruction-based neural metrics into three components: the backbone, residual function, and aggregation function.

Problem setup. Given K RGB images $\mathcal{I} = \{I_i\}_{i=1}^{K}$, with pixel grids Ω_i^{img}, we seek a ground-truth-free score of their consistency with a single static, rigid scene. Dynamic and non-rigid scenes are outside our scope. Lower neural scores indicate greater consistency.

4.1 A backbone–residual–aggregation family

We write a reconstruction-based neural metric as

$$m(\mathcal{I}) = \mathcal{A}\big(\psi\big(B(\mathcal{I}), \phi(\mathcal{I})\big)\big)\,. \tag{1}$$

The backbone B supplies geometry and pixel correspondences, ϕ extracts image features, ψ measures feature disagreement at those correspondences, and $\mathcal{A}$ aggregates the residuals. Geometry therefore determines *which pixels to compare*, while the residual compares their features.

Backbone. The backbone provides reconstructed points $\mathbf{x}_p$ and camera projections π_i. MASt3R reconstructs pairs, evaluated over all $\binom{K}{2}$ image pairs; VGGT and Fast3R process all K views jointly. We use MASt3R to match our evaluated MEt3R baseline; DUSt3R results are in the appendix.

Residuals. Frozen DINOv2 features, upsampled with FeatUp, give dense maps $F_i = \phi(I_i) \in \mathbb{R}^{768 \times H_i \times W_i}$. For corresponding pixels u and v, cosine dissimilarity is $r_{ij}(u, v) = 1 - \frac{\langle F_i(u), F_j(v) \rangle}{\|F_i(u)\| \, \|F_j(v)\| + \varepsilon}$, where $\varepsilon > 0$ ensures numerical stability. We form residuals in two ways.

Warp-based residuals (W). We lift pixel u in view i to its reconstructed point $\mathbf{x}_{i,u}$ and project it into view j. Collecting both directions gives $\mathcal{E}_{\text{warp}} = \biguplus_{i \neq j} \{r_{ij}(u, \pi_j(\mathbf{x}_{i,u})) : u \in \mathcal{V}_{i \to j}\}$, where $\mathcal{V}_{i \to j} \subseteq \Omega_i^{\text{img}}$ contains pixels accepted by the depth-consistency test, and $\biguplus$ retains residual multiplicities. These correspondences come from the predicted geometry.

Point-consistency residuals (PC). For each point $\mathbf{x}_p$ observed in views $\mathcal{S}_p$, with $s_p = |\mathcal{S}_p| \geq 2$, we average feature disagreement across those views: $e_p^{\text{PC}} = [s_p(s_p - 1)]^{-1} \sum_{i,j \in \mathcal{S}_p,\, i \neq j} r_{ij}(\pi_i(\mathbf{x}_p), \pi_j(\mathbf{x}_p))$. The collection $\mathcal{E}_{\text{PC}} = \{e_p^{\text{PC}} : s_p \geq 2\}$ contains one residual per point. We evaluate PC with VGGT and Fast3R, and W with all three backbones.

Aggregation. Given residuals $\mathcal{E} = \{e_a\}_{a=1}^N$ with empirical distribution P, we compare P with the ideal zero-residual distribution δ_0. The mean baseline is $m_{\text{Base}}(\mathcal{I}) = N^{-1} \sum_a e_a = W_1(P, \delta_0)$ for non-negative scalar residuals. Averaging can dilute sparse inconsistencies: a single outlier affects only $2/K$ of all view pairs.

We also test aggregations that depend on more than the mean. For a kernel κ and $N \geq 2$, we use $\widehat{\text{MMD}}_\kappa^2(P, \delta_0) = [N(N-1)]^{-1} \sum_{a \neq b} \kappa(e_a, e_b) - 2N^{-1} \sum_a \kappa(e_a, 0) + \kappa(0, 0)$. The RBF kernel is $\kappa_{\text{RBF}}(e, e') = \exp(-(e - e')^2/(2\sigma^2))$, with median-heuristic bandwidth or fixed $\sigma = 0.15$. The inverse multiquadric kernel is $\kappa_{\text{IMQ}}(e, e') = (1 + (e - e')^2)^{-1/2}$. Energy aggregation uses $m_{\text{Energy}}(\mathcal{I}) = 2\,\mathbb{E}_{e \sim P}|e| - \mathbb{E}_{e, e' \sim P}|e - e'|$.

Metric names specify backbone–residual–aggregation: MASt3R-W-IMQ uses warp residuals with IMQ-MMD, while Fast3R-PC-Energy uses point-consistency residuals with energy distance. Our evaluated MEt3R baseline is MASt3R-W-Base.

Table 2: **Robustness analysis of 3D consistency metrics on the SysCON3D Benchmark.** Cohen's d (averaged over $K \in \{3, \ldots, 21\}$) and *win rate* (fraction of K values with $d > 0$) comparing consistent (L_0) sets against five 3D inconsistencies. Positive d indicates the metric correctly assigns higher (worse) scores to inconsistent inputs. L_1–L_3 denote increasing levels of cross-scene corruption. The repeated-views (Identical Images) degenerate diagnostic is reported separately in app. table 11. Top-3 values per column are shaded.

		increasing cross-scene corruption $\longrightarrow$										
Backbone	Variant	Mixed-1-Outlier [L_1]		Mixed-Controlled [L_2]		Mixed-Random [L_3]		Patched Gaussian		Gaussian Noise		Overall
		Avg d	Win% (d)	Avg d	Win% (d)	Avg d	Win% (d)	Avg d	Win% (d)	Avg d	Win% (d)	Win% (d)
MASt3R	W-Energy	+0.03	57%	+1.01	86%	+3.15	100%	+0.94	100%	-3.36	0%	69%
MASt3R	W-MMD	-0.78	14%	-1.10	14%	+0.53	71%	-0.61	14%	-0.93	0%	23%
MASt3R	W-IMQ	+0.12	57%	+1.25	100%	+3.56	100%	+1.16	100%	-2.64	0%	71%
Fast3R	W-Base	-0.04	29%	+1.04	86%	+2.68	100%	-0.66	14%	-9.52	0%	46%
Fast3R	W-Energy	-0.07	43%	+0.66	86%	+2.53	100%	-0.62	0%	-6.09	0%	46%
Fast3R	W-MMD	-0.19	14%	+0.17	71%	+0.97	100%	-0.90	0%	-3.75	0%	37%
Fast3R	W-IMQ	-0.09	43%	+0.86	86%	+2.30	100%	-0.48	0%	-5.48	0%	46%
Fast3R	PC-Base	+0.16	57%	+1.41	100%	+3.42	100%	+0.33	71%	-3.24	0%	66%
Fast3R	PC-Energy	+0.08	43%	+1.52	100%	+3.87	100%	+0.62	100%	-2.75	0%	69%
Fast3R	PC-MMD	-0.24	29%	+0.38	71%	+1.01	86%	-0.63	57%	-0.25	43%	57%
Fast3R	PC-IMQ	+0.18	43%	+1.67	100%	+4.46	100%	+0.42	71%	-2.44	0%	63%
VGGT	W-Base	+0.14	57%	+0.81	86%	+2.36	100%	-0.65	0%	-9.18	0%	49%
VGGT	W-Robust	+0.22	57%	+1.23	100%	+2.63	100%	-0.67	0%	-9.12	0%	51%
VGGT	W-Energy	+0.12	71%	+0.34	86%	+1.78	86%	-0.49	0%	-6.43	0%	49%
VGGT	W-MMD	-0.07	43%	+0.08	43%	+0.89	100%	-0.26	0%	-4.60	0%	37%
VGGT	W-IMQ	+0.11	57%	+0.53	86%	+1.94	100%	-0.60	0%	-5.76	0%	49%
VGGT	PC-Base	-0.17	43%	+0.37	57%	+1.58	100%	+0.00	71%	-6.89	0%	54%
VGGT	PC-Energy	-0.30	29%	+0.33	57%	+1.41	100%	+0.07	57%	-5.97	0%	49%
VGGT	PC-MMD	-0.46	14%	-0.47	14%	+0.08	43%	+0.20	43%	+1.31	100%	43%
VGGT	PC-IMQ	-0.25	29%	+0.39	71%	+1.53	100%	+0.02	57%	-4.37	0%	51%
SED		+0.15	57%	-0.16	43%	-0.62	0%	+0.09	57%	+161.74	86%	49%
TSED		-0.25	14%	+0.43	86%	+0.75	71%	-0.35	0%	+2.39	100%	54%

4.2 Robustness Results on SysCON3D

Table 2 reports group separation across corruption types and view counts; Figure 2 shows the score trends. Alongside the neural metrics, we report SED and TSED (Yu et al., 2023), which require known fundamental matrices. SED measures median symmetric epipolar distance over SIFT matches; TSED thresholds this distance to count consistent pairs.

Aggregation improves separation. With MASt3R, replacing the mean with IMQ-MMD increases the overall win rate from 23% (8/35 comparisons) to 71% (25/35): roughly three times as many correct comparisons between corrupted and consistent group means. MASt3R-W-IMQ separates L_2, L_3, and patched Gaussian from consistent sets at every tested K. Energy aggregation performs similarly, reaching 69% overall. RBF-MMD is less effective for warp residuals, with overall win rates of 23–37%.

The gains depend on the component combination. With mean aggregation, Fast3R-W-Base and VGGT-W-Base reach 46% and 49%, compared with MEt3R's 23%. Neither retains this advantage over MASt3R with IMQ aggregation. Changing the residual also helps in some combinations: Fast3R-PC-Energy reaches 69%, compared with 46% for Fast3R-W-Energy.

Basic failures remain. Despite its higher overall win rate, MASt3R-W-IMQ still scores Gaussian noise as *more consistent* than real multiview sets at every tested K, and separates single-outlier sets in only four of seven comparisons. VGGT-PC-MMD separates Gaussian noise at every K, but succeeds in only one of seven comparisons for each of L_1 and L_2. Attention-based view rejection with a rejected-view penalty, our VGGT-W-Robust adaptation of RobustVGGT (Han et al., 2025), reaches 51% overall but fails both noise conditions at every K (app. section I). Two anomalous entries merit brief explanation: VGGT PC-MMD achieves $d = +1.31$ on Gaussian noise due to RBF kernel saturation, & SED reports $d = +161.74$ as SIFT finds no keypoints on noise images.

Implications. The neural metrics improve over MEt3R but remain unreliable. Every tested variant still fails at least one basic corruption test. Even MASt3R-W-IMQ, which achieves the highest overall win rate, scores Gaussian noise as more 3D-consistent than real multiview observations.

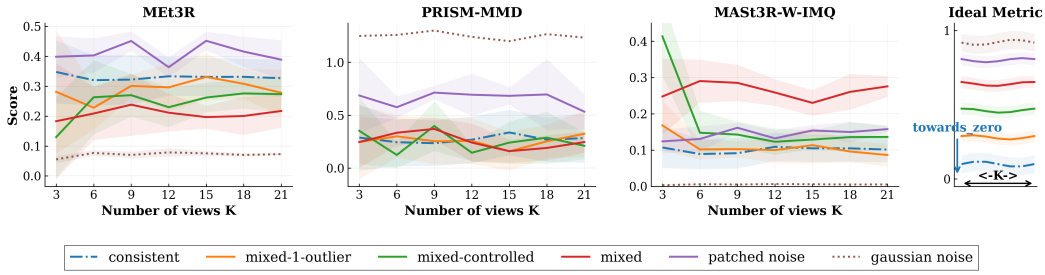


Figure 2: **Metric trends across SysCON3D scene types.** Each panel plots metric score (lower=better) vs. view count K across SysCON3D scene types; a robust metric should separate these bands in order of increasing 3D inconsistency. **Panel 1:** MEt3R conflates L_0 with cross-scene mixtures and scores Gaussian noise as more consistent than L_0. **Panel 2:** PRISM-MMD – recovers the expected SysCON3D ordering best (app. table 12) but fails on cross-scene mixtures (L_1–L_3), correctly separating only noise. **Panel 3:** MASt3R W-IMQ – distributional aggregation cleanly separates L_0 from cross-scene mixtures, but still scores noise as more consistent due to degenerate reconstructions (app. section H). **Panel 4:** Ideal: L_0 lowest, with progressively higher non-overlapping bands for L_1–L_3 and Gaussian noise.

4.3 Human Alignment on Real NVS Outputs

We next ablate improvements on SysCON3D translate with agreement with human judgments on 3D consistency of generated views.

Evaluation setup. We evaluate eight NVS methods spanning diffusion-based generation (Zhou et al., 2025; Yu et al., 2024; Cao et al., 2025; You et al., 2024; Wu et al., 2025a) and feed-forward reconstruction (Xu et al., 2025; Ziwen et al., 2025; Chen et al., 2024) on 33 scenes: 24 from DL3DV (Ling et al., 2024) and nine from Mip-NeRF360 (Barron et al., 2022). Each method receives

$K \in \{3, 6, 9\}$ input views and produces a 50-frame 360° orbit. Inference and trajectory details are in app. Section G.

Human study. 20 participants provide 1310 pairwise comparisons of anonymized NVS videos, with source images shown alongside. They vote separately on 3D consistency, visual realism, and plausibility. We combine these votes using weighted Elo (Glickman & Jones), with respective weights $2 : 1 : 1$, to obtain composite human rankings. We measure alignment using Spearman's ρ between metric-induced and human rankings across the eight methods. The protocol and interface are in app. Section F.

Table 3: Metric-induced rankings vs. human rankings of 8 NVS methods on 24 DL3DV scenes across three view-count splits (K=3, 6, 9). We compare three 3D consistency metrics based on data-driven reconstruction backbones: MEt3R, PRISM-MMD (PRISM), and MASt3R-W-IMQ (IMQ) - selected for its highest overall win rate against inconsistent scene types (table 2). Colors encode ranks 1 (best, green) to 8 (worst, red); rows sorted by K=3 human rank (**H**). MASt3R-W-IMQ trails PRISM-MMD marginally at K=3 but achieves the highest human alignment at K=6 and K=9, while MEt3R shows weak or negative correlation at K=3. Bottom row: Spearman rank correlation (ρ) measuring alignment with human ranking.

Method	K=3				K=6				K=9			
	H↓	MEt3R↓	IMQ↓	PRISM↓	H↓	MEt3R↓	IMQ↓	PRISM↓	H↓	MEt3R↓	IMQ↓	PRISM↓
DepthSplat	1	0.318 7	0.089 5	0.298 5	1	0.304 5	0.083 3	0.295 5	1	0.298 3	0.084 3	0.298 5
S-V-Camera	2	0.259 1	0.065 1	0.168 1	2	0.266 1	0.069 1	0.152 2	3	0.270 1	0.068 1	0.129 1
Long-LRM	3	0.375 8	0.111 8	0.391 7	3	0.396 8	0.131 8	0.339 7	2	0.381 8	0.125 7	0.351 7
MVGenMaster	4	0.294 4	0.079 4	0.171 3	5	0.284 2	0.074 2	0.123 1	5	0.286 2	0.077 2	0.140 2
ViewCrafter	5	0.287 2	0.078 2	0.187 4	4	0.296 3	0.090 4	0.200 3	4	0.318 5	0.100 4	0.204 3
Difix3D	6	0.311 5	0.097 7	0.459 8	8	0.337 7	0.118 7	0.400 8	8	0.356 7	0.131 8	0.388 8
NVS-Solver	7	0.290 3	0.079 3	0.169 2	7	0.302 4	0.101 5	0.224 4	7	0.307 4	0.115 6	0.228 4
MVSplat360	8	0.318 6	0.096 6	0.325 6	6	0.331 6	0.107 6	0.331 6	6	0.341 6	0.111 5	0.317 6
Rank Corr. (ρ)↑	–	-0.10	0.14	**0.19**	–	0.26	**0.48**	0.31	–	0.24	**0.48**	0.24

Results. We select MASt3R-W-IMQ for its highest overall win rate on SysCON3D. On DL3DV (Table 3), it improves over MEt3R at every K, but reaches only $\rho = 0.14$ at $K = 3$ and 0.48 at $K = 6, 9$. Humans rank DepthSplat first at every K, while MASt3R-W-IMQ ranks Stable Virtual Camera first. **Better benchmark separation does not resolve the disagreement with human judgments.** We next investigate why these failures persist.

4.4 Why Neural Metrics Fail: Hallucinated Geometric Support

To understand why the improved neural metrics still fail, we examine the reconstructions they rely on. We find that VGGT (Wang et al., 2025), DUSt3R (Wang et al., 2024), MASt3R (Leroy et al., 2024), and Fast3R (Yang et al., 2025) **can hallucinate geometry and cross-view support for unrelated scenes and random noise** (Figure 3). For example, in L_3, approximately 89% of view pairs come from different scenes and should not overlap, yet DUSt3R reports zero overlap for only 26% of all pairs (Table 7). Metrics such as MEt3R then compare features at these unsupported correspondences. DINOv2/FeatUp features can be similar across unrelated surfaces, while noise patches can produce nearly indistinguishable features. These false correspondences can therefore receive small residuals, leading the metric to report high consistency without valid geometric support. **The problem begins before aggregation: the metric is already comparing pixels that do not represent the same physical point.** Better aggregation improves some scores but leaves these failures unresolved. We therefore turn to COLMAP-based metrics (Section 5), which use geometric verification, registration, dense support, and reconstruction failure as explicit evaluation signals.

5 Classical Verification Metrics

The neural failures motivate an alternative based on explicit geometric verification. COLMAP (Schönberger & Frahm, 2016) can reject matches, fail registration, or produce little dense support. We turn these outcomes into consistency scores, where higher is better.

Reconstruction outputs. Let $\mathcal{T}$ contain all attempted images, $\mathcal{R} \subseteq \mathcal{T}$ the registered views, and $\mathcal{J} \subseteq \mathcal{R}$ those with readable photometric and geometric depth maps, $D_{p,i}$ and $D_{g,i}$. These come from

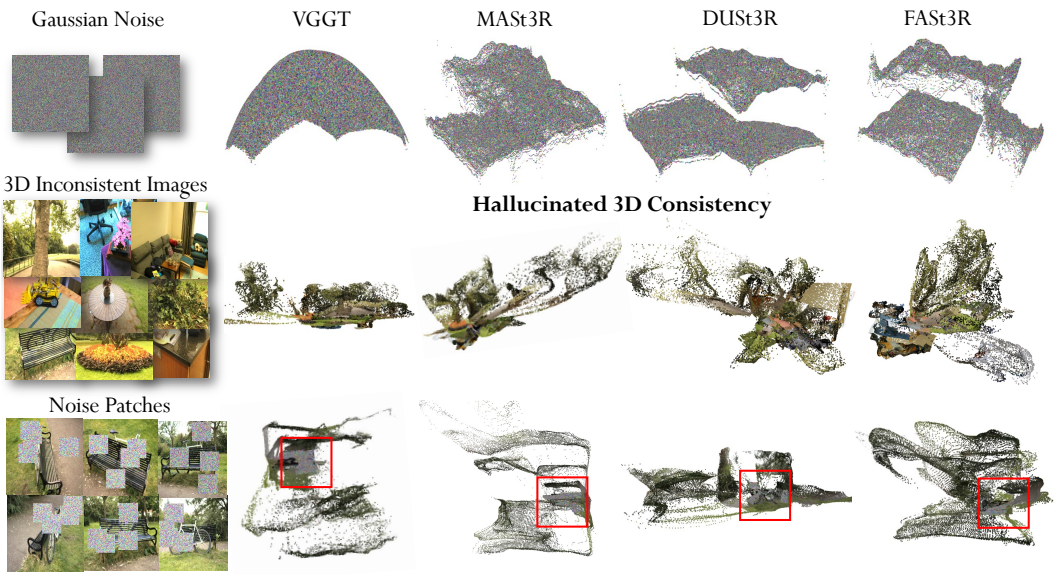


Figure 3: **Learned backbones hallucinate geometric support** on noise and cross-scene mixtures. On noise, they produce recurring, model-specific shapes, such as VGGT's umbrella-like structure (Section H.4).

