MEt3R: Measuring Multi-View Consistency in Generated Images

Mohammad Asim[1] Christopher Wewer[1] Thomas Wimmer[1,2] Bernt Schiele[1] Jan Eric Lenssen[1]
[1]Max Planck Institute for Informatics, Saarland Informatics Campus [2]ETH Zurich
{masim, jlenssen}@mpi-inf.mpg.de

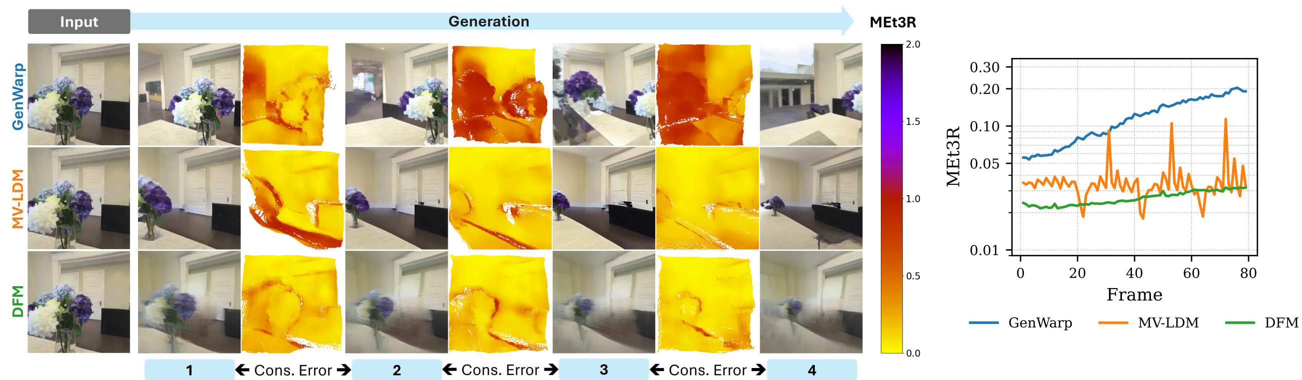


Figure 1. We introduce **MEt3R**, a metric for multi-view consistency between pairs of generated images, which is independent of image quality and content and does not require camera poses. **Left**: generated images from different generative models, conditioned on the first frame, with MEt3R score map (Cons. Error) indicating levels of inconsistencies between consecutive images i and $i+1$. **Right**: pair-wise consistency scores, evaluated for consecutive frames in a sliding window, averaged over multiple sequences. The pattern in MV-LDM's consistency clearly shows artifacts from using anchor frames that are generated first, highlighting the high signal-to-noise ratio of MEt3R.

Abstract

We introduce MEt3R, a metric for multi-view consistency in generated images. Large-scale generative models for multi-view image generation are rapidly advancing the field of 3D inference from sparse observations. However, due to the nature of generative modeling, traditional reconstruction metrics are not suitable to measure the quality of generated outputs and metrics that are independent of the sampling procedure are desperately needed. In this work, we specifically address the aspect of consistency between generated multi-view images, which can be evaluated independently of the specific scene. Our approach uses DUSt3R to obtain dense 3D reconstructions from image pairs in a feed-forward manner, which are used to warp image contents from one view into the other. Then, feature maps of these images are compared to obtain a similarity score that is invariant to view-dependent effects. Using MEt3R, we evaluate the consistency of a large set of previous methods for novel view and video generation, including our open, multi-view latent diffusion model. Code is available online: `geometric-rl.mpi-inf.mpg.de/met3r/`.

1. Introduction

Generative models, such as diffusion [14, 34] or flow-based [21] models, are trained to sample from a given data distribution, which makes them ideal candidates for stochastic inverse problems, such as reconstruction from incomplete information [11, 38, 45]. However, they raise the inherent challenge that, for individual samples, no ground truth is available to measure the quality of generations with pairwise distance metrics. As a result, metrics such as FID [13], KID [1], and CMMD [17] have been proposed to measure the quality of generated images without the need for a paired ground truth.

