Appreciate the View: A Task-Aware Evaluation Framework for Novel View Synthesis

Saar Stern[1] Ido Sobol[1] Or Litany[1,2]
[1] Technion [2] NVIDIA

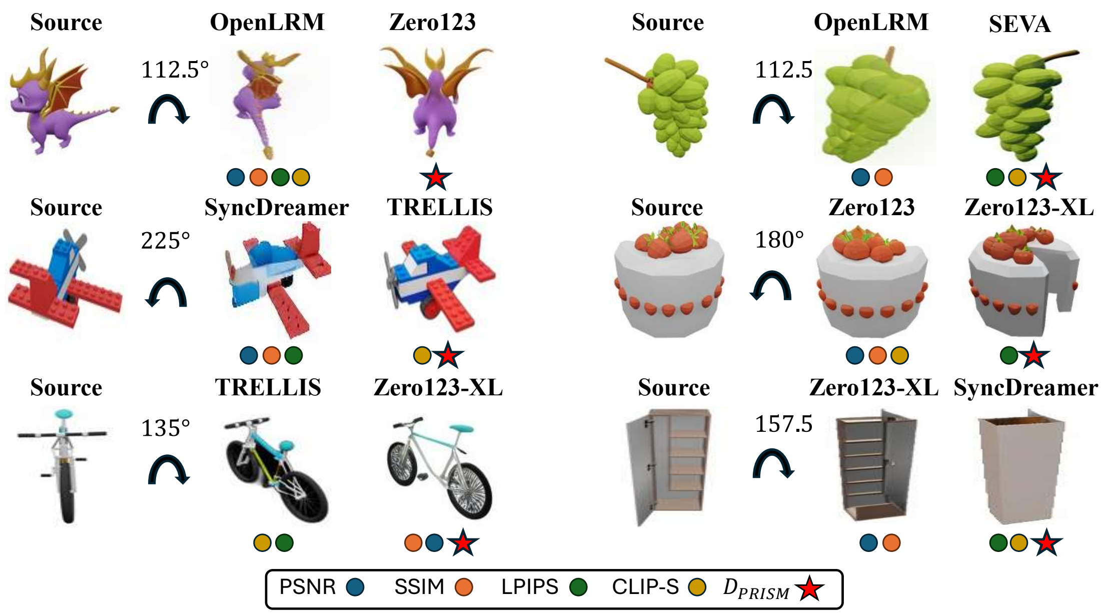

Figure 1. Standard metrics (PSNR, SSIM, LPIPS, CLIP-S) often mis-rank incorrect generations in novel view synthesis. Our metric, D_{PRISM}, penalizes these incorrect outputs, aligning more closely with human judgments. Each pair shows outputs from different NVS models under the same input, with the output favored by each metric indicated.

Abstract

The goal of Novel View Synthesis (NVS) is to generate realistic images of a given content from unseen viewpoints. But how can we trust that a generated image truly reflects the intended transformation? Evaluating its reliability remains a major challenge. While recent generative models, particularly diffusion-based approaches, have significantly improved NVS quality, existing evaluation metrics struggle to assess whether a generated image is both realistic and faithful to the source view and intended viewpoint transformation. Standard metrics, such as pixel-wise similarity and distribution-based measures, often mis-rank incorrect results as they fail to capture the nuanced relationship between the source image, viewpoint change, and generated output. We propose a task-aware evaluation framework that leverages features from a strong NVS foundation model, Zero123, combined with a lightweight tuning step to enhance discrimination. Using these features, we introduce two complementary evaluation metrics: a reference-based score, D_{PRISM}, and a reference-free score, MMD_{PRISM}. Both reliably identify incorrect generations and rank models in agreement with human preference studies, addressing a fundamental gap in NVS evaluation. Our framework provides a principled and practical approach to assessing synthesis quality, paving the way for more reliable progress in novel view synthesis. To further support this goal, we apply our reference-free metric to six NVS methods across three benchmarks: Toys4K, Google Scanned Objects (GSO), and OmniObject3D, where MMD_{PRISM} produces a clear and stable ranking, with lower scores consistently indicating stronger models. See our project page.

1. Introduction

Novel View Synthesis (NVS) is a fundamental problem in computer vision, requiring the generation of high-quality images from unseen viewpoints while maintaining structural, visual, and semantic consistency with the source view. This capability is essential for applications such as virtual and augmented reality (AR) [1], as well as robotics [22], where reconstructing objects and scenes from sparse observations is critical. Recent advances in generative models, particularly diffusion-based approaches [12, 33], have significantly improved NVS, enabling realistic image synthesis from limited inputs and accommodating large viewpoint shifts [19, 38]. However, these models remain imperfect, often struggling with geometric consistency and appearance preservation even under mild viewpoint changes. Several strategies have been proposed to improve view synthesis capabilities including using more training data [5], training multi-view models [47], incorporate strong 3D priors [41] and training-free approaches that enhance generation at inference time [32]. However, progress is hindered by the lack of robust evaluation — without reliable metrics, improvements remain ambiguous. Existing metrics fail to capture the relationship between the source image, the intended viewpoint transformation, and the generated output. Reference-based metrics such as PSNR, SSIM [37], and LPIPS [46] require a "true" target image and incorrectly penalize plausible variations that deviate from it. Reference-free metrics like FID [11] evaluate the generated view in isolation, ignoring consistency with the source view. Consistency-based metrics, such as Met3r [2], rely on correspondences between overlapping regions of the source and target, which makes them unreliable for large viewpoint changes where overlap is minimal or absent.