PatchMatch passes without and with geometric consistency, using the same SfM reconstruction and scale. Each run uses K images for SysCON3D, or K input views plus 25 generated frames for NVS.

Depth agreement. On the dense-map grid Ω_i, pixels with finite depths and $D_{g,i}(u) > 10^{-5}$ receive

$$q_i(u) = 1 - \mathrm{clip}\left(\frac{|D_{p,i}(u) - D_{g,i}(u)|}{0.2\, D_{g,i}(u)}, 0, 1\right). \tag{2}$$

All other pixels receive zero. Agreement decreases linearly to zero at 20% relative depth error.

Dense support. Let $Q_i = \sum_{u\in\Omega_i} q_i(u)$. Geometric–Photometric Consistency averages support per view, $\mathrm{GPC} = |\mathcal{J}|^{-1} \sum_{i\in\mathcal{J}} \frac{Q_i}{|\Omega_i|}$, while Integrated Consistency Mass pools pixels, $\mathrm{ICM} = \frac{\sum_{i\in\mathcal{J}} Q_i}{\sum_{i\in\mathcal{J}} |\Omega_i|}$. Both condition on successful densification. The failure-aware variant, $\mathrm{ICM}_{\mathrm{all}} = \frac{\sum_{i\in\mathcal{J}} Q_i}{\sum_{i\in\mathcal{T}} M_i}$, retains failed images in its denominator, where M_i is the image pixel count stored in the COLMAP database. Scores are zero when $\mathcal{J}$ is empty.

Viewpoint coverage. We penalize narrow viewing arcs using $\text{W-GPC} = \mathrm{GPC}_\omega = \mathrm{GPC}\,\omega/360°$. Here ω is $360°$ minus the largest circular azimuth gap between registered cameras around the median sparse-point center.

Controlled corruptions. On selected SysCON3D tests at $K = 21$, mean W-GPC decreases from 0.0117 for consistent sets to 0.0108 for single-outlier sets and 0.0049 for controlled mixtures (Table 19). Noise and repeated images score zero under the failure convention. The group means follow the expected ordering, although the single-outlier gap is small.

Human alignment. At $K = 9$, W-GPC achieves Spearman $\rho = 0.98$ with composite human rankings on both datasets (Table 4), compared with 0.24 for MEt3R and 0.48 for MASt3R-W-IMQ on DL3DV. We report W-GPC with ICM and registration rate, $|\mathcal{R}|/|\mathcal{T}|$, to distinguish limited coverage, weak dense support, and registration failure.

These scores measure verified support, not proof of consistency. COLMAP can fail on valid low-texture, repetitive, or low-overlap scenes, and scores become less discriminative when all outputs are poor. MASt3R-W-IMQ remains a secondary learned diagnostic.

Table 4: COLMAP based metrics vs. human rankings of 8 NVS methods on DL3DV (24 scenes) and MipNeRF360 (9 scenes) for the $K=9$ view split. Weighted GPC measures geometric photo-consistency weighted by angular coverage;ICM is Integrated Consistency Mass, a pixel-weighted mean of dense support.; Ang. Cov. measures angular coverage of registered views. Colors encode ranks 1 (best, green) to 8 (worst, red); rows sorted by DL3DV human rank (**H**). Bottom row reports Spearman correlation ρ between human and metric-induced method rankings.

Method	DL3DV				MipNeRF360			
	H↓	W-GPC↑	ICM↑	Ang. Cov.↑	H↓	W-GPC↑	ICM↑	Ang. Cov.↑
DepthSplat	1	0.288 **1**	0.531 **1**	170.4 **1**	1	0.475 **1**	0.636 **1**	247.0 **1**
Long-LRM	2	0.136 **2**	0.404 **2**	94.1 **2**	2	0.128 **2**	0.378 **2**	79.3 **2**
S-V-Camera	3	0.056 **3**	0.351 **3**	53.4 **3**	4	0.021 **3**	0.363 **3**	21.1 **5**
ViewCrafter	4	0.036 **4**	0.238 **7**	46.2 **4**	3	0.017 **4**	0.275 **5**	25.5 **3**
MVGenMaster	5	0.035 **5**	0.302 **6**	42.2 **5**	5	0.015 **5**	0.267 **6**	25.2 **4**
MVSplat360	6	0.027 **6**	0.319 **5**	27.2 **7**	6	0.013 **6**	0.356 **4**	13.5 **7**
NVS-Solver	7	0.018 **8**	0.350 **4**	18.9 **8**	8	0.003 **8**	0.142 **8**	13.1 **8**
Difix3D	8	0.019 **7**	0.237 **8**	32.6 **6**	7	0.010 **7**	0.203 **7**	16.1 **6**
Rank Corr. (ρ)↑	–	**0.98**	0.76	0.93	–	**0.98**	0.88	0.95

6 Discussion

Our results highlight a reliability issue in multiview 3D consistency evaluation: neural metrics can score inconsistent inputs as highly consistent and disagree substantially with human rankings. SysCON3D shows that these failures are systematic, rather than isolated examples. Our backbone-residual-aggregation analysis yields stronger neural metrics, but the improvements do not resolve the problem. Even our best variant still scores Gaussian noise as more consistent than real multiview observations. The analysis identifies a deeper issue: VGGT, DUSt3R, MASt3R, and Fast3R can hallucinate geometry and cross-view support for unrelated images and noise. Downstream metrics then compare features at unsupported correspondences and can mistake their similarity for 3D consistency. **This is not simply an aggregation artifact: the metric is evaluating matches that the images do not support.** This motivates our COLMAP-based alternative. COLMAP is a tad slower and can fail on valid low-texture, specular, repetitive, or low-overlap imagery. However, rejected matches, failed registration, missing dense maps, and limited coverage provide explicit, inspectable signals. Our metrics convert these signals into consistency scores and achieve the strongest alignment with composite human rankings in our experiments. A failed reconstruction indicates missing verified support, not proof that the images are inconsistent. These findings argue for caution when learned reconstructions or favorable metric scores are treated as geometric evidence, particularly with noisy or uncurated inputs. This matters in applications such as forensic reconstruction and autonomous driving, where unsupported geometry could be mistaken for fact. Geometric verification and human review remain important safeguards. A reliable learned alternative must do more than produce geometry: it must recognize when the input images do not support its predictions. **Models used as evaluators need to know their limitations.**

AI Use Statement

Large Language Models (LLMs) were utilized strictly for grammatical corrections and language refinement during the drafting of this manuscript. They did not contribute to the conceptualization or ideation of the research.

Ethics Statement

Our research studies the robustness of multiview 3D consistency metrics and the risk of hallucinated geometric support in learned reconstruction models. We advocate responsible use of these technologies and caution against treating generated reconstructions or favorable metric scores as verified evidence, particularly in safety-critical applications. The limitations and recommended safeguards

are discussed in section 6, and the human evaluation protocol is described in section F. We condemn the use of these technologies to create misleading or malicious content.

Reproducibility Statement

We define the neural metric family in Section 4.1, the COLMAP-based scores in Section 5, and the construction and evaluation of SysCON3D in Section 3. Additional method details, the human evaluation protocol, adaptations of novel view synthesis methods, and COLMAP metric details are provided in sections E to G and L. The associated source code, benchmark construction scripts, and SysCON3D data will be made publicly available upon acceptance of the paper.

References

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 3

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5470–5479, 2022. 3, 6

Chenjie Cao, Chaohui Yu, Shang Liu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Mvgenmaster: Scaling multi-view generation from any image via 3d priors enhanced diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6045–6056, 2025. 3, 6

Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. *Advances in Neural Information Processing Systems*, 37:107064–107086, 2024. 3, 6, 34

Shivam Duggal, Yushi Hu, Oscar Michel, Aniruddha Kembhavi, William T. Freeman, Noah A. Smith, Ranjay Krishna, Antonio Torralba, Ali Farhadi, and Wei-Chiu Ma. Eval3d: Interpretable and fine-grained evaluation for 3d generation. *CVPR*, 2025. 3

Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981. ISSN 0001-0782. doi: 10.1145/358669.358692. URL https://doi.org/10.1145/358669.358692. 3

Stephanie Fu, Mark Hamilton, Laura E. Brandt, Axel Feldmann, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=GkJiNn2QDF. 2, 3

Mark E Glickman and Albyn C Jones. Rating the chess rating system. 7, 21

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(null):723–773, March 2012. ISSN 1532-4435. 3

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974. doi: 10.1080/01621459.1974.10482962. URL https://doi.org/10.1080/01621459.1974.10482962. 3

Jisang Han, Sunghwan Hong, Jaewoo Jung, Wooseok Jang, Honggyu An, Qianqian Wang, Seungryong Kim, and Chen Feng. Emergent outlier view rejection in visual geometry grounded transformers. *arXiv preprint arXiv:2512.04012*, 2025. 1, 6, 34

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm. 3

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1): 73–101, March 1964. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732. 3

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 3

Jack Langerman, Denys Rozumnyi, Yuzhong Huang, and Dmytro Mishkin. Explaining human preferences via metrics for structured 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26944–26953, 2025. 21

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024. 2, 3, 7

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024. 6

Shalini Maiti, Lourdes Agapito, and Filippos Kokkinos. Gen3deval: Using vllms for automatic evaluation of generated 3d objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18552–18562, 2025. 3

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2, 3

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3, 7, 39

Daniel J Simons and Christopher F Chabris. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *perception*, 28(9):1059–1074, 1999. 21

Saar Stern, Ido Sobol, and Or Litany. Appreciate the view: A task-aware evaluation framework for novel view synthesis. *arXiv preprint arXiv:2511.12675*, 2025. 3

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3, 7

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2, 3, 7

Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26024–26035, 2025a. 2, 3, 6

Sibo Wu, Congrong Xu, Binbin Huang, Geiger Andreas, and Anpei Chen. Genfusion: Closing the loop between reconstruction and generation via videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b. 3

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16453–16463, 2025. 1, 3, 6

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025. 2, 3, 7

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 767–783, 2018. 3

Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. *arXiv preprint arXiv:2405.15364*, 2024. 3, 6

Jason J Yu, Fereshteh Forghani, Konstantinos G Derpanis, and Marcus A Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7094–7104, 2023. 3, 6

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 2, 3, 6

Jensen (Jinghao) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025. 3, 6

Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4349–4359, 2025. 3, 6

APPENDIX CONTENTS

A Notation and Common Terms **14**

B FAQs **16**

C Detailed Contributions **19**

D How to use our metrics to evaluate multiview 3D consistency **19**

E Linear-Time PC Residuals and Two-Sample MMD **20**

Linear-time PC residuals. 20

Two-sample MMD with a reference set. 20

F Human Evaluation Study: Implementation Details **20**

Platform and interface. 20

Matchup sampling and session flow. 20

Voting and Elo updates. 21

Rubric examples and logging. 21

Statistics 21

Note. 21

G Implementation Details: Adapting NVS Methods to Sparse-View Orbits **22**

G.1 Common: Orbital Trajectory from Sparse Cameras **22**

G.2 DepthSplat. **22**

G.3 MVSplat360. . . . . . **22**

G.4 Long-LRM. **23**

G.5 Difix3D **23**

G.6 MVGenMaster **23**

Monocular depth alignment. 23

G.7 NVS Solver **23**

Multi-view depth fusion. 23

G.8 Stable Virtual Camera (SEVA) **24**

G.9 ViewCrafter **24**

Multi-reference image conditioning. 24

Single-pass orbit generation. 24

H Reconstruction Behavior on SysCON3D Benchmark **24**

H.1 Diagnostics, definitions, and main patterns **24**

H.2 What each diagnostic reveals **25**

Reconstruction geometry. 25
Backbones hallucinate overlap on cross-scene pairs. 25
Residual separation after overlap is hallucinated. 26
Confidence and ghost mass. 27
H.3 Implications for robust evaluation 27
H.4 Signature Hallucination Patterns 34
H.5 Hallucinated correspondences on real NVS output 34

I Detailed Robustness Analysis 34
RobustVGGT metric. 34
RobustVGGT behavior. 34
VGGT PC-MMD on Gaussian noise. 34
SIFT-SED on Gaussian noise. 37
Single-outlier detection. 37
Per-module interaction effects. 37
Ranking with explicit outlier rejection. 37

J Rank Concordance Analysis 37
Kendall's τ. 37
Probabilistic pairwise concordance (PPC). 37
Results. 38

K Evaluation of Parametric Metric Family on MipNeRF360 39

L COLMAP-based 3D Consistency Metrics: Extended Notation and Coverage 39
L.1 Motivation: why classical geometric verification 39
L.2 Notation and COLMAP outputs 39
L.3 Angular coverage 40
Compute cost. 40
L.4 COLMAP failures 40
L.5 COLMAP Metric Results 41

A Notation and Common Terms

We use the same notation as in the main text. Neural metrics report *inconsistency* scores, where lower is better; COLMAP-based metrics report *consistency* scores, where higher is better. Both families evaluate support for a single static, rigid scene.

Symbol / Term	Meaning
Inputs and benchmark	
$\mathcal{I} = \{I_i\}_{i=1}^{K}$	RGB image set evaluated by the generic neural metric formulation.
K	Number of images in the generic metric formulation and in each SysCON3D set. In NVS experiments, K instead denotes the number of source views supplied to the NVS method; the full COLMAP input count is $\lvert\mathcal{T}\rvert$.
$i, j;\ u, v;\ p;\ a, b$	Indices for views; pixel locations; reconstructed points; and residual samples, respectively.
Ω_i^{img}	Pixel grid of image I_i in the neural formulation, distinguished from the COLMAP dense-map grid Ω_i.
SysCON3D	Systematically Controlled Outlier and Noise 3D benchmark: a controlled corruption protocol built from existing Mip-NeRF360 scenes to diagnose multiview consistency metrics (Section 3).
L_0	Consistent set containing K distinct views of one scene.
L_1	Single-outlier set: $K-1$ views from one scene and one foreign view. Exactly $2/K$ of its unordered view pairs are cross-scene.
L_2	Controlled mixture with approximately 30% foreign views. The cross-scene pair fraction depends on the number of views assigned to each source scene.
L_3	Random mixture with views sampled independently from nine scenes. Uniform scene sampling gives an expected cross-scene pair fraction of $8/9 \approx 89\%$.
SysCON3D-M	Cross-scene mixture subset: L_1, L_2, and L_3.
SysCON3D-N	Core noise subset: patched Gaussian and full Gaussian noise. The noise is sampled from $\mathcal{N}(0.5, 0.2^2)$ and clipped to $[0, 1]$; patched Gaussian replaces four patches per image.
Additional diagnostics	Uniform noise and repeated identical images are studied separately from the five-corruption, 35-comparison robustness evaluation. Repeated images provide degenerate multiview evidence, but are not necessarily inconsistent with a static scene.
Neural metric family	
$B, \psi, \mathcal{A}$	Reconstruction backbone, residual construction, and aggregation function in Equation (1). The backbone supplies correspondences; the residual compares features at those correspondences.
$\phi,\ F_i$	Frozen image-feature extractor and its dense feature map $F_i = \phi(I_i)$, using DINOv2 with FeatUp.
$\mathbf{x}_p,\ \mathbf{x}_{i,u},\ \pi_i$	Reconstructed point indexed by p; reconstructed point associated with pixel u in view i; and projection into view i.
$r_{ij}(u, v)$	Cosine dissimilarity between $F_i(u)$ and $F_j(v)$. This measures feature disagreement, not depth error.
$\mathcal{V}_{i\to j}$	Pixels in Ω_i^{img} accepted by the depth-consistency test when warping from view i to view j.
$\mathcal{S}_p,\ s_p$	Views observing reconstructed point $\mathbf{x}_p$, and their number $s_p = \lvert\mathcal{S}_p\rvert$. Point-consistency residuals require $s_p \geq 2$.
W / PC	Warp-based residuals compare features at geometry-induced pixel correspondences. Point-consistency residuals average feature disagreement across the views observing each reconstructed point.
e_p^{PC}	Mean pairwise feature dissimilarity for point $\mathbf{x}_p$ across its observing views $\mathcal{S}_p$.
$\mathcal{E} = \{e_a\}_{a=1}^{N}$	Collection of N scalar residuals, retaining repeated values. $\mathcal{E}_{\mathrm{warp}}$ and $\mathcal{E}_{\mathrm{PC}}$ denote the two residual constructions.
$P,\ \delta_0$	Empirical residual distribution $P = N^{-1}\sum_{a=1}^{N}\delta_{e_a}$ and the ideal distribution concentrated at zero. Here δ_{e_a} is a Dirac mass at e_a.
κ	Kernel used for MMD aggregation. We use RBF and inverse multiquadric (IMQ) kernels.
Base / MMD / IMQ / Energy	Mean residual; RBF-kernel MMD^2; IMQ-kernel MMD^2; and energy distance, respectively. Distributional scores compare P with δ_0.
Metric names	Backbone–residual–aggregation. For example, MASt3R-W-IMQ uses MASt3R, warp-based residuals, and IMQ-MMD aggregation.