Recently, a trend is to repurpose video [3, 15] and image [29, 35] diffusion models for generation of 3D scenes and objects, by generating multiple views from different camera poses [11, 41, 52], with or without given images as conditioning. Compared to direct generation of 3D representations [6, 25, 31], such multi-view generative models can be trained on images and videos, and their pixel-aligned representation allows for more efficient models and better scalability. However, they have only a weak to non-existent inductive bias to produce actually 3D consistent re-

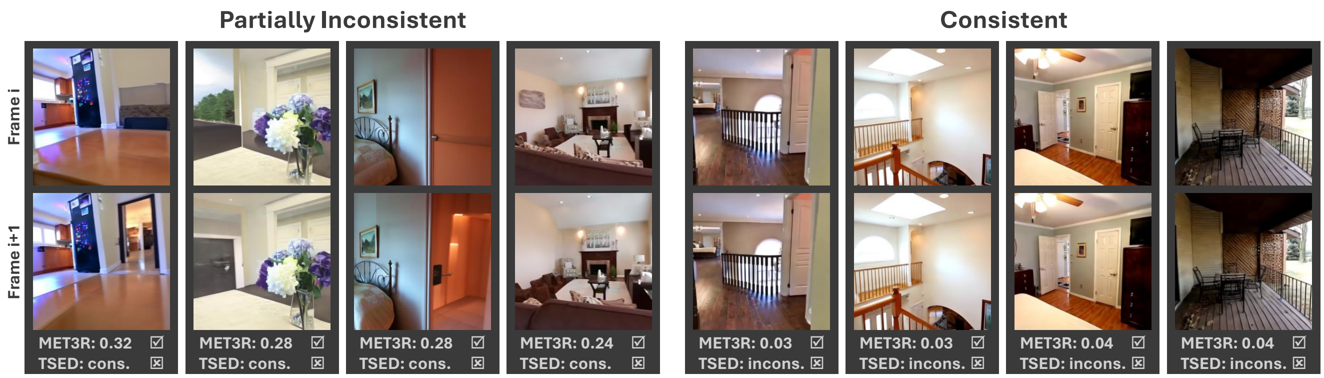


Figure 2. **Existing metrics.** A comparison between MEt3R and TSED [48] scores obtained from individual image pairs generated by GenWarp [32]. TSED misses obvious, partial multi-view inconsistencies and is biased to small violations of epipolar geometry. In contrast, MEt3R correctly captures clear 3D inconsistencies and is robust to insignificant artifacts almost invisible to the human eye.

sults, which is of large importance for the subsequent lift into 3D. A reliable metric to evaluate the multi-view consistency of such generations is critically needed to advance these models further. Fortunately, similar to general image quality, 3D consistency between views can be evaluated without the existence of paired ground-truth data. Existing metrics such as TSED [48], though, fail to reliably perform such evaluation, as shown in Fig. 2. In this work, we propose a metric to measure 3D consistency, which is independent of the specific scene and model used to generate the images, works under changing lighting conditions, does not require camera poses, is differentiable, and is a gradual measure of consistency instead of a binary one.

MEt3R utilizes DUSt3R [42] to obtain dense reconstructions from image pairs in a common 3D space. It then projects features of one image into the view of the other using the reconstructed point maps and computes feature similarity between the obtained images. As feature extractors, DINO [4] + FeatUp [10] are used to obtain high-resolution features from input images that are robust view-dependent effects, such as lighting, while preserving semantics and image-level structures, to quantify 3D consistency. We further introduce an open-source multi-view latent diffusion model (MV-LDM) to be used in our studies, which is able to generate good quality and consistent scenes. MEt3R is evaluated in different scenarios to validate its usefulness and robustness. It is used to benchmark existing methods that generate videos and multiple views of objects and scenes, with and without an intermediate 3D representation, as well as our MV-LDM. We show that MV-LDM performs well in the quality vs. consistency trade-off and find that MEt3R is a reliable metric that aligns well with the theoretical expectations of consistency among the different classes of scene generation methods. In contrast to previous metrics, it can distinguish perfectly consistent from almost consistent sequences and can robustly capture fine-grained changes in consistency over time.