To address this gap, we advocate for the use of deep features, which have become standard in evaluating image generation and reconstruction quality [11, 14]. Specifically, we seek features that encode the source–target–viewpoint relationship, ensuring sensitivity to degradations in consistency and quality. We hypothesize such features already exist in strong NVS backbones, and leverage Zero123-XL [19]. Notably, diffusion model features have recently been shown to possess strong discriminative power [36, 42, 44], making them a promising basis for evaluating NVS quality.

To validate whether these features capture the desired relationships, we construct a new dataset, VIEWMATCH, consisting of source images paired with positive and negative target views. Positive samples preserve visible regions from the source and plausibly inpainted occluded parts, while negative samples contain alterations to visible regions that break consistency. This benchmark enables quantitative assessment of whether features separate valid from invalid generations — a key requirement for reliable evaluation. Our results show that, unlike conventional features [26, 29] that lack viewpoint awareness, ours separate plausible from implausible views; lightweight tuning on the train split of VIEWMATCH further improves discrimination. This requires minimal training data yet improves discrimination between correct and incorrect samples.

Using the refined features, we form two task aware evaluation metrics: D_{PRISM} and $\text{MMD}_{\text{PRISM}}$ which are reference-based and reference-free, respectively.

To assess the quality of our refined features, we conducted a human study in which participants compared NVS results across multiple models. Our reference-based metric produces rankings of NVS models that align well with human preferences. Our main contributions are:

- **Leveraging features from an NVS foundation model** to assess adherence to both the source image and the relative viewpoint transformation, providing an evaluation signal suitable for a generative task.
- **Constructing a benchmark** of carefully generated positive and negative pairs to (a) systematically tests whether candidate evaluation methods can distinguish between plausible and implausible novel views; (b) fascilitate a lightweight finetuning that enhances the discriminative power of our extracted features.
- **Demonstrating the effectiveness of our proposed metric** in ranking NVS models, aligning closely with human preferences.

By addressing a fundamental gap in NVS evaluation, our approach provides a robust metric for assessing synthesis quality, paving the way for more reliable advancements in the field. To further support this goal, we apply our reference-free metric to six NVS methods across three benchmarks: Toys4K, Google Scanned Objects (GSO), and OmniObject3D, where $\text{MMD}_{\text{PRISM}}$ produces a clear and stable ranking, with lower scores consistently indicating stronger models.

2. Related Work

2.1. Generative Novel View Synthesis

Recent advances in generative modeling [12, 30] have enabled impressive progress in novel view synthesis (NVS), where the goal is to render unseen views from sparse observations. Methods can be grouped into two categories. Image-based approaches synthesize views directly in pixel (or latent) space, from single-image models such as Zero123 [19] to multi-view extensions [20, 47]. In contrast, 3D-based methods leverage geometry representations—radiance fields [25], meshes, Gaussian splats [16, 27], or latent geometry spaces [41, 45]. Other work optimizes 3D representations under 2D supervision via score distillation sampling (SDS) [24, 28]. Despite differences in representation and training, all aim to generate plausible novel views, and thus fall within the scope of our evaluation

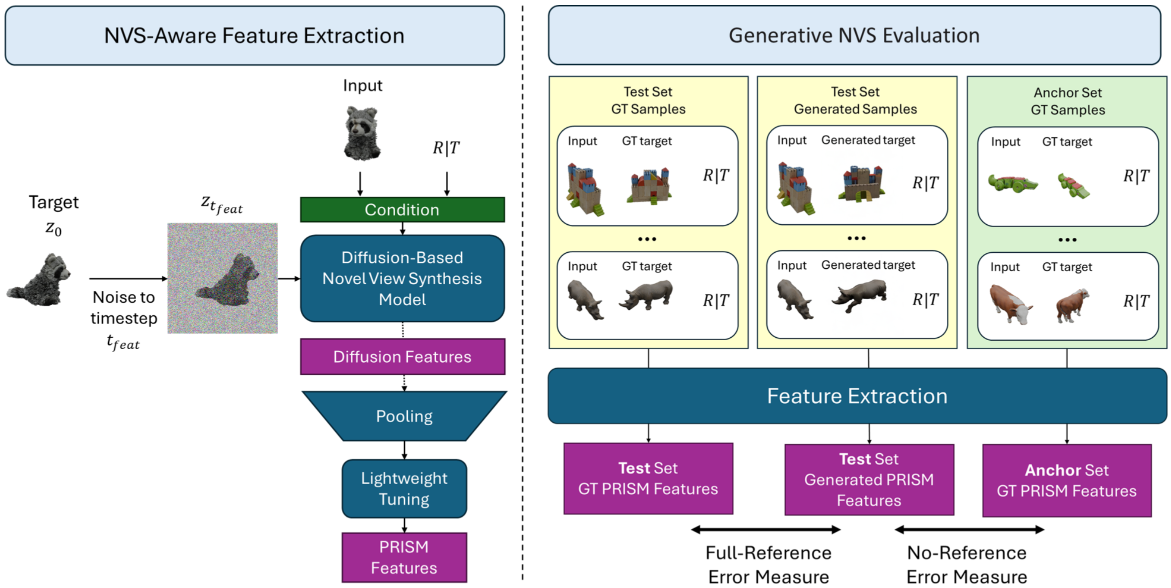

Figure 2. **Method Overview**. (Left) Feature extraction: given source, target, and camera transformation, we noise the target image and extract features from a diffusion-based NVS model. These are pooled and tuned into f_{PRISM}. (Right) Evaluation framework: **Full-Reference:** measure distance between f_{PRISM} of a predicted triplet and its ground-truth counterpart. **No-Reference:** compute MMD between f_{PRISM} from generated triplets and an anchor set of real triplets.

framework.