Symbol / Term	Meaning (continued)
COLMAP-based metrics	
$\mathcal{T},\ \mathcal{R},\ \mathcal{J}$	Attempted images, registered images, and registered images with readable photometric and geometric depth maps, respectively: $\mathcal{J} \subseteq \mathcal{R} \subseteq \mathcal{T}$.
$D_{p,i},\ D_{g,i}$	Photometric and geometric depth maps from COLMAP PatchMatch passes without and with geometric consistency, respectively. Both use the same SfM reconstruction and scale.
Ω_i	Full dense-map pixel grid for view i, including invalid pixels.
M_i	Image pixel count stored in the COLMAP database, used in the attempted-set denominator of $\mathrm{ICM}_{\mathrm{all}}$. It need not equal the dense-map pixel count $\lvert\Omega_i\rvert$.
$q_i(u)$	Per-pixel depth-agreement score from Equation (2). It lies in $[0, 1]$ and is zero for nonfinite depths, $D_{g,i}(u) \leq 10^{-5}$, or relative depth disagreement of at least 20%.
Q_i	Summed agreement over view i: $Q_i = \sum_{u \in \Omega_i} q_i(u)$.
GPC	Geometric–Photometric Consistency: $\mathrm{GPC} = \lvert\mathcal{J}\rvert^{-1} \sum_{i \in \mathcal{J}} Q_i / \lvert\Omega_i\rvert$. Each densified view receives equal weight.
ICM	Integrated Consistency Mass: $\mathrm{ICM} = \frac{\sum_{i \in \mathcal{J}} Q_i}{\sum_{i \in \mathcal{J}} \lvert\Omega_i\rvert}$. This is a pixel-weighted mean of reliable dense support, conditional on available depth maps, not a count or volume of reconstructed 3D points.
$\mathrm{ICM}_{\mathrm{all}}$	Attempted-set normalization: $\mathrm{ICM}_{\mathrm{all}} = \frac{\sum_{i \in \mathcal{J}} Q_i}{\sum_{i \in \mathcal{T}} M_i}$. Failed images contribute zero support but remain in the denominator.
ω	Angular coverage of registered camera centers around the median sparse-point center, measured as 360° minus the largest circular azimuth gap.
W-GPC / GPC_ω	Coverage-weighted GPC: $\mathrm{GPC}_\omega = \mathrm{GPC}\,\omega/360^\circ$. Our primary COLMAP-based score combines dense agreement with viewpoint coverage.
Registration rate	Fraction of attempted images registered by SfM: $\lvert\mathcal{R}\rvert/\lvert\mathcal{T}\rvert$.
Failure convention	When $\mathcal{J}$ is empty, the dense-support scores are zero. Otherwise, GPC and ICM average over $\mathcal{J}$, while $\mathrm{ICM}_{\mathrm{all}}$ normalizes over all attempted images.
Robustness, diagnostics, and human alignment	
Cohen's d	Standardized difference between corrupted and consistent group means: $d = (\bar{m}_g - \bar{m}_{L_0})/s_{\mathrm{pooled}}$, where s_{pooled} is the pooled within-group standard deviation. Scores are oriented so that larger means worse; positive d then indicates correct separation.
Win rate	Fraction of comparisons in which a corrupted group receives a worse mean score than L_0. Per-corruption rates use seven view counts; the overall rate uses $5 \times 7 = 35$ comparisons. This is not individual-set classification accuracy.
$\mathrm{FNR}_{\mathrm{cross}}$	Cross-scene overlap hallucination rate: $\Pr(\text{predicted overlap} \mid \text{cross-scene pair})$. It measures missed rejection of cross-scene pairs; the ideal value is zero. It is distinct from the zero-overlap rate over all pairs.
Ghost mass	Fraction of rendered pixels supported by wrong-scene points in the Fast3R and VGGT diagnostics. This measures rendered support from incorrect scene provenance, not physical 3D mass.
τ	Kendall rank agreement between the metric's ordering of corruption-group mean scores and the prescribed benchmark ordering.
ρ	Spearman correlation between metric-induced and human method rankings. It is not the neural residual function, which is denoted by ψ.
Human rankings	Composite rankings from weighted Elo aggregation of pairwise votes on 3D consistency, visual realism, and plausibility, with weights $2:1:1$. Reported human-alignment correlations compare rankings across eight NVS methods.

B FAQs

1. **What does SysCON3D test?**
   SysCON3D is a controlled stress test for 3D consistency metrics. It is not intended to model the full distribution of real NVS failures. Instead, it tests a basic reliability property: image

sets with weaker or contradictory evidence for a shared 3D scene should receive worse consistency scores than clean multiview sets. The corruptions are deliberately simple so that the expected ordering is known.

2. **Are cross-scene mixtures and Gaussian noise too artificial?**
   They are diagnostic rather than naturalistic corruptions. Their purpose is to test whether a metric can reject inputs that should not support a coherent multiview reconstruction. This is analogous to using controlled corruptions to expose robustness failures before trusting a metric on real generated outputs. Further, these inputs can occur, in practice, in noisy input situations, for example data scraped from the internet, or adversarial setups.
3. **Why include repeated views if they may come from the same scene?**
   Repeated views are not necessarily inconsistent with the scene identity, but they are degenerate as multiview evidence. A metric for multiview 3D consistency should not treat repeated or near-identical frames as strong evidence for a well-supported 3D reconstruction. We therefore use repeated views as a degenerate diagnostic rather than as a cross-scene corruption. Further, we have observed such outputs in NVS methods when they failed to generalize to the input.
4. **What do we mean by hallucinated geometric support?**
   We use this term when a reconstruction backbone produces nontrivial 3D structure, overlap, or correspondences for inputs that do not provide valid multiview evidence, such as unrelated scenes or random noise. The issue is not merely that the reconstruction looks visually odd; it is that downstream metrics then compute residuals on unsupported correspondences.
5. **Is the failure caused by aggregation or by the reconstruction backbone?**
   Both matter, but our results show that the hardest failures often originate before aggregation. Distributional aggregation improves robustness over the MEt3R-style mean, showing that metric design matters. However, if the backbone has already hallucinated unsupported geometry, the residuals are computed on an invalid set of correspondences, which no scalar aggregation can fully repair.
6. **Why does MASt3R-W-IMQ improve robustness but still fail on noise?**
   MASt3R-W-IMQ uses distributional aggregation and is more sensitive to outliers and residual-distribution shape than a mean score. This improves separation for cross-scene mixtures. On Gaussian noise, however, learned backbones can still produce dense geometry, and feature residuals can become uniformly small or uninformative. In that setting, the residual distribution itself lacks the signal needed for aggregation to recover the correct ordering.
7. **Why use DINOv2/FeatUp features if they can fail on cross-scene matches?**
   We use DINOv2/FeatUp because they are part of the MEt3R-style evaluation pipeline and provide strong dense semantic features. Our analysis shows a limitation of this choice for 3D consistency: semantic similarity is not the same as physical point identity. Visually similar regions from different scenes can yield low residuals even when the underlying correspondence is geometrically invalid.
8. **Does COLMAP simply win because it fails to reconstruct hard cases?**
   No. In our COLMAP-based metrics, failed registration or densification is counted as zero verified 3D support, not as a missing or ignored value. This makes failure part of the consistency score. The metric therefore measures verified reconstructability under classical multiview geometry: registered cameras, dense support, depth agreement, and viewpoint coverage.
9. **Is COLMAP a perfect 3D consistency metric?**
   No. COLMAP can fail on low-texture, specular, repetitive, blurred, or low-overlap imagery even when the scene is physically valid. We therefore interpret COLMAP-based metrics as failure-aware geometric verification metrics, not as universal ground truth. Their value is that failure modes are observable and measurable, unlike learned backbones that may always return geometry.
10. **Why report W-GPC, ICM, and registration rate together?**
    W-GPC combines dense geometric–photometric agreement with angular coverage and is our primary COLMAP-based score. ICM measures the total amount of reliable dense 3D

mass. Registration rate reports sparse geometric support. Reporting them together separates different failure modes: insufficient sparse support, poor dense agreement, low 3D mass, and narrow viewpoint coverage.

11. **When should MASt3R-W-IMQ be used instead of COLMAP-based metrics?**
MASt3R-W-IMQ is useful as a fast learned diagnostic, especially when COLMAP cannot be run or when inputs are known to be clean and free of outliers. It should not override a complete COLMAP failure. If COLMAP finds little geometric support but a learned metric reports high consistency, that sample should be flagged for human review. Also, if a robust 3D reconstruction backbone becomes available in future, MASt3R-W-IMQ will be a viable and sufficient metric.

12. **Are PRISM, SED, and TSED directly comparable to our metrics?**
They are included as external baselines with different assumptions. PRISM uses a learned object-centric embedding distribution and an anchor set. SED/TSED use explicit epipolar verification from SIFT correspondences and require known or estimated fundamental matrices. We include them to contextualize robustness, but our main comparison is among scene-level, ground-truth-free multiview consistency metrics.

13. **Why evaluate human alignment over method-level rankings?**
The human study asks which NVS methods produce more 3D-consistent outputs, so we compare metric-induced method rankings to human method rankings. The correlations are therefore computed over $n = 8$ methods. This is a compact validation of whether metrics agree with human judgments at the level used for model comparison; it is not a per-frame perceptual-quality study.

14. **Why separate 3D consistency, realism, and plausibility in the human study?**
A generated video can be realistic but geometrically inconsistent, or geometrically stable but visually imperfect. We ask participants to judge these axes separately so that metric alignment is evaluated against the intended construct: scene-level 3D consistency rather than general visual appeal.

15. **Do these results mean learned 3D backbones are poor reconstruction models?**
No. The result concerns their use as automatic evaluators or sources of geometric evidence under corruptions. Learned backbones are powerful reconstruction models and can support useful metrics when paired with better residual aggregation. The failure arises when they return confident-looking geometry for inputs that provide little or no valid multiview support.

16. **What is the main practical recommendation?**
For robust evaluation of NVS outputs, we recommend reporting W-GPC with ICM and registration rate. MASt3R-W-IMQ should be reported as a complementary learned diagnostic. Large disagreement between COLMAP-based verification and learned-backbone metrics is itself informative and should trigger inspection, especially in noisy or uncurated settings.

17. **Why is hallucinated geometry a problem?**
Learned reconstruction models can use priors to complete geometry when the image evidence is weak. This is useful for reconstruction, but it becomes risky when the same model is used as an evaluator. If the backbone invents 3D structure or cross-view overlap, a metric built on that backbone will compute residuals on unsupported correspondences and may score an invalid view set as consistent. This can distort NVS benchmark rankings and can be dangerous in applications where geometry is treated as evidence, such as forensic or crime-scene reconstruction, medical reconstruction, autonomous driving, and robotics. In such settings, learned 3D predictions should be checked with geometric verification, calibrated uncertainty, or human review.

18. **Why not apply RANSAC-style geometric checks to matches predicted by VGGT, DUSt3R, or MASt3R?**
Classical matching separates evidence from verification: SIFT proposes discrete matches, and RANSAC tests whether they support a geometry. Learned 3D backbones predict geometry, cameras, and correspondences together, so their matches are already tied to the model's own reconstruction. If the model hallucinates a plausible scene, a geometric check can end up validating that internal reconstruction rather than independent image evidence. COLMAP gives a separate verification signal through matches, inliers, registration, dense

support, and viewpoint coverage; when these fail, the view set lacks verified multiview consistency.

C Detailed Contributions

1. We introduce a **unified parametric framework** for ground-truth-free 3D consistency metrics, showing that existing and new neural metrics can be decomposed into three components: a *reconstruction backbone*, a *residual function*, and an *aggregation function*. This formulation clarifies the design space of multi-view consistency metrics and enables principled variants beyond pairwise mean aggregation.
2. We construct **SysCON3D**, a robustness benchmark for 3D consistency evaluation, which systematically injects controlled cross-scene outliers and synthetic corruptions into multi-view image sets. This benchmark enables direct testing of whether a metric can distinguish geometrically consistent scenes from increasingly inconsistent ones.
3. Through SysCON3D, we uncover a **systematic and previously underappreciated failure mode** of modern data-driven 3D reconstruction backbones: rather than rejecting impossible inputs, they hallucinate non-trivial 3D structure and spurious cross-view consistency for cross-scene mixtures, repeated images, and random noise. Consequently, evaluation metrics built on top of these learned backbones can inherit these biases and become unreliable.
4. To address this limitation, we develop **robust COLMAP-based 3D consistency metrics** that avoid learned priors and instead rely on classical geometric verification. These metrics provide a more interpretable, failure-aware, and robust measure of scene-level multi-view consistency.
5. We design a **structured human preference study** for evaluating 3D consistency, with explicit protocols that distinguish 3D consistency from visual realism and plausibility. This yields a more targeted human reference for assessing how well automatic metrics align with human judgments of scene-level consistency.
6. We perform a **comprehensive empirical evaluation** on SysCON3D, Mip-NeRF360, and DL3DV, together with our human study, and show that the proposed metrics substantially improve robustness over prior work while also aligning more closely with human judgments. In particular, neural distributional metrics improve over MEt3R, and COLMAP-based metrics achieve the strongest human alignment among the evaluated metrics in our structured analysis.

D How to use our metrics to evaluate multiview 3D consistency

We recommend using both our classical and neural metrics in conjunction, as they provide a complementary analysis of the multiview data being evaluated.

We recommend **W-GPC** as the primary COLMAP-based metric because it combines dense geometric-photometric agreement with viewpoint coverage. When all compared methods are evaluated on the same target views, ICM is a useful complementary measure of total reliable 3D mass and should be reported alongside W-GPC (or used to scale W-GPC). Very low registration rate typically indicates that the generated view set is not reconstructable under classical multi-view geometry, in which case COLMAP-based scores naturally approach zero. In such cases, MASt3R-W-IMQ may be reported as a secondary diagnosis, but it should not override a COLMAP failure, as learned backbones can assign spuriously favorable scores to noisy or inconsistent inputs.

Notably, MASt3R-W-IMQ can still be used for evaluation if we are certain that the input images contain no noise or outliers, or if the COLMAP pipeline cannot be run for any unusual reason. Further, if a new robust 3D reconstruction backbone becomes available that does not hallucinate 3D like VGGT et al., MASt3R-W-IMQ becomes a more robust measure. In practice, the registration rate is the threshold we use to decide whether to fall back on neural metrics: a low registration rate usually means the input views are largely or completely 3D-inconsistent, and COLMAP scores tend to zero, in which case the neural variant can still be reported, with the caveat above.

E Linear-Time PC Residuals and Two-Sample MMD

Linear-time PC residuals. Once geometry and visibility are available, PC residuals require one feature lookup per observing view. For a point $\mathbf{x}_p$ observed in views $\mathcal{S}_p$, let $s_p = |\mathcal{S}_p| \geq 2$ and $f_{i,p} = F_i(\pi_i(\mathbf{x}_p))$. For nonzero features, define $\hat{f}_{i,p} = f_{i,p}/\|f_{i,p}\|_2$ and accumulate $\mathbf{z}_p = \sum_{i\in\mathcal{S}_p} \hat{f}_{i,p}$. Expanding the squared norm gives $\sum_{i\neq j}\langle \hat{f}_{i,p}, \hat{f}_{j,p}\rangle = \|\mathbf{z}_p\|_2^2 - s_p$. For exact cosine dissimilarity, without the numerical stabilizer, the PC residual in Section 4.1 is therefore

$$e_p^{\mathrm{PC}}\big|_{\varepsilon=0} = 1 - \frac{\|\mathbf{z}_p\|_2^2 - s_p}{s_p(s_p-1)}. \tag{3}$$

This avoids enumerating view pairs and reduces per-point residual computation from $O(s_p^2)$ to $O(s_p)$ for fixed feature dimension. The identity assumes exact unit normalization; it does not hold unchanged when ε is added to the product of feature norms.

Two-sample MMD with a reference set. The aggregation in Section 4.1 compares the residual distribution P with δ_0, the ideal distribution concentrated at zero. When ground-truth images $\mathcal{I}^{\mathrm{gt}}$ are available, we can instead compare against reference residuals obtained using the same backbone, feature extractor, and residual construction.

Let P and Q be the empirical distributions of generated residuals $\{e_a\}_{a=1}^N$ and reference residuals $\{e'_c\}_{c=1}^M$. For $N, M \geq 2$, the off-diagonal MMD estimator gives a reference-based discrepancy score:

$$\widehat{\mathrm{MMD}}_\kappa^2(P,Q) = \frac{\sum_{a\neq b}\kappa(e_a,e_b)}{N(N-1)} - \frac{2}{NM}\sum_{a=1}^{N}\sum_{c=1}^{M}\kappa(e_a,e'_c) + \frac{\sum_{c\neq d}\kappa(e'_c,e'_d)}{M(M-1)}. \tag{4}$$

The finite-sample estimate can be negative. This expression applies to both RBF and IMQ kernels; energy distance also admits a two-sample extension.

F Human Evaluation Study: Implementation Details

This section provides implementation details for the pairwise human evaluation described in section 4.3. The study comprises 1310 pairwise comparisons from 20 participants, including 2 experts in sparse-view 3D reconstruction. For each view-count setting $K \in 3, 6, 9$, participants compare two anonymized videos from the same scene and vote on three axes: 3D consistency, visual realism, and plausibility. *3D consistency* asks whether the rendered views look like they all come from a single coherent 3D scene rather than the scene morphing or drifting as the camera moves. *Visual realism* asks whether each individual frame looks like a real, sharp photograph rather than a blurry, artifact-ridden, or AI-generated image. *Plausibility* asks whether the synthesized frames at novel views and input observations appear to come from the same 3D scene. We aggregate these votes using a weighted Elo protocol, assigning 2 points to 3D consistency and 1 point to each of the other two axes.

Platform and interface. The study uses a static GitHub Pages frontend, a Cloudflare Worker backend, a D1 (SQLite) database, and Cloudflare R2 for input images and videos. Participants see the K input views, along with two anonymized orbit videos labeled A and B (fig. 4); method names are never exposed, and asset URLs use obfuscated keys. The frontend keeps both videos synchronized, loops them automatically, offers shared speed controls, and speeds up long clips to keep each comparison short. A manual with rubric examples is available during the session.

Matchup sampling and session flow. Sampling is coverage-biased at two levels. For a chosen view count $K \in \{3, 6, 9\}$, the backend first prefers scenes with the fewest completed games. Within that scene, it prefers methods with the fewest prior appearances, excludes rubric pairs and pairs already seen by the same participant, and randomizes the left-right assignment. After each vote, the winner enters a winner-stays block of at most three rounds: the same scene and K are kept, the winner stays on the same side, and a new challenger is drawn from the least-played unseen methods for that scene. Participants can break the block with `Skip` or `New pair`; `Skip` clears the block and does not log a game. When a block ends, the frontend rotates to the next K value to spread effort across 3, 6, and 9 views. Only after a participant exhausts all unique pairs at a given K does the backend allow repeats.