In summary, our **contributions** include:
- a simple yet effective metric for measuring 3D consistency of generated views without given camera poses,
- a comprehensive analysis of existing methods that generate videos and multiple views of objects and scenes, and
- an open-source multi-view latent diffusion model, which performs best in the quality vs. consistency trade-off.

Our code and models are publicly available.

2. Related Work

We introduce a metric to evaluate the 3D consistency of multi-view generations. Thus, we review existing methods that generate multi-view representations of scenes and give an overview of existing quality metrics in this setting.

Multi-view Generative Models. Recent success in 2D image generation using generative models like diffusion [29] has sparked interest in generating 3D scenes. As the scarcity of high-quality training data and the complexity of 3D representations present a challenge for direct text-to-3D generative methods, recent methods explore repurposing image or video generation models as supervision signal or initialization for 3D generation [5, 11, 12, 20, 22, 26, 28, 32, 33, 38, 41, 45, 48, 52].

3D-aware image generation methods can be grouped into methods for pose-conditioned single-view generation [20, 32, 43, 48, 52], simultaneous multi-view image generation [11, 28, 33] and methods that use an internal 3D representation of the scene as prior for generation [5, 22, 38, 44, 45]. Further distinction can be made between models that are trained on single-asset 3D datasets [20, 22, 33, 43], such as Objaverse [7], and models trained on full 3D scenes [5, 11, 28, 32, 38, 45, 48, 52]. Our introduced metric is agnostic to how images are generated. In our experiments, we perform a comprehensive evaluation of consistency for images generated by openly available models, including those that model the joint distribution of input and single output views [32, 48], multiple output views [28], and methods that use an internal 3D representation [38] to enforce consistency.

Existing Metrics. Existing metrics used for quantifying image generation outputs include distribution-based metrics, such as the Fréchet Inception Distance (FID) [13], Kernel Inception Distance (KID) [1], Inception Score (IS) [30], or the CLIP Maximum Mean Discrepancy (CMMD) [17]. While these metrics are used to measure the alignment of generated samples with a target distribution using pretrained feature extractors, they do not measure 3D consistency, which is of utmost importance for multi-view generative models. To this end, Xie et al. [46] proposed using the Fréchet Video Distance (FVD) [39] to measure the quality of generated sequences with moving camera.

To explicitly measure 3D consistency, Watson et al. [43] proposed to train a NeRF [24] from a subset of generated views and compare rendered novel views with the remaining generated set of images. This metric comes with several drawbacks, as it requires a large amount of generated images, does not work on sparsely observed scenes, is expensive to compute, and difficult to interpret: are dissimilarities between generated views and rendered novel views from the trained NeRF caused by inconsistencies in the multi-view generation pipeline or insufficient quality of the NeRF training? As an alternative, Yu et al. [48] proposed TSED, a metric that checks whether image features detected in pairs of generated images respect the epipolar constraint, given the relative camera pose. As can be seen in Fig. 2, it has certain limitations, e.g., it deems two images consistent when it finds enough matching features, ignoring obvious inconsistencies in the images. In contrast, MEt3R does not require camera poses as inputs, and we find that it is more aligned with perceptual assessment when looking at the results of individual methods.

3. MEt3R: Measuring Consistency

In this section, we introduce MEt3R, our feed-forward metric to measure multi-view consistency. Given two images as input, a metric for multi-view consistency should (1) penalize image pairs that are not consistent, and (2) must not penalize pairs that are consistent but deviate from a given ground truth or do not follow a desired distribution. Thus, we develop MEt3R to be orthogonal to image quality metrics, e.g., FID [13], and to pixel-wise reconstruction metrics, e.g., PSNR.