2.2. Evaluating Generative Image Models

Evaluating NVS remains challenging because multiple outputs can be correct.

Full-reference metrics such as PSNR, SSIM [37], LPIPS [46] and CLIP-S [10] compare against ground-truth targets, but penalize valid variations. Masked or silhouette-based extensions reduce bias [16, 32], yet remain limited. Multi-view consistency methods [4, 39] require multiple target views, making them impractical for single-view evaluation.

No-reference metrics avoid the need for ground truth, typically comparing feature distributions (IS, FID, KID, CMMD [3, 11, 14, 31]). However, they ignore the conditioning source and viewpoint transformation, rewarding realism over correctness. JFID and JFDD [8] incorporate source–target pairs but does not explicitly account for viewpoint change. Geometry-aware approaches such as 3DiM [38], TSED [43], and Met3r [2] enforce multi-view consistency but either require many views or overlook the plausibility of hallucinated regions. Overall, existing metrics fail to provide reliable single-source-to-single-target evaluation.

2.3. Diffusion Features for Discriminative Tasks

Diffusion features capture rich structural and semantic cues and have proven useful for tasks such as correspondence, segmentation, and shape matching [36, 42]. Recent work shows that pooled diffusion features support 3D reasoning and lightweight adaptation [21, 44]. Applied to NVS, these features naturally encode the interplay between source, target, and viewpoint, making them a promising foundation for evaluation.

3. Novel View Synthesis Diffusion Features

In this section, we introduce **PRISM** (*Pose-aware Representation for Image Synthesis Monitoring*)—a compact, triplet-aware embedding for evaluating single-source to single-target novel view synthesis (NVS). Given a source image I_{src}, a target image I_{tgt}, and a known relative transformation π, NVS aims to generate I_{tgt} consistent with both I_{src} and π. Since the task is ill-posed, evaluation must account for both geometric alignment and the plausibility of synthesized, unobserved regions. PRISM encodes this triplet into a fixed-length feature $f_{\text{PRISM}}(I_{\text{src}}, I_{\text{tgt}}, \pi)$ that enables robust, discriminative assessment.

Recent work has shown that diffusion models contain rich geometric signals, often accessed by extracting intermediate features and training task-specific classifiers [44]. We extend this idea to the novel view synthesis setting and instead learn a compact embedding via contrastive training, enabling representation-level reasoning over source, pose, and target image triplets.

An overview of the PRISM feature extraction pipeline is shown on the left side of Fig. 2. At a high level, PRISM uses a conditional NVS diffusion model to extract multi-scale features from an NVS triplet and projects them into a low-

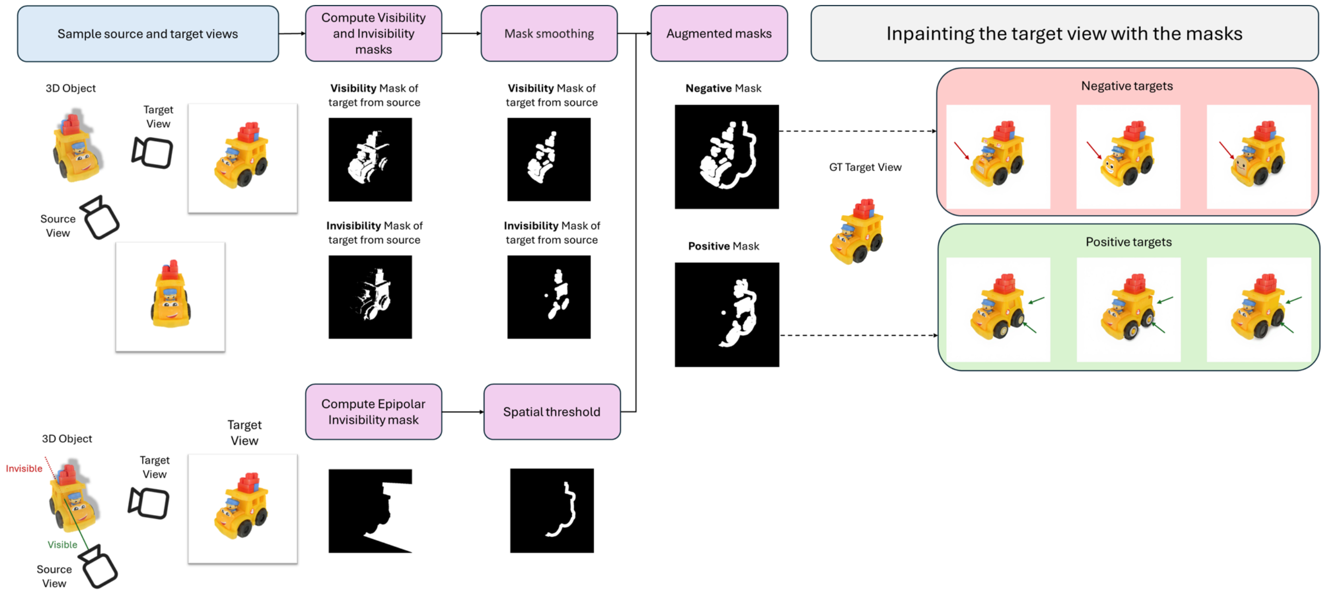

Figure 3. Overview of our VIEWMATCH creation process of positive and negative target examples. (Top Left) Given a 3D mesh and source and target viewpoints, we extract visibility and invisibility masks of the target view from the source, based on the visible faces of the target from the source. (Bottom Left) Given a 3D mesh and source and target viewpoints, we extract an epipolar invisibility mask, representing the unseen regions from the target view, beyond the object. (Right) We augment the visibility and invisibility masks with parts of the epipolar masks, to enable shape changes, and pass the true target and the created masks to an inpainting model.