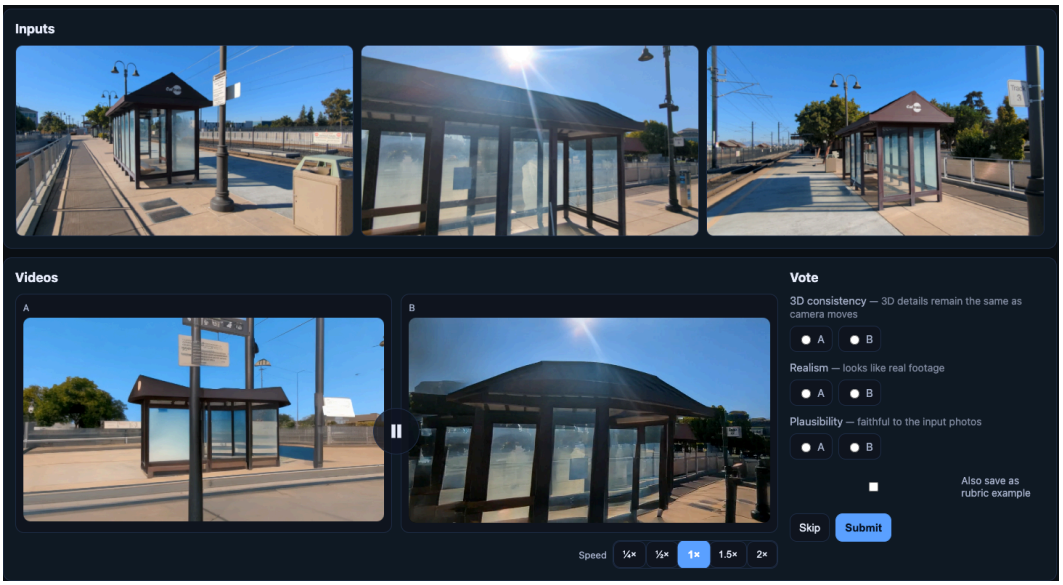


Figure 4: **Human-evaluation interface.** Participants see the K input views (top) alongside two anonymized 360° orbit videos (A/B) played in a synchronized loop with shared speed controls. Method names are never revealed. Each comparison requires forced A/B votes on three axes – 3D consistency, visual realism, and plausibility (no per-axis ties). Method ratings are aggregated with a weighted Elo protocol, giving 3D consistency $2\times$ weight and the other two axes $1\times$ each. After each vote, the winner stays on the same side for up to three additional rounds against new challengers within the same scene/K, mainly to reduce cognitive load on the participants. The study collected 1310 pairwise comparisons from 20 participants (including 2 experts in sparse-view 3D reconstruction) across $K \in \{3, 6, 9\}$ input views on DL3DV and Mip-NeRF360 scenes.

Voting and Elo updates. Each comparison requires forced A/B choices on three axes: 3D consistency, realism, and plausibility. There are no per-axis ties. The Elo outcome is binary: the side that wins at least two of the three axes wins (and stays on for up to 3 rounds), and the other side loses, so the game itself has no draw state. However, when updating ratings for the participating methods, we assign $2\times$ weight to 3D consistency and $1\times$ weight to realism and plausibility, so ties are possible in the final scoring system. Ratings follow standard Elo (Glickman & Jones): given current ratings R_A and R_B, the expected score is $E_A = 1/(1 + 10^{(R_B - R_A)/400})$, and ratings update as $R'_A = R_A + 32 \cdot (S_A - E_A)$. All methods start with a base rating of 500. Dataset-specific leaderboards are computed by replaying the logged games after filtering to the relevant dataset.

Rubric examples and logging. Admins can run a separate rubric mode or promote an ordinary comparison to a rubric example, but only after writing reasons for all three axes. These examples populate the manual/examples panel shown to participants and are excluded from future matchup sampling to prevent calibration content from distorting coverage. All votes are timestamped, and participants can later review, edit, or delete their past submissions; leaderboards are then recomputed from the updated game log.

Statistics The study is based on 1310 games played across 20 participants, involving 2 experts in sparse-view 3D reconstruction. The top-4 methods for a specific dataset and view split are usually involved in 20-30% more games than the bottom-4 methods, as the winner of each game stays on for a maximum of 3 rounds to reduce cognitive load on participants.

Note. Human evaluation in our setting is subject to known perceptual biases. In particular, prior work suggests that observers can be disproportionately influenced by high-frequency structures such as edges and corners (Langerman et al., 2025), and may also miss important inconsistencies due to inattentional blindness (Simons & Chabris, 1999). As a result, annotators may occasionally underweight background 3D errors or over-prefer generations with sharper local detail despite poorer

global 3D consistency. To reduce this source of noise, we explicitly instructed participants to prioritize *3D consistency* and presented counterexamples highlighting such failure modes. Finally, we observe that human discrimination also degrades when all candidate generations are of very low quality: ranking among several poor 3D outputs becomes intrinsically error-prone, similar to the behavior of our COLMAP-based metrics under severely degraded inputs.

G Implementation Details: Adapting NVS Methods to Sparse-View Orbits

Adapting the 8 NVS methods (DepthSplat, MVSplat360, Long-LRM, Difix3D, MVGenMaster, NVS Solver, Stable Virtual Camera) to produce 360°orbital renderings from only $P \in \{3, 6, 9\}$ sparse input images required method-specific modifications, which we describe below. One component - the orbital camera trajectory - is shared across all methods and is presented first.

G.1 Common: Orbital Trajectory from Sparse Cameras

All methods follow the same procedure to convert a sparse set of calibrated cameras into a smooth 360°orbital path. Given camera centers $\{c_i\}$ and unit forward directions $\{v_i\}$, we first estimate a shared look-at point p^* by minimizing the sum of squared perpendicular distances from each camera ray:

$$p^* = \arg\min_p \sum_{i=1}^{P} \left\| (I - v_i v_i^\top)(p - c_i) \right\|^2 . \tag{5}$$

Setting the gradient to zero gives the 3×3 linear system

$$\underbrace{\left(\sum_{i=1}^{P} I - v_i v_i^\top \right)}_{A} p^* \;=\; \underbrace{\sum_{i=1}^{P} (I - v_i v_i^\top)\, c_i}_{b}, \tag{6}$$

which we solve as $p^* = A^{-1}b$ (or via least-squares when A is rank-deficient). We then estimate an up direction u from the mean camera orientation, compute a median orbital radius r and height h from the camera positions projected onto the plane perpendicular to u, and generate N evenly spaced orbit poses via Rodrigues rotation of an initial planar direction $\hat{d}_0$ about u:

$$c(\theta) = p^* + r \left[\cos\theta\, \hat{d}_0 + \sin\theta\, (u \times \hat{d}_0) \right] + h\, u, \qquad \theta \in [0, 2\pi). \tag{7}$$

Each orbit camera looks back at p^*, producing a continuous azimuthal sweep even when the inputs cover a narrow angular range.

G.2 DepthSplat.

DepthSplat required no architectural or algorithmic modifications. Because the model is trained on randomly sampled context sets of varying size, it directly accepts 3, 6, or 9 input views and renders arbitrary target cameras - including the orbit trajectory - without retraining or fine-tuning. The only practical consideration is checkpoint selection: a model trained on 2-6 context views is used for the 3 and 6-view settings, while one trained on 4-10 views is used for 9 views, to match the training distribution.

G.3 MVSplat360.

MVSplat360 pairs a Gaussian-splatting encoder with a video-diffusion refiner (SVD) that enhances the rendered frames. Two modifications were needed to apply this refiner to a long orbit sequence.

First, the SVD model accepts a fixed window of $W = 28$ frames, whereas our orbits contain $T = 50$ frames. We divide the sequence into overlapping windows with a stride of $W - \delta$ ($\delta = 6$ frames of overlap). Each window is refined independently. Where two windows cover the same frame, we keep the prediction with the higher temporal-context score

$$s = n_{\text{back}} + 0.25\, n_{\text{fwd}}, \tag{8}$$

where n_{back} and n_{fwd} count preceding and following frames within that window, respectively. This favours predictions that have more causal context, matching the autoregressive nature of SVD. To ensure consistency in overlapping regions, the diffusion noise tensors are pre-sampled for the full T-frame sequence and sliced per window, so overlapping frames receive identical noise realisations.

Second, the classifier-free guidance scale is raised from the default value of 2.5 to 5.0. With only 3–9 input views, the Gaussian-splatting renderings may contain artifacts; a higher guidance scale steers the refiner closer to these renderings, reducing hallucination in unsupported regions.

G.4 LONG-LRM.

Long-LRM predicts a full set of 3D Gaussians in a single forward pass through a transformer/Mamba-2 backbone that was pre-trained on sequences of $M = 32$ views.

The main adaptation for sparse inputs is cyclic view repetition. Given N sparse input views, we construct a length-M input sequence by tiling:

$$\tilde{\mathbf{I}}_j = \mathbf{I}_{j \bmod N}, \quad j = 0, \ldots, M-1. \tag{9}$$

The corresponding poses and intrinsics are repeated in the same order. Because the model's self-attention aggregates information across tokens regardless of ordering, duplicate views do not degrade prediction quality. They instead reinforce the available observations. The predicted Gaussians are then rendered from each orbit camera in a single rasterisation pass.

G.5 DIFIX3D

Difix3D is a post-rendering diffusion model that removes artifacts from 3D Gaussian Splatting (3DGS) outputs by conditioning on a reference photograph. Its original pipeline renders novel views along an interpolated path between the training cameras and applies the diffusion fixer per frame, selecting the nearest training view as the reference. We replace the interpolated path with the orbital trajectory described above, so the same render-then-fix pipeline now produces full $360°$sequences from P inputs without other algorithmic changes.

G.6 MVGENMASTER

MVGenMaster is a multi-view diffusion model that conditions on reference images, their camera rays, and 3D positional encodings (3DPE) warped via metric depth maps. The original inference pipeline uses Depth Pro to obtain metric depth and supports chunked generation when the target sequence exceeds the model's context window. We adapt it to the sparse-view orbit setting with one key change.

Monocular depth alignment. The 3DPE module warps reference-view features into each target viewpoint using metric depth. In the original pipeline, Depth Pro provides metric depth directly. In our sparse-view benchmark, we instead use Depth Anything V2, which outputs disparity rather than metric depth. We align the disparity maps to world scale using the camera geometry: we set the inverse-depth scale factor as $s = r_{\text{med}}/\text{median}(1/d)$, where r_{med} is the median camera-to-lookat distance. This calibration gives the 3DPE module sufficient metric accuracy to produce plausible view warps from sparse inputs.

G.7 NVS SOLVER

NVS Solver is a training-free method that steers a pretrained video diffusion model toward geometric consistency using gradient-based guidance at each denoising step. The original pipeline supports single-image and two-image conditioning via depth-based warping with mask-weighted gradient refinement. We extend it to handle $P > 2$ input views.

Multi-view depth fusion. For each target orbital pose, we forward-warp *every* reference image into the target view using its depth map and known camera transformation, then fuse the reprojected

pixels by averaging over valid contributions:

$$\tilde{I}_t(x) = \frac{\sum_{i=1}^{P} m_{i\to t}(x)\, \Pi(I_i;\, D_i,\, T_{i\to t})(x)}{\sum_{i=1}^{P} m_{i\to t}(x) + \epsilon}, \tag{10}$$

where Π denotes depth-based warping and $m_{i\to t}(x)$ indicates pixel validity. Regions with no valid reprojection are flagged as holes via a dilated binary mask M_t. The fused image and mask are then passed through the existing gradient-refinement mechanism, originally designed for one or two source views. This converts the sparse-view extrapolation problem into a structured inpainting task without modifying the underlying denoising procedure.

G.8 Stable Virtual Camera (SEVA)

SEVA conditions a video diffusion model on Plücker ray embeddings and supports multi-view conditioning, two-pass generation with anchor frames, and distance-aware classifier-free guidance out of the box. We introduce a new task mode, `img2trajvid_m-prob`, that combines multi-view conditioning (as in SEVA's existing sparse-view task) with preset trajectory priors (as in its single-view trajectory task). Given P input views, we estimate the orbital camera path from their poses and feed it as the target trajectory, while the model's existing two-pass sampling and adaptive guidance handle the long-sequence generation without modification. No changes to the model or inference algorithm were needed.

G.9 ViewCrafter

ViewCrafter uses a video diffusion model conditioned on point-cloud renderings and an image embedding to synthesize novel views. The original model reconstructs a point cloud from DUSt3R, renders it under target cameras as a geometric prior, and conditions the diffusion model on a single reference image via cross-attention. We add multi-reference conditioning and an orbit rendering mode for sparse-view 360° generation.

Multi-reference image conditioning. The original model encodes a reference image into an embedding, which is concatenated with the text embedding for cross-attention. We extend this to P references by independently encoding each image into an embedding e_i and aggregating before injection. We support averaging ($e = \frac{1}{P}\sum_i e_i$), which produces a compact summary, and concatenation ($e = [e_1; \ldots; e_P]$), which preserves per-view detail at the cost of a longer attention context. Both modes require no architectural changes and keep all observed views in context throughout generation, which is especially important for viewpoints far from any single input on the orbit.

Single-pass orbit generation. ViewCrafter renders the point cloud under the full orbital trajectory and runs a single diffusion pass over all frames, which is both more efficient and more temporally consistent than the chunked generation used by methods like SEVA. Chunking into fixed-length clips re-runs diffusion independently per chunk, introducing boundary flicker and drift where consecutive clips meet. A single pass avoids these artifacts, though it is only feasible when the trajectory fits within the model's sequence length.

H Reconstruction Behavior on SysCON3D Benchmark

To understand why learned 3D consistency metrics fail on 3D-inconsistent inputs, we inspect the reconstruction outputs produced by the backbones themselves on SysCON3D. We evaluate the 4 backbones - MASt3R, DUSt3R, Fast3R, and VGGT over $K \in \{3, 6, 9\}$ across every scene type in SysCON3D for our analysis.

H.1 Diagnostics, definitions, and main patterns

We use four complementary diagnostics. First, table 6 summarizes coarse point-cloud geometry and confidence from the sweep over different scene types from SysCON3D. Second, table 7 measures

Table 6: Coarse reconstruction statistics on 3D-inconsistent inputs from the controlled $K \in \{3, 6, 9\}$ sweep. BBox Vol, Conf, and Depth are averaged over $K \in \{3, 6, 9\}$; Points shows the K=6 value (scales as $\sim$ 50K $\times$ K for the DUSt3R, MASt3R and $\sim$ 270K $\times$ K for VGGT). We use this table as geometric context; the hallucination-specific overlap, residual, and ghost diagnostics are reported separately below.

Input Type	Engine	Points	BBox Vol	Conf	Depth
Mixed	MASt3R	301K	4086	1.06	2.77
	DUSt3R	301K	3.7	1.11	0.47
	Fast3R	301K	0.5	1.07	0.20
	VGGT	1.6M	21	1.05	0.78
One-Outlier	MASt3R	301K	3975	2.29	3.20
	DUSt3R	301K	6.3	1.59	0.53
	Fast3R	301K	1.7	1.31	0.25
	VGGT	1.6M	109	1.04	1.20
Gaussian Noise	MASt3R	301K	1.3	1.00	1.54
	DUSt3R	301K	0.05	1.81	0.52
	Fast3R	301K	0.00	1.21	0.20
	VGGT	1.6M	0.4	1.00	0.70
Patched	MASt3R	301K	643	1.58	2.27
	DUSt3R	301K	1.0	1.22	0.52
	Fast3R	301K	0.4	1.18	0.23
	VGGT	1.6M	18	1.05	1.00

pairwise overlap hallucination via

$$\mathrm{FNR}_{\mathrm{cross}} = P(\mathrm{overlap} \mid \mathrm{cross\text{-}scene}),$$

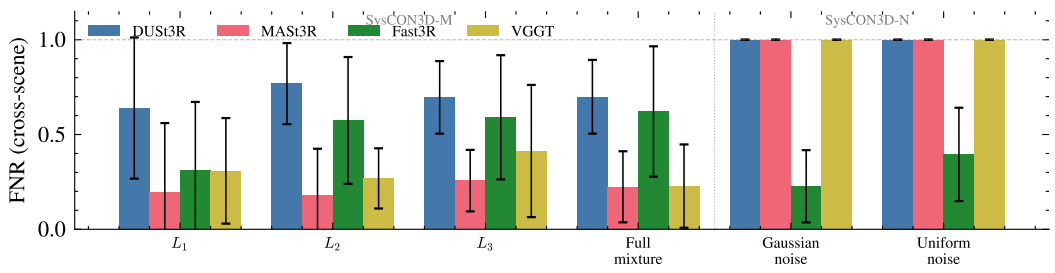
whose ideal value is 0. Third, table 9 splits DINOv2 cosine residuals into same-scene and cross-scene pairs after the backbone has already declared overlap. Fourth, table 10 reports confidence and, for Fast3R and VGGT, ghost mass: the fraction of rendered pixels supported by wrong-scene points. We focus on the main SysCON3D scene types L_0–L_3, and Gaussian noise, while using identical images and patched Gaussian as control cases.

Across all four tools, the same picture emerges. All backbones reconstruct 3D-inconsistent inputs instead of abstaining. On cross-scene mixtures, MASt3R hallucinates the least overlap and DUSt3R the most, while Fast3R and VGGT start from comparatively high same-scene zero-overlap on L_0. On Gaussian noise, MASt3R, DUSt3R, and VGGT still treat essentially every pair as overlapping, so the failure begins before any feature-space aggregation. Cross-scene DINOv2 residuals are larger than same-scene residuals, but their standard deviations overlap substantially, especially for Fast3R and VGGT, which limits separability once spurious correspondences survive. Confidence is therefore only a weak failure signal: MASt3R and VGGT move toward the floor on inconsistent inputs, whereas DUSt3R remains highly confident on Gaussian noise.

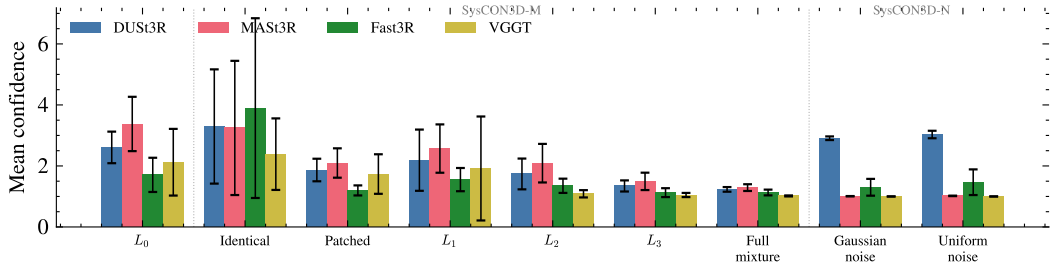
H.2 What each diagnostic reveals

Reconstruction geometry. Table 6 shows that every backbone produces dense point clouds even on inconsistent inputs. The number of points is determined mainly by image resolution, so Gaussian noise still yields 301K points for MASt3R, DUSt3R, and Fast3R, and 1.6M points for VGGT at K=6. Geometry collapses rather than disappearing: Gaussian noise shrinks the bounding-box volume to nearly zero, while mixed inputs produce large mash-up reconstructions that merge incompatible scenes into a single frame. The examples in figs. 8, 18, 20 and 21 show the characteristic modes. Full-Mixed, L_3 and L_2 stitch multiple scenes together; L_1 keeps a dominant scene while absorbing the foreign view; patched noise corrupts local support; and Gaussian noise collapses into planes or domes.

Backbones hallucinate overlap on cross-scene pairs. Table 7 and Table 8 are complementary. Table 7 is the coarse scene-level view: it compares the expected fraction of cross-scene pairs in each level against the backbone's overall zero-overlap rate across *all* pairs. If a backbone cleanly rejected



(a) $\mathrm{FNR}_{\mathrm{cross}}$ on SysCON3D-M and SysCON3D-N.



(b) Mean confidence across SysCON3D scene types.

Figure 5: **Hallucination diagnostics across the four backbones.** Left: cross-scene overlap hallucination rises sharply on L_2, L_3, and Gaussian noise. Right: confidence decreases on harder cross-scene mixtures, but it is not a reliable rejection signal on SysCON3D-N, most notably for DUSt3R.