An overview of MEt3R is shown in Fig. 3. Given two images $\mathbf{I}_1, \mathbf{I}_2$ as input, we first process them with DUSt3R to obtain dense 3D point maps for $\mathbf{I}_1$ and $\mathbf{I}_2$. Then, we obtain DINO [4] features on the original images and upscale them using FeatUp [10]. We use the predicted point maps to unproject the upscaled features of both images into the 3D coordinate frame of $\mathbf{I}_1$ and render them separately onto the 2D image plane of the 1^{st} camera to obtain two projections. Lastly, we compute feature similarity on the projected features, leading to cosine similarity scores, which we denote as $S(\mathbf{I}_1, \mathbf{I}_2)$ and $S(\mathbf{I}_2, \mathbf{I}_1)$.

MEt3R Definition. Given the scores $S(\mathbf{I}_1, \mathbf{I}_2)$ and $S(\mathbf{I}_2, \mathbf{I}_1)$, we can define MEt3R as,

$$\text{MEt3R}(\mathbf{I}_1, \mathbf{I}_2) = 1 - \frac{1}{2}\Big(S(\mathbf{I}_1, \mathbf{I}_2) + S(\mathbf{I}_2, \mathbf{I}_1)\Big), \quad (1)$$

which gives $\text{MEt3R}(\cdot, \cdot) \in [0, 2]$, lower is better, due to $S(\cdot, \cdot) \in [-1, 1]$, and is symmetric. We found S to already behave approximately symmetric. Thus, in practice, $\text{MEt3R}(\cdot, \cdot)$ can also be approximated well by only computing one direction of S in case of runtime constraints. We now provide the details for the DUSt3R reconstruction in Sec. 3.1 and feature similarity in Sec. 3.2

3.1. Stereo Reconstruction with DUSt3R

The core of our method relies on pose-free stereo reconstruction of pixel-aligned point clouds. Given an image pair $\mathbf{I}_1$, $\mathbf{I}_2$, the DUSt3R [42] model Ψ regresses pixel-aligned 3D point clouds $\mathbf{X}_1 \in \mathbb{R}^{H \times W \times 3}$ and $\mathbf{X}_2 \in \mathbb{R}^{H \times W \times 3}$:

$$\mathbf{X}_1, \mathbf{X}_2 = \Psi(\mathbf{I}_1, \mathbf{I}_2), \quad (2)$$

where point locations of both, $\mathbf{X}_1$ and $\mathbf{X}_2$ are given in the camera space of $\mathbf{I}_1$. It does so by employing a shared ViT [8] backbone to extract image features. Then, both feature maps are decoded by separate transformer decoders with cross-view attention that encodes a multi-view prior and shares important information between views. Finally the decoded features are regressed into point maps $\mathbf{X}_i$. For more details, please refer to the original work [42].

DUSt3R does not require camera poses, which is inherited by MEt3R. While MASt3R [19] additionally finds potentially useful feature correspondences between the two images, we do not make use of them in our method and hence stick with DUSt3R.

3.2. High-Resolution Feature Similarity

Since both generated point maps contain points in the canonical coordinate frame of $\mathbf{I}_1$, we can use the point maps to project pixel-aligned features from camera space of $\mathbf{I}_2$ into that of $\mathbf{I}_1$. Instead of performing this projection and the subsequent comparison directly in RGB pixel space, we found it more suitable to perform them in feature space. The reason are view-dependent effects, such as different lighting, which often occurs in natural videos and negatively impacts RGB comparisons. We provide a detailed comparison between both approaches in Sec. 5.4.