dimension l_2-normalized embedding space. This projection is trained using a curated dataset of positive and negative triplets (see Sec. 3.2), where positive triplets correspond to plausible synthesized views, and negatives to implausible ones – generated under the same input conditions but containing geometric or appearance distortions. This enables the embedding to capture fine-grained differences in task adherence and forms the basis for our evaluation pipeline. Once extracted, PRISM can be applied in two evaluation modes as illustrated on the right side of Fig. 2:

Reference-Based Evaluation. Given a predicted triplet $(I_{\text{src}}, \tilde{I}_{\text{tgt}}, \pi)$ and a corresponding ground-truth triplet $(I_{\text{src}}, I_{\text{tgt}}, \pi)$, we compute the distance between their PRISM embeddings, normalized to $[0, 1]$:

$$\mathrm{D}_{\text{PRISM}} = \frac{1}{2}\left\| f_{\text{PRISM}}(I_{\text{src}}, \tilde{I}_{\text{tgt}}, \pi) - f_{\text{PRISM}}(I_{\text{src}}, I_{\text{tgt}}, \pi) \right\| \tag{1}$$

where lower values indicate stronger alignment with the reference.

Reference-Free Evaluation. We compare sets of PRISM embeddings from generated triplets $\mathcal{F}_{\text{gen}}$ and an anchor set $\mathcal{F}_{\text{anch}}$, using maximum mean discrepancy (MMD):

$$\mathrm{MMD}_{\text{PRISM}} = \mathrm{MMD}(\mathcal{F}_{\text{gen}}, \mathcal{F}_{\text{anch}}) \tag{2}$$

Lower values indicate stronger alignment. MMD estimates a distance between the underlying distributions, We choose it over Frechet distance (FD) due to its efficient GPU implementation, following CMMD [14]; full details are provided in the Appendix Sec. 13.

The next two subsections describe the construction of PRISM in detail. In Sec. 3.1, we explain how intermediate triplet-aware features are extracted from a conditional diffusion model. In Sec. 3.3, we outline the contrastive training procedure used to project these features into the final embedding space.

3.1. Diffusion Features Extraction

We extract features from a conditional diffusion model trained for novel view synthesis. While our framework is generally applicable, we focus here on Zero123 [19] as the backbone, as it is a well-established method for object-level NVS trained on a large-scale dataset. The target image I_{tgt} is encoded into a latent representation z_0, and Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is added at one timestep $t \in [0, T]$ to obtain:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon. \tag{3}$$

We perform a single denoising step with U-Net f_θ, conditioned on the source image I_{src} and relative pose π, and extract activations from each model block $b \in [1, B]$ (details about Zero123 U-Net architecture are in the Appendix Sec. 9.1):

$$F_b = f_{\theta_b}(z_t, t; I_{\text{src}}, \pi) \in \mathbb{R}^{H_b \times W_b \times C_b}, \tag{4}$$

Each activation map F_b is then normalized and spatial pooled :

$$v_b = \frac{1}{H_b W_b} \sum_{i,j} \frac{F_b[i,j,:]}{\|F_b[i,j,:]\|}, \tag{5}$$

and concatenated across blocks into a global vector:

$$v = \text{Concat}(v_{t,1}, \dots, v_B) \in \mathbb{R}^C, \quad C = \sum_{b=1}^{B} C_b \tag{6}$$

3.2. The ViewMatch Benchmark

Although diffusion features encode rich geometry-aware signals, they are not optimized to separate correct from incorrect generations. To learn such discriminative representations, we require supervision from contrastive pairs.

We therefore construct ViewMatch, which provides aligned positive and negative triplets for training. Positives preserve consistency with the source and camera motion, while negatives introduce deliberate violations in content, geometry, or appearance.

We select 40 objects from the GSO dataset [7], each rendered from multiple viewpoints. For every source–target pair, we create positive and negative examples by inpainting the target view using visibility-based masks. The resulting triplets are split into train and test sets for contrastive tuning and evaluation. See Appendix for full details.

Visibility and Invisibility Masks. These are masks rendered from the target view, indicating which parts of the object surface are visible from the source view. Specifically, the visibility mask shows which parts in the target view correspond to visible regions from the source view, as these regions should remain unchanged when generating a target view. The invisibility mask shows which parts in the target view correspond to occluded or invisible regions from the source view, as these regions can differ from the specific GT target view while still being geometrically and semantically consistent with the source view and camera transformation. We generate these masks by analyzing per-viewpoint visibility and applying post-processing to refine object boundaries. Further implementation details are provided in the Appendix.