Table 7: Expected cross-scene pair fraction versus the backbone's overall detected zero-overlap pair fraction, averaged across $K \in \{3, \ldots, 21\}$. This is a coarse proxy because the measured zero-overlap rate mixes both same-scene and cross-scene pairs. A negative delta indicates that the backbone shows overlap where none should exist.

Level	Expected	MASt3R	DUSt3R	Fast3R	VGGT
L_0 (consistent)	0%	4%	2%	39%	21%
L_1 (one outlier)	25%	22% (-3)	8% (-17)	40% ($+15$)	32% ($+7$)
L_2 (controlled)	60%	47% (-13)	15% (-45)	39% (-22)	45% (-16)
L_3 (random mixture)	89%	68% (-21)	26% (-63)	41% (-48)	71% (-18)

foreign views while preserving same-scene matches, the two quantities would be similar. Instead, all four backbones under-detect zero-overlap on the harder mixture levels. DUSt3R is the clearest failure: on L_3, it reports only 26% zero-overlap, against an expected 89%, a -63pp gap. At the same level, MASt3R, Fast3R, and VGGT still under-report zero-overlap by 21pp, 48pp, and 18pp. The apparent positive deltas for Fast3R and VGGT on L_1 do not indicate better cross-scene rejection. They arise because these backbones already drop many same-scene pairs on L_0 itself, with baseline zero-overlap rates of 39% and 21%, respectively.

Table 8 removes that same-scene confound by conditioning only on cross-scene pairs, so $\mathrm{FNR}_{\mathrm{cross}} = P(\text{overlap} \mid \text{cross-scene})$ directly measures hallucinated correspondences. Under this cleaner view, the backbone ordering is consistent with the coarse table: MASt3R is the least permissive on cross-scene mixtures, with $\mathrm{FNR}_{\mathrm{cross}}$ of 19.8%, 17.9%, 25.7%, and 22.4% on L_1, L_2, L_3, while DUSt3R is the most permissive, ranging from 64.0% on L_1 to 76.9% on L_2. Fast3R and VGGT lie in between at L_1–L_3, but their L_0 baselines remain high at 39.2% and 18.1%, respectively, which explains why they can look less problematic in table 7 than they are on truly cross-scene pairs. Gaussian noise is even more extreme: MASt3R, DUSt3R, and VGGT hallucinate overlap on every pair, and Fast3R still hallucinates overlap on 22.7% of Gaussian-noise pairs. Patched Gaussian is not a cross-scene condition and therefore does not appear in table 8; its effect is visible through the same-scene zero-overlap controls, where MASt3R and DUSt3R stay near zero but Fast3R rises to 72.4% and VGGT to 30.4%.

Table 8: **Cross-scene overlap failure rates in the hallucination sweep.** Values are mean ± std per scene type and are reported in percent. $\mathrm{FNR}_{\mathrm{cross}} = P(\mathrm{overlap} \mid \mathrm{cross\text{-}scene})$ is the fraction of cross-scene pairs that the backbone still treats as overlapping, so the ideal value is 0. FPR_{L_0} reports the same-scene zero-overlap baseline on consistent scenes, which is already non-zero for Fast3R and VGGT on far-apart views.

Backbone	FPR_{L_0}	L_1 $\mathrm{FNR}_{\mathrm{cross}}$	L_2 $\mathrm{FNR}_{\mathrm{cross}}$	L_3 $\mathrm{FNR}_{\mathrm{cross}}$	Full mix $\mathrm{FNR}_{\mathrm{cross}}$
MASt3R	0.2 ± 0.8	19.8 ± 36.2	17.9 ± 24.6	25.7 ± 16.2	22.4 ± 18.8
DUSt3R	0.5 ± 1.5	64.0 ± 37.2	76.9 ± 21.4	69.6 ± 19.1	69.9 ± 19.5
Fast3R	39.2 ± 34.1	31.2 ± 35.9	57.4 ± 33.5	59.1 ± 32.8	62.1 ± 34.4
VGGT	18.1 ± 17.7	30.8 ± 27.8	26.9 ± 15.9	41.3 ± 34.9	22.8 ± 22.0

Table 9: **Mean DINOv2 residuals on same-scene and cross-scene pairs.** Values are mean ± std of per-sample residual means. We report L_1, L_2, and L_3 because they each contain both same-scene and cross-scene pairs; full mixture is omitted because it has no same-scene pairs. Cross-scene residuals are consistently larger, but the gap remains modest relative to the variance, especially for Fast3R and VGGT.

Backbone	L_1 same	L_1 cross	L_2 same	L_2 cross	L_3 same	L_3 cross
MASt3R	0.345 ± 0.055	0.780 ± 0.144	0.321 ± 0.083	0.741 ± 0.047	0.339 ± 0.136	0.767 ± 0.088
DUSt3R	0.388 ± 0.103	0.815 ± 0.145	0.332 ± 0.092	0.796 ± 0.093	0.352 ± 0.154	0.797 ± 0.066
Fast3R	0.643 ± 0.136	0.928 ± 0.099	0.597 ± 0.155	0.842 ± 0.080	0.635 ± 0.193	0.869 ± 0.062
VGGT	0.652 ± 0.117	0.909 ± 0.076	0.621 ± 0.112	0.845 ± 0.098	0.646 ± 0.165	0.824 ± 0.096

Residual separation after overlap is hallucinated. Once a backbone decides that two views overlap, same-scene and cross-scene residuals are not cleanly separable. Cross-scene means are larger for every backbone and level, but the gaps remain modest relative to the standard deviations. MASt3R and DUSt3R show the clearest separation, with cross-scene residuals roughly 0.42–0.46 above same-scene residuals across L_1–L_3. Fast3R and VGGT are less separated because their same-scene residuals are already high, around 0.60–0.65, while cross-scene residuals stay below 0.93. This overlap is precisely why aggregation helps on L_2 and L_3 but still struggles on L_1: many cross-scene pairs survive with DINOv2 dissimilarities that remain inside the same broad range as genuine pairs.

Confidence and ghost mass. Figure 5 and table 10 show that confidence partially tracks degradation on cross-scene mixtures, but it is not a reliable rejection signal. All backbones assign high confidence to the identical-image control and lower confidence by L_3. However, DUSt3R remains highly confident on Gaussian noise (2.91 ± 0.06), while MASt3R and VGGT sit essentially at their minimum. For the global backbones, ghost mass exposes another failure that confidence alone misses: on L_3, roughly 40–50% of rendered pixels are supported by wrong-scene points, and VGGT's ghost mass rises to 0.80 ± 0.10 on Gaussian noise.

H.3 Implications for robust evaluation

Together, these diagnostics show that the learned backbones hallucinate 3D consistency rather than measuring it. For our neural metric family (backbone, aggregation, residual) in Section 4.1, the primary failure happens before aggregation: the reconstruction module B invents geometry or overlap on 3D-inconsistent inputs, and the residual module ρ inherits either semantically plausible DINOv2 matches on cross-scene pairs or feature collapse on noise. No choice of aggregation $\mathcal{A}$ can fully recover the correct ordering once the residual set is already degenerate. This is why distributional aggregation improves L_2 and L_3, but still fails on single outliers, on patched noise for some global backbones, and on Gaussian noise for nearly all learned variants.

This failure mode directly motivates our shift toward classical evaluators. SIFT and COLMAP can reject matches, fail to register views, or return only sparse geometry when the input set is inconsistent. That abstention signal is precisely what the learned backbones lack on SysCON3D. Our COLMAP-based metrics in Section 5 exploit this behavior by turning match verification, registration rate, and

Table 10: **Confidence and ghost-mass diagnostics.** Confidence is the per-sample mean confidence reported by each backbone, shown as mean ± std. We include identical-image and patched-Gaussian controls, plus representative hard cases (L_3, full mixture, and Gaussian noise). Ghost mass is the fraction of rendered pixels sourced from wrong-scene points for the global backbones. Confidence often moves in the right direction, but even near the floor the models still produce structured reconstructions and low residuals.

Backbone	L_0 conf.	Identical conf.	Patched conf.	L_3 conf.	Gaussian conf.	L_3 ghost	Full mix ghost	Gaussian ghost
MASt3R	3.38 ± 0.89	3.25 ± 2.20	2.10 ± 0.48	1.49 ± 0.28	1.01 ± 0.01	–	–	–
DUSt3R	2.61 ± 0.52	3.29 ± 1.87	1.87 ± 0.37	1.34 ± 0.18	2.91 ± 0.06	–	–	–
Fast3R	1.71 ± 0.56	3.90 ± 2.95	1.20 ± 0.16	1.12 ± 0.15	1.30 ± 0.28	0.39 ± 0.27	0.51 ± 0.26	0.37 ± 0.17
VGGT	2.12 ± 1.09	2.39 ± 1.17	1.73 ± 0.65	1.05 ± 0.07	1.00 ± 0.00	0.41 ± 0.17	0.40 ± 0.25	0.80 ± 0.10

dense depth agreement into the evaluation signal, yielding a more robust and reliable notion of 3D consistency for NVS evaluation.

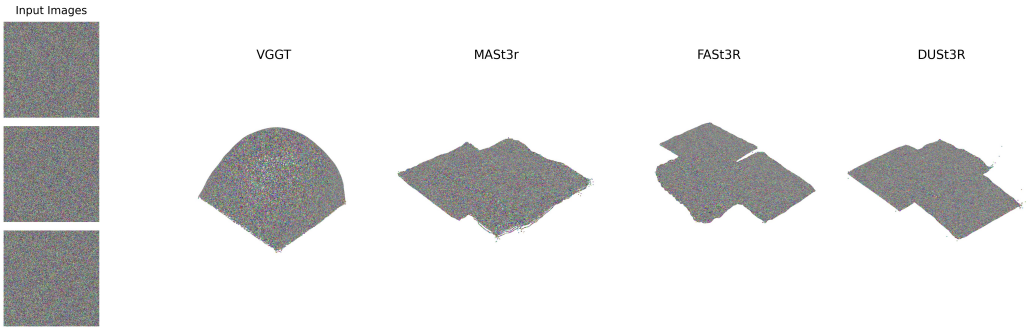


Figure 6: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Gaussian noise with $K = 3$. (Seed: 3045)

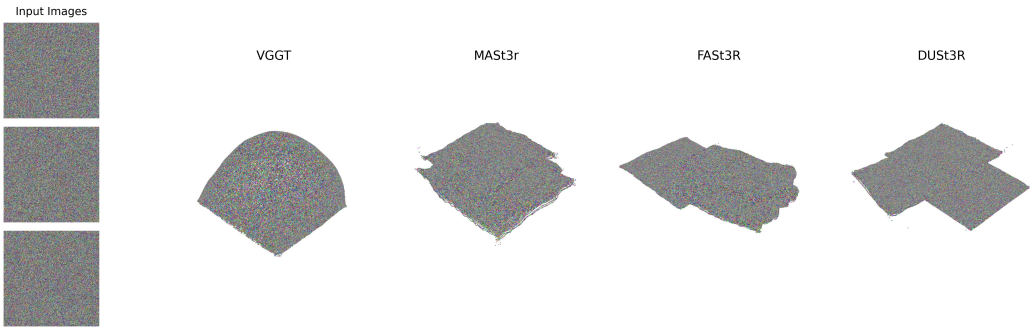


Figure 7: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Gaussian noise with $K = 3$. (Seed: 3145)

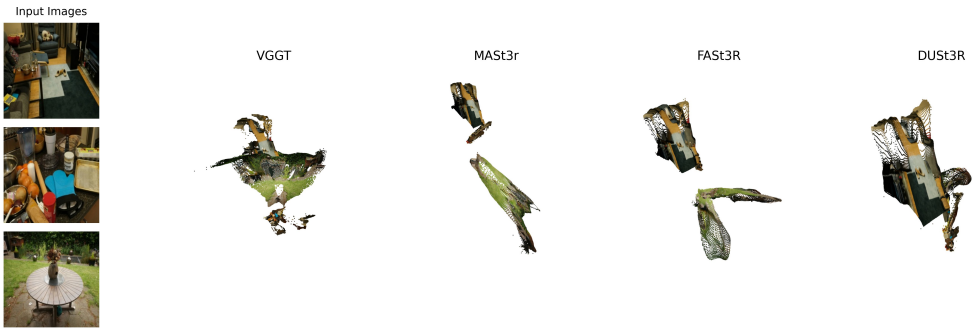


Figure 8: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Random mixture (L_3) with $K = 3$.



Figure 9: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Random mixture (L_3) with $K = 3$.



Figure 10: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Single-outlier (L_1) with $K = 3$.



Figure 11: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Single-outlier (L_1) with $K = 3$.

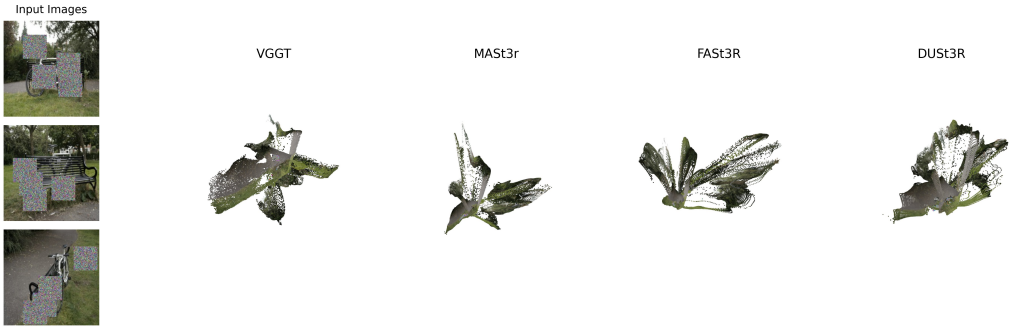


Figure 12: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Patched noise / Patched images with $K = 3$. (Seed: 4045)



Figure 13: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Patched noise / Patched images with $K = 3$. (Seed: 4145)

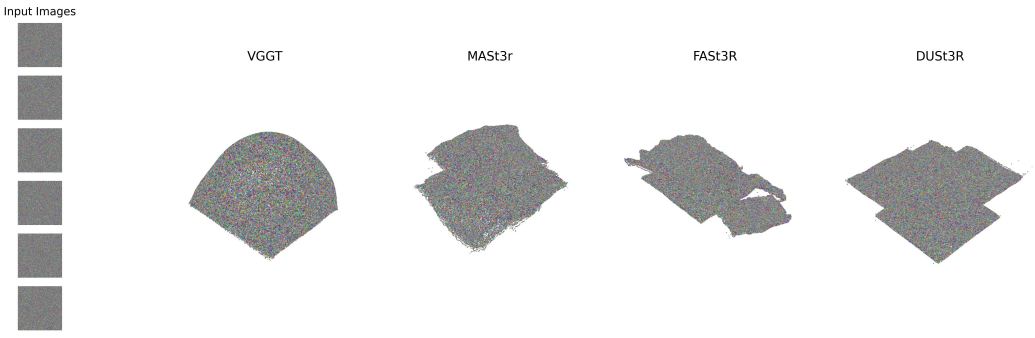


Figure 14: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Gaussian noise with $K = 6$. (Seed: 3048)

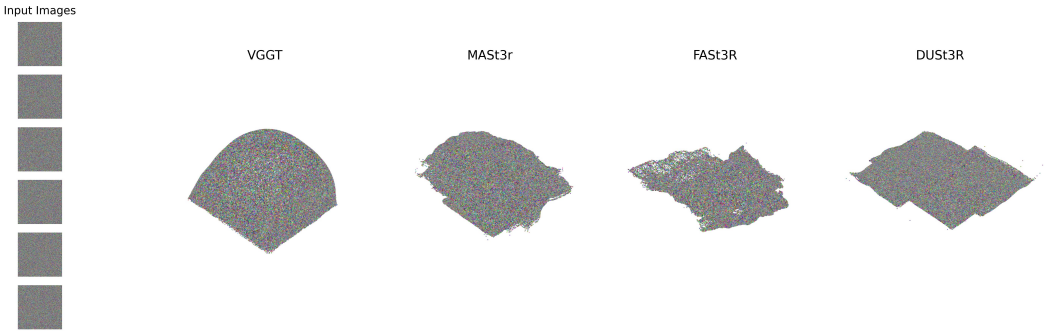


Figure 15: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Gaussian noise with $K = 6$. (Seed: 3148)

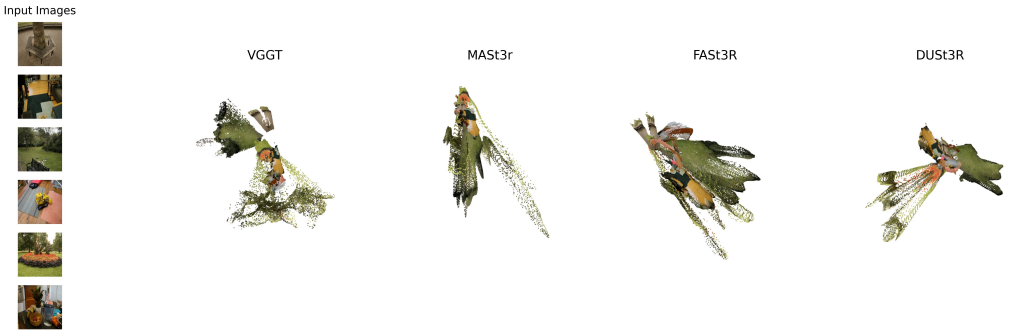


Figure 16: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Random mixture (L_3) with $K = 6$.

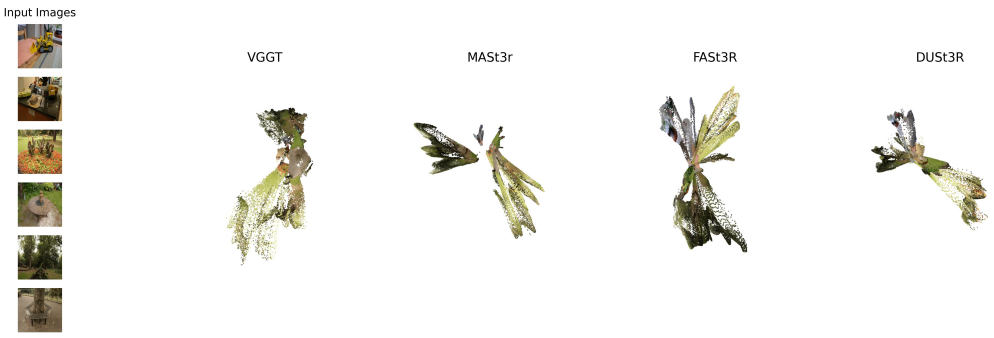


Figure 17: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Random mixture (L_3) with $K = 6$.



Figure 18: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Single-outlier (L_1) with $K = 6$.