Epipolar Mask. Inspired by iNVS [15], this mask is rendered from the target view and highlights regions that lie in 3D volumes invisible from the source view. We cast rays from the source camera onto the visible object surface; each continues into occluded space beyond the first intersection. Segments that pass through the object are discarded, and remaining occluded rays are projected onto the target view. These epipolar masks are used to augment visibility-based masks, enabling more diverse inpainting regions beyond the object silhouette. To avoid unrealistic edits, we constrain these masks to a neighborhood around the object. As these regions are unobserved in the source, any inpainting within them remains geometrically valid.

Generating Positive and Negative Examples. We generate positives by inpainting the invisibility and epipolar masks using FLUX [17], preserving geometric validity. Negatives are created by inpainting visibility and epipolar masks, corrupting regions that should remain unchanged. In both cases, the inpainting is text-guided using BLIP [18] prompts from the source image, ensuring semantic consistency and making the distinction more challenging. The use of epipolar masks extends edits beyond the ground-truth object area, increasing diversity and difficulty. See Fig. 3, Fig. 4, and the appendix for examples.

With these positive and negative triplets available, we can now perform contrastive training, using ViewMatch to fine-tune the projection head that maps raw diffusion features into the final PRISM embedding.

3.3. Contrastive Fine-Tuning

The raw diffusion features v (Eq. (6)) encode multi-scale cues from a source–pose–target triplet, but they are not explicitly optimized to distinguish plausible from implausible generations. To specialize them for evaluation, we train a lightweight two-layer MLP with ReLU activation using supervision from ViewMatch (Sec. 3.2). The MLP projects v into a compact, ℓ_2-normalized embedding:

$$f_{\text{PRISM}} = \frac{h(v)}{\|h(v)\|}. \tag{7}$$

The compact size of f_{PRISM} is important: it allows efficient distributional comparison via MMD, while reducing storage overhead when saving anchor sets.

Training follows a triplet contrastive strategy. For each GT triplet $(I_{\text{src}}, I_{\text{tgt}}, \pi)$ in the train split of ViewMatch, we form contrastive examples by pairing the anchor (GT) with:

- **Positive:** a plausible synthesized target consistent with the source and pose.
- **Negative:** either (i) an implausible target from ViewMatch, or (ii) a target view rendered at the wrong pose ("misaligned-angle").

The embedding is optimized with a margin-based triplet loss:

$$\mathcal{L}(a,p,n) = \max\big(d(a,p) - d(a,n) + m,\, 0\big), \tag{8}$$

where $d(s,s') = \|s - s'\|_2$ and m is a margin hyperparameter (further details are provided in the experimental setup). This encourages GT-aligned embeddings to be close to positives and far from both ViewMatch and misaligned-angle negatives.

4. Experiments

We evaluate D_{PRISM} and $\text{MMD}_{\text{PRISM}}$ across a series of controlled analyses and cross-dataset comparisons. The

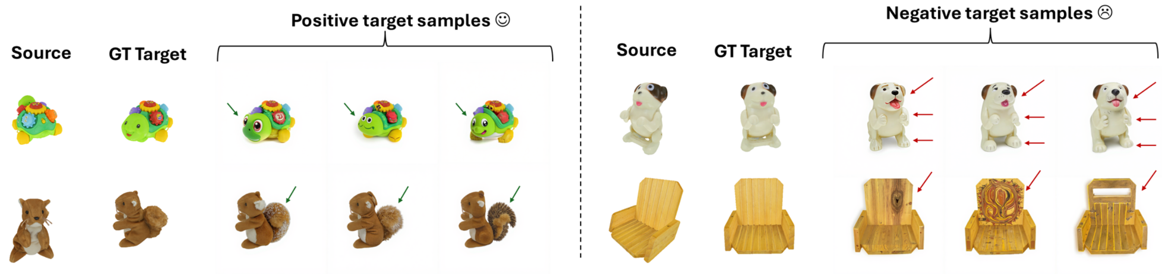

Figure 4. Examples from VIEWMATCH. Each group shows a source view, its ground-truth target, and three generated samples. Positives preserve consistency; negatives violate it through targeted inpainting as described in Sec. 3.2.

datasets we rely on are:

- **VIEWMATCH.** Our benchmark of positive and inpainted negative triplets, used to test discriminability, assess full-reference metrics, and validate no-reference evaluation.
- **Misaligned-GSO.** A controlled set where ground-truth targets are rendered at incorrect azimuths, isolating viewpoint sensitivity.
- **GSO [7].** Used for ranking (200/200 anchor/test split), for the user study with five diverse NVS models, and for degradation experiments.
- **Toys4K [34] and OmniObject3D [40]** Large-scale object datasets; we sample 40 and 20 categories respectively to benchmark rankings across domains.

Training Details. We optimize the projection head with AdamW (learning rate 1×10^{-4}, batch size 32) for 100 epochs, using a margin of $m = 1$ in the triplet loss. The embedding dimension is set to 2048, which is roughly one quarter of the raw diffusion feature size, providing a balance between discriminability and compactness for efficient MMD computation and anchor storage. Unless stated otherwise, all reported results use features extracted at $t = 0$.

Evaluation Roadmap. The subsections that follow build on these datasets in sequence: (Sec. 4.1–Sec. 4.3) establish discriminability and no-reference validation on VIEWMATCH, (Sec. 4.4) reports alignment with human judgments from the GSO-based study with five NVS models, (Sec. 4.5) ranks NVS models across GSO, Toys4K, and OmniObject3D, and (Sec. 4.6) analyzes robustness to pose misalignment and image degradation.