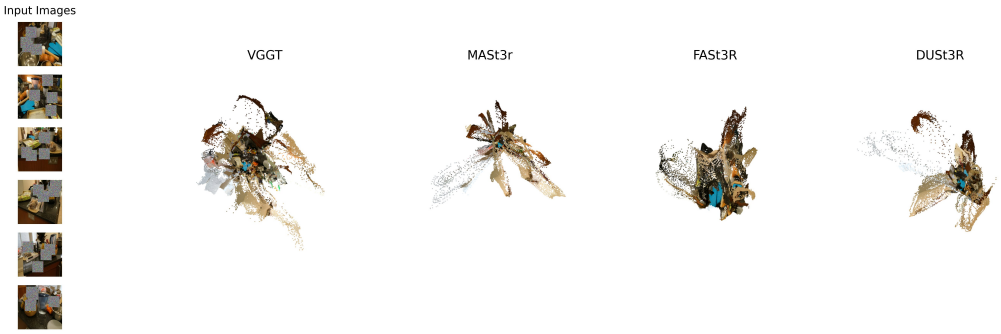


Figure 19: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Patched noise / Patched images with $K = 6$. (Seed: 4048)

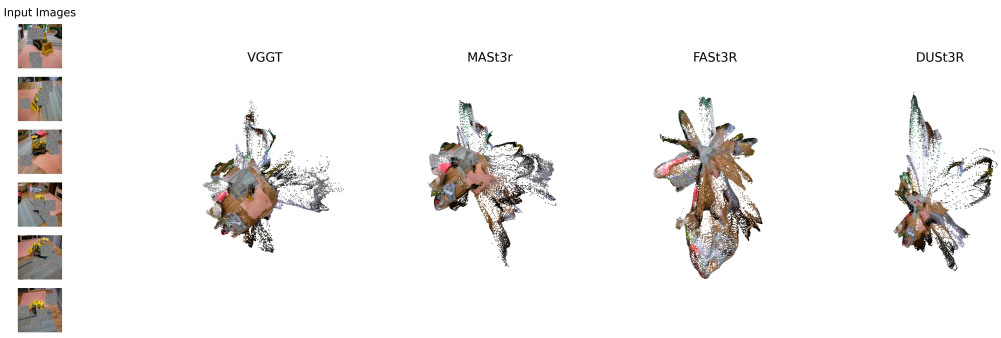


Figure 20: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Patched noise / Patched images with $K = 6$. (Seed: 4148)

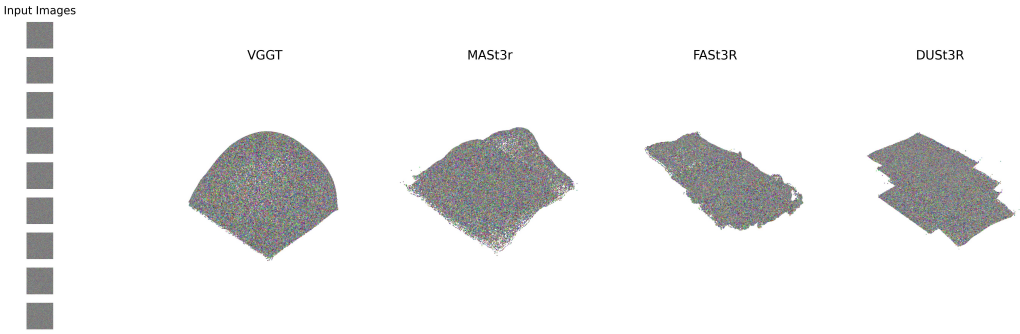


Figure 21: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Gaussian noise with $K = 9$. (Seed: 3051)



Figure 22: Analysis of 3D reconstruction backbones on SysCON3D benchmark. Input is Random mixture (L_3) with $K = 9$.

H.4 Signature Hallucination Patterns

Our hallucination analysis reveals that each learned 3D reconstruction model tends to produce a characteristic fallback geometry under severely inconsistent inputs. Although related artifacts can sometimes appear for real image sets with poor geometric correspondence, they are most clearly exposed by Gaussian-noise inputs. In that setting, the reconstructed geometry is remarkably stable across different noise realizations and numbers of views, indicating that it is dominated by the model prior rather than image evidence. VGGT, for instance, consistently produces a canopy-like structure, while MASt3R, DUSt3R, and Fast3R tend to reconstruct variants of a small set of approximately planar surfaces with model-specific spacing. This suggests that hallucinated geometry may exhibit backbone-specific signatures determined by architecture and training, which may be useful to study in future work.

H.5 Hallucinated correspondences on real NVS output

The diagnostics above are measured on SysCON3D, where we control the inputs directly. The same hallucination mode also appears in real NVS outputs. fig. 23 shows pixel correspondences returned by the two backbones on two pairs of views generated by MVSplat360 (Chen et al., 2024), which, in our experiments, drift and hallucinate more aggressively than the other NVS methods we tested. Within each pair, we deliberately pick frames from consecutive orbit cameras whose generated content shares no perceptible geometric or semantic overlap – a case where any backbone should return very few matches. Both MASt3R and VGGT instead link many low-residual (DINO/FeatUp feature distance small) pixel pairs across patches that clearly do not depict the same surface. When the backbone unprojects these matches, the two unrelated views are merged into a single shared point cloud, which is the same mechanism behind the metric failures in table 2: the residual aggregation cannot tell hallucinated matches apart from genuine ones, so the final score reports near-perfect 3D consistency on inputs that share none.

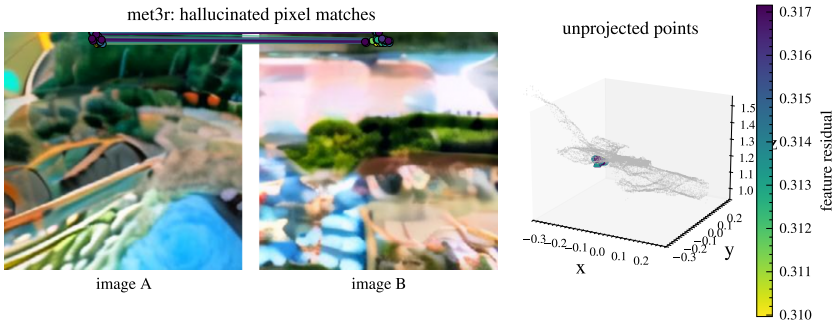
I Detailed Robustness Analysis

This subsection introduces a metric based on the RobustVGGT method (Han et al., 2025), and then explains some anomalous entries in tables 2 and 11.

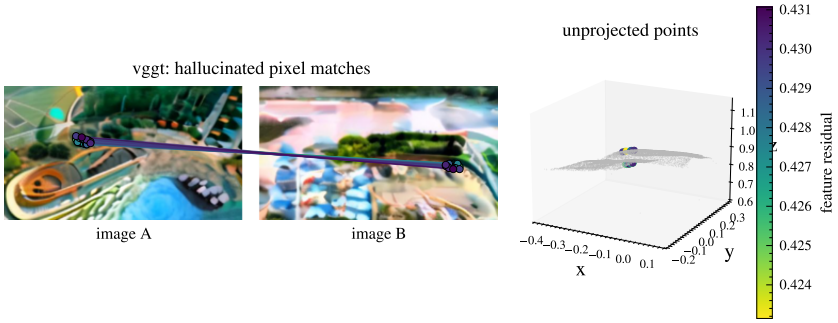
RobustVGGT metric. RobustVGGT (Han et al., 2025) is a test-time reconstruction method for mixed-scene VGGT inputs: it probes late global-attention features, scores each view by how well it agrees with an anchor view, rejects likely distractors, and reruns reconstruction on the retained subset. We adapt this into a scalar metric, *VGGT/W-Robust*, as follows. We hook layer 23 of the VGGT aggregator and compute one score per view from two late signals: the mean attention mass received from the anchor-view patches, and the late-feature cosine similarity to the anchor. We min-max normalize the two terms, average them with equal weights, and keep views whose combined score exceeds 0.4 while always retaining the anchor. We then rerun the standard weighted VGGT metric on the retained subset and add an explicit penalty proportional to the rejected-view fraction, so the metric cannot improve by simply discarding evidence and ignoring the rejected views.

RobustVGGT behavior. The VGGT/W-Robust row in table 2 shows that the added rejection signal helps on cross-scene mixtures (100% win on L_2 and L_3) but fails on patched Gaussian and Gaussian noise (0% win each), for an overall win rate of only 51%. The same pattern holds in table 11: VGGT/W-Robust still ranks repeated identical views below true consistent sets (d=-0.86, 14% win). Late-layer outlier rejection, therefore, helps when the failure is driven by real-world distractor views, but not when the backbone collapses on degenerate inputs.

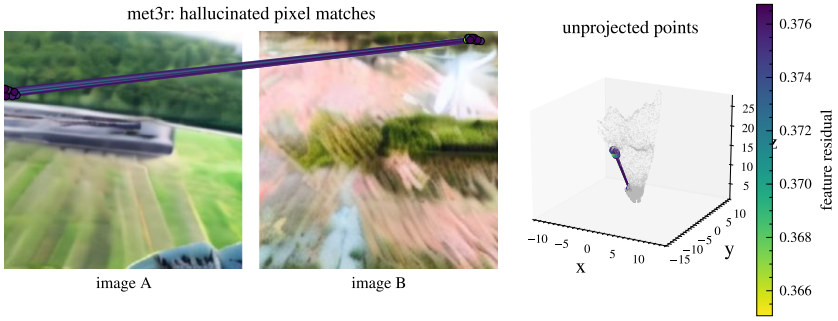
VGGT PC-MMD on Gaussian noise. In table 2, VGGT PC-MMD is the only data-driven variant that achieves a positive Cohen's d ($+1.31$, 100% win rate) on Gaussian noise. This is an artifact of RBF kernel saturation rather than genuine sensitivity to noise. On noise inputs with zero visual overlap, PC residuals are uniformly high (~ 1.0), and the RBF kernel $k(e, e') = \exp(-(e-e')^2/2\sigma^2)$ with σ=0.15 saturates near its maximum for nearby high residuals, pushing MMD to ~ 1.55. On real consistent scenes, PC residuals have intermediate values that keep MMD near ~ 1.45, so noise



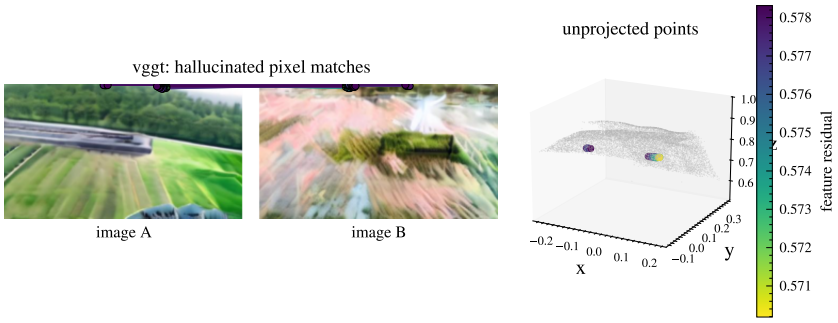
(a) Scene 1, MASt3R correspondences.



(b) Scene 1, VGGT correspondences.



(c) Scene 2, MASt3R correspondences.



(d) Scene 2, VGGT correspondences.

Figure 23: **Hallucinated correspondences between unrelated NVS frames.** Pixel correspondences returned by MASt3R (top of each pair) and VGGT (bottom) on two pairs of views generated by MVSplat360. The two views in each pair come from consecutive orbit cameras, and their generated content shows no perceptible geometric or semantic overlap, yet both backbones return many low-residual matches across unrelated patches. Unprojecting these spurious matches merges the two views into a shared point cloud, thereby biasing downstream consistency scores.

Table 11: **Robustness on the identical-image degenerate diagnostic.** Cohen's d (averaged over $K \in \{3, \ldots, 21\}$) and win rate (fraction of K values with $d > 0$) compare consistent sets (L_0) against repeated identical views. Positive d means the metric correctly assigns higher (worse) scores to repeated copies than to genuine consistent sets. Most metrics fail this test: identical views yield a trivially perfect but degenerate reconstruction, so they receive lower (better) scores than genuine consistent sets (negative d); MMD aggregations fail most severely (e.g., VGGT PC-MMD d=-19.72). IMQ variants are the exception, with Fast3R W-IMQ (d=$+9.88$) and VGGT W-IMQ (d=$+11.56$) reliably separating the two. MEt3R fails (d<0, 0% win rate); among external baselines, PRISM-MMD detects the diagnostic (d=$+2.92$). The top-3 values per column are shaded.

Backbone	Variant	Identical Images	
		Avg d	Win% (d)
MASt3R	W-Energy	-4.19	0%
MASt3R	W-MMD	-8.89	0%
MASt3R	W-IMQ	-2.78	0%
Fast3R	W-Base	-0.53	14%
Fast3R	W-Energy	-0.49	14%
Fast3R	W-MMD	-0.75	29%
Fast3R	W-IMQ	+9.88	100%
Fast3R	PC-Base	-4.28	0%
Fast3R	PC-Energy	-3.72	0%
Fast3R	PC-MMD	-7.92	0%
Fast3R	PC-IMQ	-2.64	0%
VGGT	W-Base	-0.85	14%
VGGT	W-Robust	-0.86	14%
VGGT	W-Energy	-1.10	0%
VGGT	W-MMD	-1.84	0%
VGGT	W-IMQ	+11.56	100%
VGGT	PC-Base	-7.25	0%
VGGT	PC-Energy	-6.31	0%
VGGT	PC-MMD	-19.72	0%
VGGT	PC-IMQ	-4.38	0%
MEt3R		-5.06	0%
PRISM-MMD		+2.92	100%
SED		-1.17	0%
TSED		-1.39	0%

accidentally exceeds consistent scores. This variant performs poorly on cross-scene mixtures (43% overall), limiting its practical utility.

SIFT-SED on Gaussian noise. SIFT-SED reports a degenerate Cohen's d of $+161.74$ on Gaussian noise. SIFT detects no keypoints on structureless noise images, causing SED to return a fixed fallback value of 17.5 with zero variance. For $K \geq 12$, consistent-scene SED is near zero ($\sim 0.1 \pm 0.07$), producing a tiny pooled standard deviation that inflates d to ~ 300 at individual K values. Averaging across all 7 values of K yields the reported 161.74. This is a measurement artifact: SED's keypoint-level evaluation is undefined on inputs without detectable features.

Single-outlier detection. Detecting a single foreign view among $K-1$ consistent ones remains difficult for all metrics. The best win rate is still 71% (VGGT W-Energy), while VGGT / W-Robust improves only to 57%, and most variants achieve $|d| < 0.2$. This is expected: with only $\sim \frac{2}{K}$ cross-scene view pairs, the signal from one outlier is diluted by the consistent majority, particularly at larger K.

Per-module interaction effects. The interaction between backbone and aggregation is still non-trivial. While global backbones outperform pairwise MASt3R at baseline weighted aggregation (46–49% vs. 23% overall), the strongest data-driven metric remains MASt3R W-IMQ at 71%. VGGT/W-Robust's gains come almost entirely from L_2, L_3, and full-mixture cases, so it does not close the gap to MASt3R W-IMQ or to the better-calibrated external baselines.

Ranking with explicit outlier rejection. This paragraph compares W-Robust against the stronger ordinal baselines. In table 12, VGGT / W-Robust reaches Kendall's τ=0.24 and PPC=0.56. This is slightly better than most VGGT variants, but it remains well below PRISM-MMD (τ=0.45, PPC=0.72) and TSED (τ=0.70, PPC=0.76). Even with explicit outlier rejection, the VGGT-based metric still does not maintain the ideal severity ordering as reliably as PRISM-MMD or TSED.

J Rank Concordance Analysis

Win rate (section 4.2) only tests whether each inconsistent group scores worse than consistent inputs. Here we ask a stricter question: does a metric recover the full severity ladder $L_0 < L_1 < L_2 < L_3 <$ Gaussian noise across the six calibration scene types? We therefore report two complementary rank-based measures.

Kendall's τ. For each $K \in \{3, 6, 9, 12, 15, 18, 21\}$, we rank the 6 scene types by their mean metric score and compare that ranking against the ideal ordering above, where lower scores mean better 3D consistency. We average Kendall's τ across all K values. We use Kendall's τ rather than Spearman's ρ, which is reported for the human-alignment tables in the main paper and appendix, because the target here is a fixed ordinal ladder, not a correlation between two empirical rankings of methods. Every pairwise inversion between scene types is meaningful, and Kendall's τ directly measures concordant versus discordant pairs. A perfect ordering gives $\tau = 1$, a random ordering is near 0, and a complete reversal gives $\tau = -1$.

Probabilistic pairwise concordance (PPC). PPC approximates separation between individual set scores. For corruption types g, h whose prescribed ordering requires g to score lower, we model independently drawn scores as Gaussian with estimated means μ_g, μ_h and standard deviations σ_g, σ_h. The estimated probability of correct ordering is

$$\widehat{\mathrm{PPC}}_{gh} = \Phi\left(\frac{\mu_h - \mu_g}{\sqrt{\sigma_g^2 + \sigma_h^2}}\right).$$

We average this quantity over the prescribed ordered pairs and view counts. This describes score-distribution separation under the approximation; it is not a confidence interval for the means or a significance test.

Table 12: **Rank concordance with the ideal scene-type ordering.** External baselines TSED and PRISM-MMD achieve the highest concordance, while MEt3R nearly perfectly inverts the intended ordering. Kendall's τ compares each metric's point-estimate ranking of the calibration scene types against the ideal ordering ($L_0 < L_1 < L_2 < L_3 \leq$ Gaussian noise). Probabilistic Pairwise Concordance (PPC) accounts for score uncertainty: for each pair (i, j) that should be separated, PPC $= \Phi\big(|\mu_i - \mu_j|/\sqrt{\sigma_i^2 + \sigma_j^2}\big)$, averaged over all pairs and all K values. TSED (τ=0.70, PPC=0.76) and PRISM-MMD (τ=0.45, PPC=0.72) lead by a wide margin; among backbone-based variants, PC-MMD aggregations perform best but remain below τ=0.30. MEt3R scores τ=−0.89 (PPC=0.20), nearly inverting the severity ordering - consistent with its inability to separate cross-scene corruption levels in table 2. Top-3 values per column are shaded.

Backbone	Variant	Kendall's τ	PPC
MASt3R	W-Energy	0.22	0.57
MASt3R	W-MMD	0.07	0.57
MASt3R	W-IMQ	0.26	0.59
Fast3R	W-Base	0.10	0.53
Fast3R	W-Energy	0.12	0.53
Fast3R	W-MMD	0.05	0.50
Fast3R	W-IMQ	0.20	0.57
Fast3R	PC-Base	0.20	0.56
Fast3R	PC-Energy	0.18	0.56
Fast3R	PC-MMD	0.28	0.62
Fast3R	PC-IMQ	0.26	0.60
VGGT	W-Base	0.10	0.50
VGGT	W-Energy	0.07	0.49
VGGT	W-MMD	0.12	0.50
VGGT	W-IMQ	0.20	0.56
VGGT	PC-Base	0.16	0.51
VGGT	PC-Energy	0.10	0.49
VGGT	PC-MMD	0.30	0.64
VGGT	PC-IMQ	0.16	0.54
VGGT	W-Robust	0.24	0.56
MEt3R		-0.89	0.20
PRISM-MMD		0.45	0.72
SED		-0.03	0.52
TSED		0.70	0.76

Results. Table 12 evaluates a stricter property than the win-rate analysis in Table 2: rank concordance with the ideal metric ranking introduced in fig. 2. TSED and PRISM-MMD clearly lead here; PC-MMD variants are the strongest backbone-based models; and MEt3R still nearly reverses the intended severity ladder. TSED achieves the best overall concordance (τ=0.70, PPC=0.76), followed by PRISM-MMD (τ=0.45, PPC=0.72). Among the backbone-based metrics, VGGT PC-MMD (τ=0.30, PPC=0.64) and Fast3R PC-MMD (τ=0.28, PPC=0.62) perform best, while most other variants remain only weakly positive with τ between 0.05 and 0.26. MASt3R W-IMQ, despite the best overall win rate in table 2, reaches only moderate concordance (τ=0.26, PPC=0.59), showing that separating consistent from inconsistent inputs is easier than recovering the full severity ordering. At the other extreme, MEt3R scores τ=−0.89 and PPC=0.20, which is close to a complete inversion of the desired ranking. SED is also near chance (τ=−0.03, PPC=0.52), whereas TSED converts the same geometric signal into a much more reliable ordinal ranking. Overall, the non-learned baselines remain much better aligned with the calibration ladder than the data-driven backbone-based metrics, reinforcing the motivation for COLMAP-based metrics.

K Evaluation of Parametric Metric Family on MipNeRF360

In table 13, we evaluate our best performing neural metric from the parametric family, MASt3R-W-IMQ, on NVS renderings of the 8 regression and diffusion-based reconstruction methods on 9 scenes of the MipNeRF360 dataset across 3, 6, and 9 views.

Table 13: **Metric rankings vs. human rankings on MipNeRF360.** IMQ shows the strongest alignment with human judgments, especially at higher view counts, while PRISM is anti-correlated across all splits. We compare MEt3R, PRISM-MMD (PRISM), and MASt3R-W-IMQ (IMQ) rankings of 8 NVS methods on 9 MipNeRF360 scenes across three view-count splits (K=3, 6, 9). IMQ is selected for its highest overall win rate against inconsistent scene types (table 2). Colors encode ranks 1 (best, green) to 8 (worst, red); rows sorted by K=3 human rank (**H**). Bottom row: Spearman rank correlation (ρ). IMQ achieves the highest ρ at every split, reaching 0.52 at K=9. PRISM shows negative correlation across all splits ($\rho \in [-0.40, -0.19]$), in contrast to its moderate positive correlation on DL3DV (table 3). Notably, DepthSplat - ranked 1st by humans - is placed 7th–8th by MEt3R and IMQ at K=3, but converges to rank 2–3 at higher K, suggesting metric-human agreement improves with more input views on MipNeRF360's wider-baseline scenes.

Method	K=3				K=6				K=9			
	H	MEt3R	IMQ	PRISM-MMD	H	MEt3R	IMQ	PRISM-MMD	H	MEt3R	IMQ	PRISM-MMD
DepthSplat	1	0.360 7	0.116 8	0.643 6	1	0.329 3	0.094 3	0.719 6	1	0.325 2	0.095 2	0.723 7
S-V-Camera	2	0.267 1	0.069 1	0.362 2	4	0.261 1	0.071 1	0.382 2	4	0.291 1	0.076 1	0.362 1
ViewCrafter	3	0.304 2	0.083 2	0.500 4	3	0.315 2	0.091 2	0.601 5	3	0.331 3	0.101 3	0.590 5
Long-LRM	4	0.381 8	0.114 7	0.728 8	2	0.444 8	0.150 7	0.739 7	2	0.426 8	0.149 7	0.702 6
MVSplat360	5	0.337 6	0.103 4	0.706 7	6	0.349 5	0.108 4	0.748 8	6	0.389 6	0.144 6	0.736 8
MVGenMaster	6	0.329 3	0.097 3	0.375 3	5	0.342 4	0.114 5	0.392 3	5	0.342 4	0.109 4	0.368 2
NVS-Solver	7	0.330 5	0.107 5	0.254 1	7	0.388 7	0.158 8	0.262 1	8	0.407 7	0.174 8	0.454 3
Difix3D	8	0.329 4	0.113 6	0.505 5	8	0.374 6	0.139 6	0.430 4	7	0.369 5	0.142 5	0.499 4
Rank Corr. (ρ)	–	0.00	**0.07**	-0.19	–	0.33	**0.43**	-0.40	–	0.38	**0.52**	-0.33

L COLMAP-based 3D Consistency Metrics: Extended Notation and Coverage

This section supplements Section 5. We clarify which reconstruction outputs enter each score and describe the angular-coverage calculation. All notation follows the main text, and higher scores indicate greater consistency. For $i \in \mathcal{J}$, let $\mathcal{U}_i = \{u \in \Omega_i : D_{g,i}(u) > 10^{-5},\ D_{g,i}(u), D_{p,i}(u) \text{ finite}\}$. We define $\text{Density}_i = |\mathcal{U}_i|/|\Omega_i|$ and $\text{Consistency}_i = |\mathcal{U}_i|^{-1} \sum_{u \in \mathcal{U}_i} q_i(u)$ when $\mathcal{U}_i$ is nonempty. Their product is $Q_i/|\Omega_i|$.

L.1 Motivation: why classical geometric verification

Our analysis shows that neural metrics can assign favorable scores to unsupported correspondences, even after improving their aggregation (Sections 4.2 and 4.4). This motivates an alternative based on geometric verification. COLMAP (Schönberger & Frahm, 2016) tests feature matches against geometric constraints and attempts to register images using the surviving correspondences. Rejected matches, failed registration, and missing dense support provide explicit evaluation signals. These failures indicate a lack of verified support, not proof that the images are physically inconsistent.

L.2 Notation and COLMAP outputs

Let $\mathcal{T}$ contain all attempted images, $\mathcal{R} \subseteq \mathcal{T}$ the images registered in the evaluated SfM reconstruction, and $\mathcal{J} \subseteq \mathcal{R}$ those with readable photometric and geometric depth maps. The registration rate is $|\mathcal{R}|/|\mathcal{T}|$.

For each $i \in \mathcal{J}$, the maps $D_{p,i}$ and $D_{g,i}$ come from COLMAP's PatchMatch passes without and with geometric consistency, respectively. Both use the same SfM reconstruction and scale; they are not independently estimated monocular depths. Readable maps need not contain valid support everywhere: invalid pixels receive zero agreement under Equation (2).

Let Ω_i be the full dense-map pixel grid and $Q_i = \sum_{u \in \Omega_i} q_i(u)$ the summed agreement in view i. GPC averages $Q_i/|\Omega_i|$ equally across views in $\mathcal{J}$, while ICM pools agreement over their pixels. Both therefore condition on available dense maps.

The attempted-set variant instead uses

$$\mathrm{ICM}_{\mathrm{all}} = \frac{\sum_{i \in \mathcal{J}} Q_i}{\sum_{i \in \mathcal{T}} M_i}, \tag{11}$$

where M_i is the image pixel count stored in the COLMAP database. Images in $\mathcal{T} \setminus \mathcal{R}$ fail registration, while those in $\mathcal{R} \setminus \mathcal{J}$ lack the required dense maps. Both contribute zero support but remain in this denominator. When $\mathcal{J}$ is empty, the dense-support scores are zero.

The two pixel counts serve different roles: $|\Omega_i|$ measures the dense-map resolution, whereas M_i measures the database image resolution. For nonempty $\mathcal{J}$,

$$\mathrm{ICM}_{\mathrm{all}} = \mathrm{ICM}\, \frac{\sum_{i \in \mathcal{J}} |\Omega_i|}{\sum_{i \in \mathcal{T}} M_i}.$$

Thus, $\mathrm{ICM}_{\mathrm{all}}$ reflects both failed support and any difference between these resolutions. Database normalization alone does not make the score resolution-invariant.

L.3 Angular coverage

Angular coverage measures how broadly the registered cameras are distributed around the reconstructed scene. Given sparse points $\{\mathbf{x}_p\}$, we use their coordinate-wise median as the center, $\mathbf{o} = \mathrm{median}_p\, \mathbf{x}_p$. For registered view i, let $\mathbf{R}_i$ and $\mathbf{t}_i$ be its world-to-camera rotation and translation. Its camera center is $\mathbf{c}_i = -\mathbf{R}_i^\top \mathbf{t}_i$.

The coverage calculation supports two projection planes: the plane spanned by the two leading principal directions of the registered camera centers, with a world-XZ fallback when degenerate, or the world-XZ plane directly. Let $\mathbf{b}_1, \mathbf{b}_2$ be orthonormal axes of the chosen plane. For a nonzero projected camera offset, its azimuth is

$$\theta_i = \left[\frac{180^\circ}{\pi} \operatorname{atan2}\big(\mathbf{b}_2^\top (\mathbf{c}_i - \mathbf{o}), \mathbf{b}_1^\top (\mathbf{c}_i - \mathbf{o})\big) \right] \bmod 360^\circ. \tag{12}$$

For $n_{\mathrm{reg}} = |\mathcal{R}| \geq 1$ with defined azimuths, sort $\theta_{(1)} \leq \cdots \leq \theta_{(n_{\mathrm{reg}})}$ and set $\theta_{(n_{\mathrm{reg}}+1)} = \theta_{(1)} + 360^\circ$. Coverage is the full circle minus the largest gap:

$$\omega = 360^\circ - \max_{1 \leq \ell \leq n_{\mathrm{reg}}} \big(\theta_{(\ell+1)} - \theta_{(\ell)}\big), \qquad \text{W-GPC} = \mathrm{GPC}_\omega = \mathrm{GPC}\, \frac{\omega}{360^\circ}. \tag{13}$$

A narrow cluster of camera directions gives low coverage; views spread around the scene give higher coverage. This measures camera-angle spread, not the fraction of the scene surface reconstructed.

Compute cost. Running our COLMAP-based metrics takes on average $\sim$5 minutes on our evaluated datasets on an RTX 3090.

A complete dense reconstruction failure receives zero. With partial reconstruction, GPC and ICM average only over views with readable dense maps. The attempted-set variant $\mathrm{ICM}_{\mathrm{all}}$ also retains failed images in its denominator. W-GPC additionally weights agreement by the angular coverage of registered cameras.

L.4 COLMAP failures

Approximately 1% of our NVS experiments on DL3DV and Mip-NeRF360 resulted in COLMAP failures. Although COLMAP can fail under standard SfM failure modes such as low texture, limited overlap, repetitive structure, or large illumination/viewpoint changes, we found that the failures in our setting were primarily caused by low-quality generated views, as illustrated in Figure 24. To rule out ordinary SIFT matching failures, we additionally repeated the affected evaluations using stronger correspondence pipelines, including SIFT+LightGlue and ALIKED+LightGlue. These

Table 14: COLMAP reconstruction quality vs. human rankings of 8 NVS methods on DL3DV (24 scenes), full-metric version for the $K=9$ view split. Colors encode ranks 1 (best, green) to 8 (worst, red); rows sorted by human rank (**H**). Bottom row: Spearman rank correlation (ρ) between human and metric-induced rankings.

Method	H↓	W-GPC↑	ICM↑	Ang. Cov.↑	GPC↑	Avg. Den.↑	Avg. Cons.↑	GPC All↑	ICM All↑	Reg. Rate↑
DepthSplat	1	0.288 ± 0.25 **1**	0.531 ± 0.16 **1**	170.4 ± 126.89 **1**	0.531 ± 0.16 **1**	0.622 ± 0.16 **1**	0.833 ± 0.09 **2**	0.294 ± 0.25 **1**	0.294 ± 0.25 **1**	0.496 ± 0.35 **1**
Long-LRM	2	0.136 ± 0.20 **2**	0.404 ± 0.17 **2**	94.1 ± 115.33 **2**	0.415 ± 0.16 **2**	0.496 ± 0.16 **3**	0.826 ± 0.10 **3**	0.138 ± 0.20 **2**	0.136 ± 0.20 **2**	0.294 ± 0.36 **2**
S-V-Camera	3	0.056 ± 0.12 **3**	0.351 ± 0.16 **3**	53.4 ± 95.19 **3**	0.353 ± 0.16 **5**	0.421 ± 0.18 **5**	0.844 ± 0.07 **1**	0.052 ± 0.14 **3**	0.052 ± 0.14 **3**	0.126 ± 0.27 **4**
ViewCrafter	4	0.036 ± 0.06 **4**	0.238 ± 0.12 **7**	46.2 ± 58.19 **4**	0.255 ± 0.11 **8**	0.328 ± 0.14 **8**	0.793 ± 0.10 **6**	0.014 ± 0.02 **6.5**	0.011 ± 0.01 **7**	0.113 ± 0.18 **6**
MVGenMaster	5	0.035 ± 0.06 **5**	0.302 ± 0.18 **6**	42.2 ± 51.68 **5**	0.313 ± 0.17 **6**	0.415 ± 0.22 **6**	0.772 ± 0.11 **7**	0.039 ± 0.08 **4**	0.046 ± 0.09 **4**	0.143 ± 0.21 **3**
MVSplat360	6	0.027 ± 0.06 **6**	0.319 ± 0.19 **5**	27.2 ± 31.09 **7**	0.370 ± 0.19 **4**	0.466 ± 0.22 **4**	0.805 ± 0.12 **4**	0.014 ± 0.02 **6.5**	0.019 ± 0.02 **6**	0.060 ± 0.05 **7**
NVS-Solver	7	0.018 ± 0.04 **8**	0.350 ± 0.25 **4**	18.9 ± 36.75 **8**	0.406 ± 0.23 **3**	0.502 ± 0.23 **2**	0.795 ± 0.13 **5**	0.026 ± 0.06 **5**	0.028 ± 0.07 **5**	0.116 ± 0.17 **5**
Difix3D	8	0.019 ± 0.01 **7**	0.237 ± 0.13 **8**	32.6 ± 25.98 **6**	0.261 ± 0.13 **7**	0.348 ± 0.15 **7**	0.748 ± 0.13 **8**	0.007 ± 0.01 **8**	0.006 ± 0.01 **8**	0.055 ± 0.03 **8**
Rank Corr. (ρ)↑	–	**0.976**	0.762	0.929	0.500	0.381	0.738	0.862	0.833	0.833

alternatives substantially relax the dependence on brute-force SIFT matching, but did not produce statistically significant changes relative to the standard SIFT-based results. This suggests that the observed failures reflect degradation in the generated imagery rather than limitations of the COLMAP matching backend.



Figure 24: Examples of COLMAP failures from our evaluation on DL3DV and MipNeRF360 generations. We observed that COLMAP failures corresponded to poor quality generations with little observable shared 3D consistency, as seen in the images.(Top) NVS Solver on DL3DV (Bottom) MVSplat360 on MipNeRF360 (stump)

L.5 COLMAP Metric Results

This section collects extended results for the COLMAP-based metrics defined in Sec. 5. tables 14 to 18 show that COLMAP metrics (W-GPC) are more human-aligned in ranking 3D consistency of NVS methods than the data-driven baselines. **Please note that the standard deviation provided in the following tables, graphs, and histograms are for metric values across scenes, which are independent of one another.** It is not the standard deviation of the metric for different runs on the same scene. The number of points in a point cloud, the amount of 3D information, size of images, etc., can vary drastically across different scenes. Therefore, this standard deviation also varies quite a bit. It is simply here to provide a general idea of what the values look like across the entire benchmark instead of listing a huge amount of tables, for the individual scene per model baseline results.

Table 15: COLMAP reconstruction quality vs. human rankings of 8 NVS methods on MipNeRF360 (9 scenes), full-metric version for the K=9 view split. Colors encode ranks 1 (best, green) to 8 (worst, red); rows sorted by human rank (**H**). Bottom row: Spearman rank correlation (ρ) between human and metric-induced rankings.

Method	H↓	W-GPC↑	ICM↑	Ang. Cov.↑	GPC↑	Avg. Den.↑	Avg. Cons.↑	GPC All↑	ICM All↑	Reg. Rate↑
DepthSplat	1	0.475 ± 0.31 **1**	0.636 ± 0.17 **1**	247.0 ± 145.55 **1**	0.636 ± 0.17 **1**	0.747 ± 0.17 **1**	0.835 ± 0.08 **2**	0.427 ± 0.33 **1**	0.427 ± 0.33 **1**	0.621 ± 0.43 **1**
Long-LRM	2	0.128 ± 0.26 **2**	0.378 ± 0.25 **2**	79.3 ± 110.39 **2**	0.381 ± 0.25 **3**	0.443 ± 0.27 **3**	0.829 ± 0.09 **3**	0.135 ± 0.26 **2**	0.136 ± 0.27 **2**	0.250 ± 0.33 **2**
ViewCrafter	3	0.017 ± 0.02 **4**	0.275 ± 0.18 **5**	25.5 ± 18.88 **3**	0.274 ± 0.18 **6**	0.362 ± 0.23 **6**	0.784 ± 0.12 **5**	0.009 ± 0.01 **6**	0.007 ± 0.01 **8**	0.083 ± 0.11 **3**
S-V-Camera	4	0.021 ± 0.04 **3**	0.363 ± 0.15 **3**	21.1 ± 41.89 **5**	0.361 ± 0.15 **4**	0.437 ± 0.18 **4**	0.839 ± 0.12 **1**	0.021 ± 0.02 **3**	0.021 ± 0.02 **3**	0.077 ± 0.09 **4**
MVGenMaster	5	0.015 ± 0.02 **5**	0.267 ± 0.25 **6**	25.2 ± 28.04 **4**	0.283 ± 0.24 **5**	0.410 ± 0.23 **5**	0.637 ± 0.19 **7**	0.009 ± 0.02 **6**	0.013 ± 0.02 **6**	0.056 ± 0.06 **7**
MVSplat360	6	0.013 ± 0.01 **6**	0.356 ± 0.24 **4**	13.5 ± 13.40 **7**	0.420 ± 0.22 **2**	0.527 ± 0.23 **2**	0.790 ± 0.14 **4**	0.012 ± 0.01 **4**	0.018 ± 0.03 **4**	0.044 ± 0.02 **8**
Difix3D	7	0.010 ± 0.01 **7**	0.203 ± 0.19 **7**	16.1 ± 12.03 **6**	0.245 ± 0.18 **7**	0.337 ± 0.18 **7**	0.692 ± 0.17 **6**	0.009 ± 0.02 **6**	0.018 ± 0.04 **5**	0.061 ± 0.04 **6**
NVS-Solver	8	0.003 ± 0.00 **8**	0.142 ± 0.09 **8**	13.1 ± 5.34 **8**	0.177 ± 0.07 **8**	0.299 ± 0.14 **8**	0.625 ± 0.14 **8**	0.007 ± 0.01 **8**	0.008 ± 0.01 **7**	0.070 ± 0.02 **5**
Rank Corr. (ρ)↑	–	**0.976**	0.881	0.952	0.690	0.690	0.714	0.805	0.563	0.786

Table 16: COLMAP reconstruction quality vs. human rankings of 8 NVS methods on DL3DV, full-metric version for the K=6 view split. Values are mean ± std over scenes. Colors encode ranks 1 (best, green) to 8 (worst, red); rows sorted by human rank (**H**). Bottom row: Spearman rank correlation (ρ) between human and metric-induced rankings.