4.1. Discriminability of NVS Diffusion Features

We first test whether raw diffusion features can distinguish valid from invalid targets in VIEWMATCH. Baselines use CLIP and DINOv2 embeddings of source and target images concatenated with the relative angle (**CAT-Angle**). As shown in Tab. 1 (right), diffusion features achieve 0.90 AUC, clearly outperforming CLIP (0.73) and DINOv2 (0.68). Results are stable across diffusion timesteps ($t = 0, 100, 200$), with full results in the Appendix. For efficiency, we use $t = 0$ features in the remainder of the paper. This demonstrates that even without tuning, diffusion features encode viewpoint-sensitive discriminative signals absent from standard image features.

Table 1. Comparison of standard metrics and linear classifiers on VIEWMATCH. Left: average scores for pointwise evaluation metrics on positive (P) and negative (N) samples. Right: AUC scores for linear classifiers trained to distinguish positive from negative triplets.

Pointwise Evaluation Metrics			Linear Classifier AUC ↑	
Metric	**P**	**N**	**Feature**	**AUC**
PSNR (dB) ↑	18.18	18.69	CLIP-CAT-Angle	0.73
SSIM ↑	0.852	0.835	DINO-CAT-Angle	0.68
LPIPS ↓	0.163	0.178	Diffusion ($t = 0$)	0.90
CLIP-S ↑	0.89	0.96	Diffusion ($t = 100$)	0.87
D_{PRISM} ↓	**0.299**	**0.703**	Diffusion ($t = 200$)	0.86

4.2. Limitations of Standard Full-Reference Metrics

A natural question is whether standard full-reference (FR) metrics can identify invalid targets. We therefore benchmark PSNR, SSIM, LPIPS, and CLIP-S on positive versus inpainted negative triplets. As shown in Tab. 1 (left), these metrics not only fail to distinguish the two, but in some cases even assign *higher* scores to negatives. This is because both positive and negative samples remain perceptually close to the ground-truth target, so pixel- or feature-based similarity scores provide little separation. These findings illustrate that conventional FR metrics are poorly suited to NVS, where validity depends on viewpoint and structural consistency rather than low-level similarity.

4.3. Reference-Free Distributional Evaluation

We next ask whether $\text{MMD}_{\text{PRISM}}$ can serve as a no-reference (NR) evaluation metric, distinguishing valid from

Table 2. No-reference validation on VIEWMATCH. Lower is better.

Method	P	N
FID	102.360	107.240
CMMD	0.814	0.792
FDD	0.701	0.725
JFID	259.120	265.180
JCMMD	1.041	1.024
JFDD	1.631	1.659
$\text{MMD}_{\text{PRISM}}$	**0.691**	**0.984**

invalid generations without ground-truth targets. To test this, we compare distributions of PRISM features from positive versus negative triplets in VIEWMATCH against an anchor distribution built from 400 held-out objects.

As shown in Tab. 2, positives achieve substantially lower $\text{MMD}_{\text{PRISM}}$ scores (0.69) than negatives (0.98), indicating reliable separation at the distributional level. In contrast, standard NR metrics such as FID, CMMD, and JFID yield nearly identical scores for positives and negatives. Because these baselines operate purely on image appearance, they ignore the viewpoint-conditioned structure of NVS and thus fail to penalize implausible generations. This highlights that $\text{MMD}_{\text{PRISM}}$ provides a practical reference-free evaluation signal sensitive to both plausibility and geometric consistency.

4.4. Alignment with Human Judgments

Human perception is the gold standard for assessing highly generative outputs. We therefore test whether D_{PRISM} aligns with human preferences. Each task presents a source image, a relative camera motion, and two candidate target views to be judged along four aspects: *Viewpoint Accuracy*, *Shared Region Consistency*, *Plausibility of New Regions*, and *Image Quality*. A blurred hint of the ground-truth target aids viewpoint assessment while concealing fine detail.

We conducted the study on 40 held-out objects from the GSO dataset, distinct from those used in the creation of VIEWMATCH. For each object we rendered two large azimuth shifts ($> 90°$), yielding 320 possible pairwise comparisons across five diverse NVS models—OpenLRM, Zero123, Zero123-XL, TRELLIS, and SEVA—covering regression-based, multi-view, and 3D generation approaches. Forty participants each completed 15 comparison tasks, producing just over 500 pairwise judgments; with four questions per task, this corresponds to roughly 2,000 aspect-level human judgments. Full details on interface, instructions, and blur generation are provided in Appendix Sec. 12.

Tab. 3 reports Pearson correlation between metric predictions and human majority votes. Standard full-reference

Table 3. **Pearson correlation** between metric predictions and human judgments. Columns denote: VP = Viewpoint Accuracy, SC = Shared Consistency, PL = Plausibility of New Regions, IQ = Image Quality. Higher is better.

Metric	VP	SC	PL	IQ
PSNR	0.071	0.007	-0.187	-0.323
SSIM	**0.279**	<u>0.270</u>	0.114	-0.117
LPIPS	0.071	0.056	-0.035	-0.251
MEt3R	-0.028	-0.016	<u>0.179</u>	0.080
CLIP-S w/ GT	-0.330	0.107	-0.084	<u>0.019</u>
CLIP-S w/ src	-0.012	-0.048	-0.034	-0.008
D_{PRISM} (Ours)	<u>0.223</u>	**0.352**	**0.205**	**0.394**

metrics (PSNR, SSIM, LPIPS) achieve weak or even negative correlation, while MEt3R and CLIP-S also fail to capture human choices. In contrast, D_{PRISM} ranks first in 3 of the 4 aspects and second in the remaining one. Although absolute correlation values remain modest—reflecting the inherent difficulty of evaluating generative tasks—this still represents a clear advance, underscoring the suitability of D_{PRISM} as a perceptual evaluation metric for NVS.

4.5. Reference-Free Ranking of NVS Models

We use $\text{MMD}_{\text{PRISM}}$ to provide a reference-free leaderboard for single-view NVS models. Our evaluation covers six methods: Zero123, Zero123-XL, SEVA, OpenLRM, TRELLIS, and SyncDreamer [20]. Rankings are computed on GSO, Toys4K, and OmniObject3D, using the anchor–test splits described in Section 4. For each model we compute $\text{MMD}_{\text{PRISM}}$ between its generated triplets and the dataset-specific anchor distribution, and report *distribution-based baselines* (FID, CMMD, FDD, and their joint variants JFID, JCMMD, JFDD, more details in Appendix Sec. 13).

Tab. 4 reports results on Toys4K, where $\text{MMD}_{\text{PRISM}}$ produces a clear and consistent ranking, with lower scores corresponding to stronger models. In contrast, the distribution-based baselines often yield compressed score ranges that make it difficult to distinguish performance levels. Importantly, this trend is not unique to Toys4K: additional experiments on GSO and OmniObject3D (Tabs. 9 and 10), with summary results shown in Tab. 5, confirm that $\text{MMD}_{\text{PRISM}}$ provides stable rankings across datasets of varying difficulty, suggesting its practicality as a general, reference-free tool for comparing NVS models.

Discussion. Qualitative inspection (Fig. 1) further supports the reported rankings (e.g. Tab. 4) through highlighting systematic differences across models. OpenLRM underperforms relative to generative methods, consistent with its feed-forward design, which limits its ability to capture diverse object appearances. TRELLIS, though producing

Table 4. **Toys4K benchmark.** Comparison of reference-free metrics for ranking NVS models. Lower is better.

Metric	OpenLRM	Z123	Z123-XL	TRELLIS	SEVA	SyncDreamer
FID	168.5738	163.3072	161.2755	164.3074	156.2047	160.4168
CMMD	118.6056	85.6184	86.1490	34.7009	37.7579	102.9813
FDD	0.7483	0.7389	0.7249	0.6690	0.7030	0.7273
JFID	379.3934	371.8979	369.4718	373.8946	364.4985	371.3178
JCMMD	1.0403	0.7972	0.8036	0.3383	0.3680	0.9351
JFDD	1.5605	1.5380	1.5188	1.4673	1.5002	1.5350
$\text{MMD}_{\text{PRISM}}$ (Ours)	0.8017	0.3930	0.3067	0.4304	0.1978	0.2216

Table 5. $\text{MMD}_{\text{PRISM}}$ results across **Toys4K**, **GSO**, and **OmniObject3D**. For GSO and OmniObject3D, results of alternative metrics are deferred to the appendix.

Dataset	OpenLRM	Z123	Z123-XL	TRELLIS	SEVA	SyncDreamer
Toys4K	0.8017	0.3930	0.3067	0.4304	0.1978	0.2216
GSO	0.8415	0.3552	0.2997	0.2858	0.1231	0.1392
OmniObject3D	1.3752	0.6376	0.5206	0.5696	0.2896	0.2030

visually compelling outputs, is an image-to-3D model repurposed for NVS via source-view angle alignment; small angle mismatches introduced in this process reduce its score under PRISM, which is sensitive to pose inconsistency. Importantly, although PRISM is built on Zero123-XL features, it does not bias toward that model: methods such as SEVA and SyncDreamer consistently achieve higher scores across all three benchmarks. Taken together, these findings demonstrate that PRISM provides a robust and reliable evaluation tool.

4.6. Additional Analysis

Sensitivity to Pose Misalignment. We evaluate pose sensitivity on a controlled set of ground-truth renderings, where target views are intentionally misaligned by varying azimuth offsets relative to the source. Both D_{PRISM} and $\text{MMD}_{\text{PRISM}}$ follow an M-shaped trend, rising with larger offsets, peaking near 180°, dipping at symmetries, and recovering toward 360°, mirroring PSNR, SSIM, and LPIPS (Appendix Fig. 12) and confirming that our features capture viewpoint structure. In contrast, no-reference baselines (e.g., FID, JFID) remain flat (Fig. 6), as they ignore source conditioning and thus score even misaligned ground-truth views as plausible. This underscores the need for viewpoint-aware evaluation.

Sensitivity to Image Degradation. We also test responsiveness to low-level corruptions. Following [6], we apply Gaussian noise, Gaussian blur, color shifts, and salt-and-pepper noise at increasing intensities (parameters in Appendix Sec. 11.1). Results for Gaussian blur are shown in Fig. 5, while the other corruptions appear in Fig. 9–11. Across all cases, D_{PRISM} degrades consistently, with scores increasing steadily as distortions intensify, confirming its sensitivity to perceptual quality.