Method	H↓	W-GPC↑	ICM↑	Ang. Cov.↑	GPC↑	Avg. Den.↑	Avg. Cons.↑	GPC All↑	ICM All↑	Reg. Rate↑
DepthSplat	1	0.251 ± 0.30 **1**	0.582 ± 0.22 **1**	128.985 ± 140.26 **1**	0.585 ± 0.22 **1**	0.671 ± 0.23 **2**	0.847 ± 0.10 **3**	0.238 ± 0.29 **1**	0.238 ± 0.29 **1**	0.381 ± 0.40 **1**
S-V-Camera	2	0.060 ± 0.13 **3**	0.386 ± 0.19 **4**	54.748 ± 96.31 **3**	0.398 ± 0.18 **4**	0.464 ± 0.20 **4**	0.855 ± 0.07 **2**	0.053 ± 0.14 **3**	0.053 ± 0.14 **3**	0.141 ± 0.28 **3**
Long-LRM	3	0.113 ± 0.18 **2**	0.334 ± 0.14 **5**	89.564 ± 120.46 **2**	0.337 ± 0.14 **5**	0.445 ± 0.16 **5**	0.763 ± 0.14 **6**	0.114 ± 0.18 **2**	0.113 ± 0.18 **2**	0.281 ± 0.35 **2**
ViewCrafter	4	0.035 ± 0.07 **5**	0.256 ± 0.17 **8**	42.834 ± 58.66 **5**	0.267 ± 0.15 **8**	0.372 ± 0.18 **8**	0.721 ± 0.12 **8**	0.019 ± 0.05 **5**	0.019 ± 0.05 **6**	0.083 ± 0.14 **6**
MVGenMaster	5	0.040 ± 0.05 **4**	0.283 ± 0.15 **7**	45.951 ± 45.28 **4**	0.316 ± 0.13 **6**	0.414 ± 0.16 **6**	0.775 ± 0.08 **5**	0.030 ± 0.07 **4**	0.033 ± 0.07 **4**	0.114 ± 0.17 **4**
MVSplat360	6	0.010 ± 0.01 **8**	0.570 ± 0.22 **2**	7.154 ± 6.17 **8**	0.582 ± 0.20 **2**	0.694 ± 0.23 **1**	0.840 ± 0.06 **4**	0.011 ± 0.01 **7**	0.011 ± 0.01 **7**	0.041 ± 0.01 **8**
NVS-Solver	7	0.013 ± 0.02 **7**	0.409 ± 0.26 **3**	14.371 ± 23.18 **7**	0.437 ± 0.25 **3**	0.497 ± 0.26 **3**	0.875 ± 0.11 **1**	0.018 ± 0.01 **6**	0.021 ± 0.02 **5**	0.089 ± 0.03 **5**
Difix3D	8	0.020 ± 0.02 **6**	0.287 ± 0.14 **6**	30.194 ± 29.85 **6**	0.304 ± 0.13 **7**	0.410 ± 0.15 **7**	0.743 ± 0.11 **7**	0.008 ± 0.01 **8**	0.009 ± 0.01 **8**	0.059 ± 0.05 **7**
Rank Corr. (ρ)↑	–	0.857	0.238	0.857	0.310	0.190	0.167	**0.929**	0.857	0.810

Table 17: COLMAP reconstruction quality vs. human rankings of 8 NVS methods on DL3DV, full-metric version for the K=3 view split. Values are mean ± std over scenes. Colors encode ranks 1 (best, green) to 8 (worst, red); rows sorted by human rank (**H**). Bottom row: Spearman rank correlation (ρ) between human and metric-induced rankings.

Method	H↓	W-GPC↑	ICM↑	Ang. Cov.↑	GPC↑	Avg. Den.↑	Avg. Cons.↑	GPC All↑	ICM All↑	Reg. Rate↑
DepthSplat	1	0.190 ± 0.27 **1**	0.552 ± 0.22 **1**	106.946 ± 133.21 **1**	0.580 ± 0.18 **1**	0.673 ± 0.17 **1**	0.848 ± 0.12 **1.5**	0.170 ± 0.27 **1**	0.170 ± 0.27 **1**	0.286 ± 0.38 **1**
S-V-Camera	2	0.023 ± 0.02 **5**	0.381 ± 0.19 **4**	32.645 ± 39.35 **5**	0.387 ± 0.18 **4**	0.450 ± 0.21 **5**	0.848 ± 0.08 **1.5**	0.022 ± 0.03 **4**	0.022 ± 0.03 **4**	0.077 ± 0.10 **5**
Long-LRM	3	0.028 ± 0.03 **3**	0.283 ± 0.17 **7**	34.183 ± 31.68 **3**	0.292 ± 0.16 **8**	0.368 ± 0.18 **8**	0.792 ± 0.13 **4**	0.021 ± 0.03 **5**	0.022 ± 0.04 **5**	0.101 ± 0.11 **4**
MVGenMaster	4	0.018 ± 0.03 **6**	0.320 ± 0.11 **5**	21.760 ± 34.98 **6**	0.334 ± 0.10 **5**	0.459 ± 0.14 **4**	0.738 ± 0.07 **8**	0.014 ± 0.02 **7**	0.015 ± 0.02 **7**	0.063 ± 0.07 **6**
ViewCrafter	5	0.038 ± 0.06 **2**	0.291 ± 0.14 **6**	42.596 ± 58.12 **2**	0.296 ± 0.14 **7**	0.387 ± 0.18 **6**	0.782 ± 0.10 **6**	0.027 ± 0.05 **2**	0.026 ± 0.05 **2**	0.111 ± 0.18 **3**
Difix3D	6	0.024 ± 0.03 **4**	0.283 ± 0.16 **8**	32.784 ± 35.54 **4**	0.300 ± 0.14 **6**	0.385 ± 0.17 **7**	0.785 ± 0.11 **5**	0.010 ± 0.01 **8**	0.010 ± 0.01 **8**	0.061 ± 0.04 **7**
NVS-Solver	7	0.013 ± 0.01 **7**	0.429 ± 0.24 **3**	14.734 ± 19.32 **7**	0.431 ± 0.24 **3**	0.518 ± 0.24 **3**	0.809 ± 0.18 **3**	0.025 ± 0.02 **3**	0.026 ± 0.02 **3**	0.144 ± 0.12 **2**
MVSplat360	8	0.013 ± 0.01 **8**	0.445 ± 0.20 **2**	10.161 ± 9.57 **8**	0.466 ± 0.16 **2**	0.605 ± 0.15 **2**	0.760 ± 0.11 **7**	0.017 ± 0.02 **6**	0.017 ± 0.02 **6**	0.055 ± 0.04 **8**
Rank Corr. (ρ)↑	–	0.667	0.000	**0.690**	-0.024	-0.048	0.548	0.405	0.357	0.476

Table 18: COLMAP reconstruction quality vs. human rankings of 8 NVS methods on MipNeRF360, full-metric version for the K=3 view split. Values are mean ± std over scenes. Colors encode ranks 1 (best, green) to 8 (worst, red); rows sorted by human rank (**H**). Bottom row: Spearman rank correlation (ρ) between human and metric-induced rankings.

Method	H↓	W-GPC↑	ICM↑	Ang. Cov.↑	GPC↑	Avg. Den.↑	Avg. Cons.↑	GPC All↑	ICM All↑	Reg. Rate↑
DepthSplat	1	0.266 ± 0.35 **1**	0.623 ± 0.28 **1**	139.076 ± 155.77 **1**	0.625 ± 0.27 **1**	0.692 ± 0.30 **1**	0.907 ± 0.06 **1**	0.241 ± 0.35 **1**	0.241 ± 0.35 **1**	0.344 ± 0.46 **1**
S-V-Camera	2	0.060 ± 0.11 **2**	0.424 ± 0.11 **3**	47.626 ± 78.69 **2**	0.425 ± 0.12 **3**	0.490 ± 0.13 **5**	0.874 ± 0.06 **2**	0.061 ± 0.08 **2**	0.061 ± 0.08 **2**	0.161 ± 0.19 **2**
ViewCrafter	3	0.025 ± 0.04 **3**	0.371 ± 0.13 **5**	29.193 ± 43.63 **3**	0.374 ± 0.13 **5**	0.474 ± 0.17 **6**	0.797 ± 0.08 **3**	0.021 ± 0.04 **4**	0.020 ± 0.04 **5**	0.092 ± 0.14 **5**
Long-LRM	4	0.006 ± 0.01 **8**	0.182 ± 0.13 **8**	10.153 ± 5.64 **8**	0.201 ± 0.12 **8**	0.310 ± 0.19 **8**	0.689 ± 0.11 **7**	0.002 ± 0.00 **8**	0.002 ± 0.00 **8**	0.048 ± 0.02 **6**
MVSplat360	5	0.011 ± 0.01 **5**	0.391 ± 0.15 **4**	10.589 ± 7.98 **7**	0.404 ± 0.15 **4**	0.571 ± 0.16 **2**	0.695 ± 0.11 **6**	0.009 ± 0.00 **6**	0.009 ± 0.00 **6**	0.043 ± 0.01 **7**
MVGenMaster	6	0.011 ± 0.01 **6**	0.319 ± 0.07 **6**	12.380 ± 5.64 **6**	0.323 ± 0.07 **6**	0.500 ± 0.06 **4**	0.642 ± 0.07 **8**	0.006 ± 0.00 **7**	0.006 ± 0.00 **7**	0.038 ± 0.00 **8**
NVS-Solver	7	0.009 ± 0.01 **7**	0.319 ± 0.22 **7**	17.001 ± 16.36 **4**	0.321 ± 0.22 **7**	0.393 ± 0.23 **7**	0.754 ± 0.17 **5**	0.020 ± 0.02 **5**	0.021 ± 0.02 **4**	0.155 ± 0.07 **3**
Difix3D	8	0.021 ± 0.02 **4**	0.453 ± 0.15 **2**	15.476 ± 11.60 **5**	0.450 ± 0.15 **2**	0.567 ± 0.10 **3**	0.776 ± 0.14 **4**	0.042 ± 0.09 **3**	0.043 ± 0.09 **3**	0.107 ± 0.17 **4**
Rank Corr. (ρ)↑	–	**0.595**	0.310	0.548	0.310	0.143	**0.595**	0.429	0.333	0.429

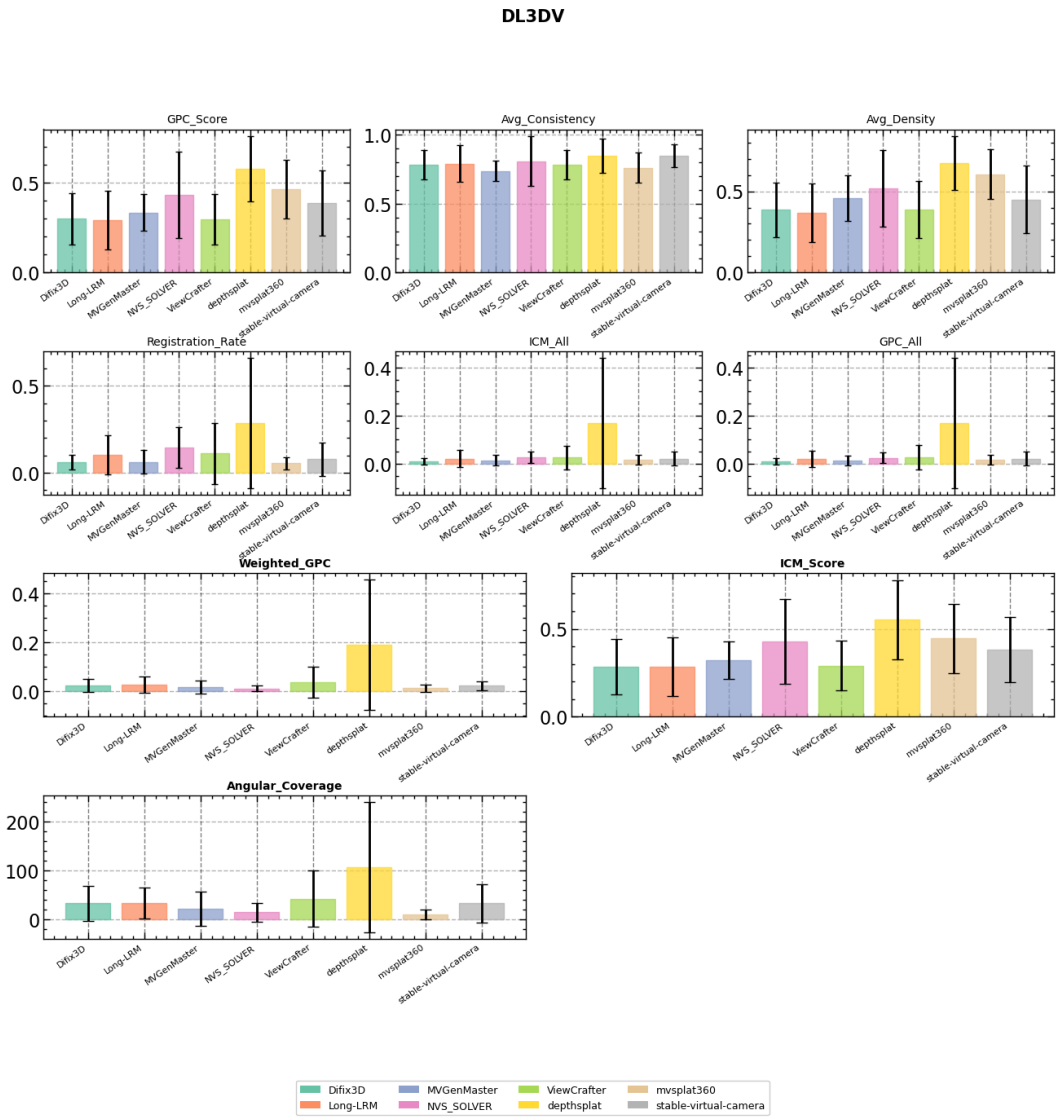


Figure 25: COLMAP based metric on DL3DV for K=3 view split.

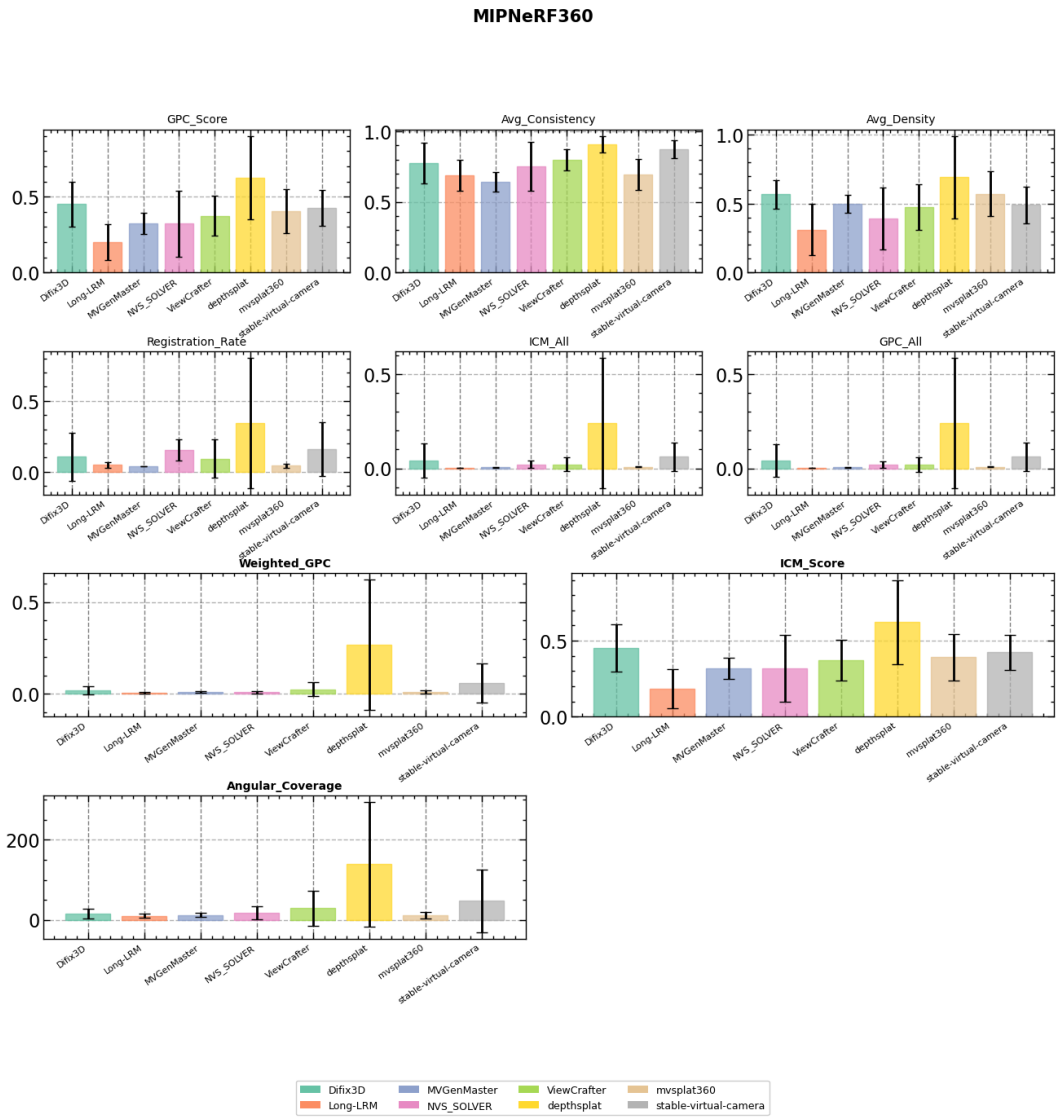


Figure 26: COLMAP based metric on MipNeRF360 for K=3 view split.

Table 19: Category-level behavior of COLMAP-based metrics on SysCON3D at $K=21$ under our custom comparison protocol. Gaussian Noise and Identical Images are assigned zero-valued metrics by convention since they do not yield meaningful COLMAP reconstructions. Rows report mean $\pm$ std across scenes. The final row reports category-level Spearman rank correlation (ρ) with the ideal inconsistency ordering: *Consistent* $>$ *Mixed-1-Outlier* $>$ *Mixed-Controlled* $>$ *Gaussian Noise* $=$ *Identical Images*. As can be seen, our COLMAP metrics are able to robustly rank the SysCON3D variants.

Variant	N	W-GPC↑	ICM↑	Coverage↑
Consistent	9	0.011685 ± 0.008193	0.286463 ± 0.088432	19.555 ± 16.216
Mixed-1-Outlier	8	0.010812 ± 0.000000	0.191046 ± 0.000000	20.330 ± 0.000
Mixed-Controlled	8	0.004896 ± 0.003891	0.210709 ± 0.126232	18.612 ± 17.540
Gaussian Noise	0	0.000000 ± 0.000000	0.000000 ± 0.000000	0.000 ± 0.000
Identical Images	0	0.000000 ± 0.000000	0.000000 ± 0.000000	0.000 ± 0.000
Spearman ρ	–	**1.000**	0.935	0.935