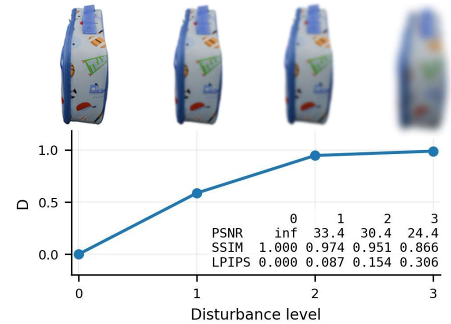

Figure 5. Degradation of D_{PRISM} (denoted as D in the plot) under Gaussian blur at increasing intensity levels.

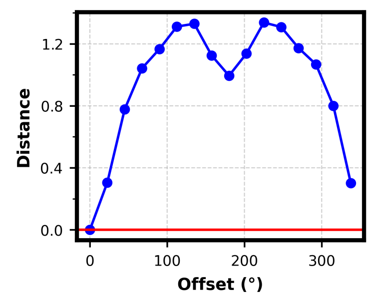

Figure 6. Metric scores vs. azimuth offset (0°–360°). Blue: our $\text{MMD}_{\text{PRISM}}$; Red: distribution-based baselines (e.g., FID, JFID). While the baselines remain flat, $\text{MMD}_{\text{PRISM}}$ shows an M-shaped curve, indicating pose sensitivity.

5. Conclusion and Limitations

We introduced task-aware diffusion features for NVS evaluation, explicitly capturing source–target–viewpoint relationships with a strong diffusion backbone and lightweight adaptation. Our metrics, D_{PRISM} and $\text{MMD}_{\text{PRISM}}$, show higher sensitivity to implausible generations than existing baselines, and enable both reference-based and reference-free evaluation. These tools provide a first step toward more reliable benchmarking of single-view NVS models, which remains a critical need for the community.

A limitation of our approach is its reliance on the Zero123-XL backbone, though the framework can be readily extended to other NVS models. Future work includes applying it to scene-level and multi-view settings.

6. Acknowledgement

Or Litany acknowledges support from the Israel Science Foundation (grant 624/25) and the Azrieli Foundation Early Career Faculty Fellowship. This research was also supported in part by an academic gift from Meta. The authors gratefully acknowledge this support.

References

[1] Daniel Andersen and Voicu Popescu. An ar-guided system for fast image-based modeling of indoor scenes. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 501–502. IEEE, 2018. 2

[2] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. *arXiv preprint arXiv:2501.06336*, 2025. 2, 3

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 3

[4] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4217–4229, 2023. 3

[5] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 4

[6] Terrance DeVries, Adriana Romero, Luis Pineda, Graham W Taylor, and Michal Drozdzal. On the evaluation of conditional gans. *arXiv preprint arXiv:1907.08175*, 2019. 8

[7] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 5, 6

[8] Noam Elata, Bahjat Kawar, Yaron Ostrovsky-Berman, Miriam Farber, and Ron Sokolovsky. Novel view synthesis with pixel-space diffusion models. *arXiv preprint arXiv:2411.07765*, 2024. 3

[9] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM, 2023. 4

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 3

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2, 3, 5

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[13] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 4

[14] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315, 2024. 2, 3, 4, 5, 6

[15] Yash Kant, Aliaksandr Siarohin, Michael Vasilkovsky, Riza Alp Guler, Jian Ren, Sergey Tulyakov, and Igor Gilitschenski. invs: Repurposing diffusion inpainters for novel view synthesis. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. 5

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 3

[17] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 5

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 5

[19] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2, 4

[20] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2, 7, 4

[21] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36: 47500–47510, 2023. 3, 2

[22] Dominic Maggio, Marcus Abate, Jingnan Shi, Courtney Mario, and Luca Carlone. Loc-nerf: Monte carlo localization using neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4018–4025. IEEE, 2023. 2

[23] Benyuan Meng, Qianqian Xu, Zitai Wang, Xiaochun Cao, and Qingming Huang. Not all diffusion model activations have been evaluated as discriminative features. In *Annual Conference on Neural Information Processing Systems*, pages 55141–55177, 2024. 2

[24] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12663–12673, 2023. 2

[25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez,

Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3, 5

[27] Chensheng Peng, Ido Sobol, Masayoshi Tomizuka, Kurt Keutzer, Chenfeng Xu, and Or Litany. A lesson in splats: Teacher-guided diffusion for 3d gaussian splats generation with 2d supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 28707–28717, 2025. 2

[28] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[31] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 3

[32] Ido Sobol, Chenfeng Xu, and Or Litany. Zero-to-hero: Enhancing zero-shot novel view synthesis via attention map filtering. *arXiv preprint arXiv:2405.18677*, 2024. 2, 3

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[34] Stefan Stojanov, Anh Thai, and James M. Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1798–1808, 2021. 6

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[36] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 2, 3

[37] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2, 3

[38] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 2, 3

[39] Sangmin Woo, Byeongjun Park, Hyojun Go, Jin-Young Kim, and Changick Kim. Harmonyview: Harmonizing consistency and diversity in one-image-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10574–10584, 2024. 3

[40] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 6

[41] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 2, 4

[42] Chenfeng Xu, Huan Ling, Sanja Fidler, and Or Litany. 3difftection: 3d object detection with geometry-aware diffusion features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10617–10627, 2024. 2, 3

[43] 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*, pages 7094–7104, 2023. 3

[44] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. A general protocol to probe large vision models for 3d physical understanding. *Advances in Neural Information Processing Systems*, 37:43468–43498, 2025. 2, 3

[45] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 2

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2, 3

[47] 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. 2, 4