Concretely, we first use DINO [4] to obtain semantic features for $\mathbf{I}_1$ and $\mathbf{I}_2$. Then, since the corresponding feature

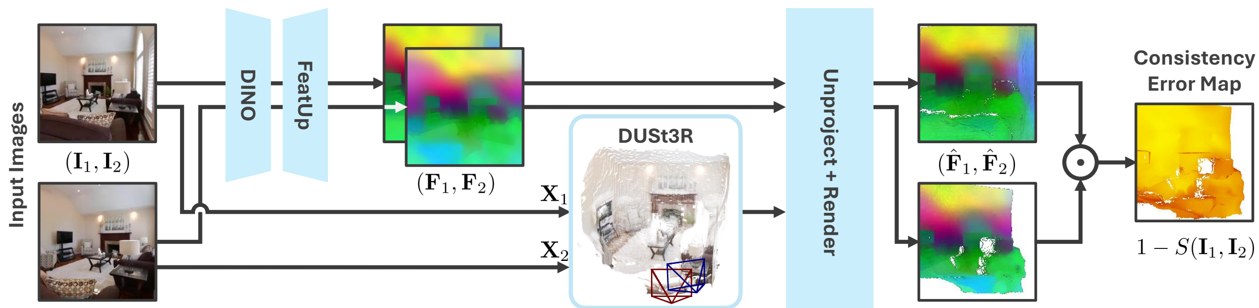


Figure 3. **Method overview.** Our metric evaluates the consistency between images $\mathbf{I}_1$ and $\mathbf{I}_2$. Given such a pair, we apply DUSt3R to obtain dense 3D point maps $\mathbf{X}_1$ and $\mathbf{X}_2$. These point maps are used to project upscaled DINO features $\mathbf{F}_1$, $\mathbf{F}_2$ into the coordinate frame of $\mathbf{I}_1$, via unprojecting and rendering. We compare the resulting feature maps $\hat{\mathbf{F}}_1$ and $\hat{\mathbf{F}}_2$ in pixel space to obtain similarity $S(\mathbf{I}_1, \mathbf{I}_2)$.

maps are of low resolution and do not represent detailed structures, we upsample them using FeatUp [10], which employs an image-adaptive upsampling, i.e., a stack of Joint Bilateral Upsamplers (JBUs) that learned to upsample low-resolution feature maps from DINO. It uses the high resolution image to transfer high frequency information to the upsampling process, allowing the upsampled features to faithfully reconstruct and preserve important details.

Let $\mathbf{F}_1$ and $\mathbf{F}_2$ denote the upsampled DINO features from images $\mathbf{I}_1$ and $\mathbf{I}_2$, respectively. We unproject both features into 3D space using the DUSt3R point maps and subsequently reproject them onto the camera frame of $\mathbf{I}_1$:

$$\hat{\mathbf{F}}_1 = \mathcal{P}(\mathbf{F}_1, \mathbf{X}_1), \qquad \hat{\mathbf{F}}_2 = \mathcal{P}(\mathbf{F}_2, \mathbf{X}_2), \tag{3}$$

where $\mathcal{P}$ assigns each 3D point the feature vector from its corresponding pixel before rendering the feature point cloud using the PyTorch3D [18] point rasterizer.

Following the projections, we obtain $S(\mathbf{I}_1, \mathbf{I}_2)$ as the weighted sum of pixel-wise similarities between $\hat{\mathbf{F}}_1$ and $\hat{\mathbf{F}}_2$:

$$S(\mathbf{I}_1, \mathbf{I}_2) = \frac{1}{|\mathbf{M}|} \sum_i^W \sum_j^H m^{ij} \frac{\hat{f}_1^{ij} \cdot \hat{f}_2^{ij}}{\|\hat{f}_1^{ij}\| \|\hat{f}_2^{ij}\|}, \tag{4}$$

where $m^{ij} := [\mathbf{M}]_{ij}$ is a boolean mask representing the overlapping region, $\hat{f}_1^{ij} := [\hat{\mathbf{F}}_1]_{ij}$ and $\hat{f}_2^{ij} := [\hat{\mathbf{F}}_2]_{ij}$.

4. Multi-View Latent Diffusion Model

Additionally to our metric, we provide an open-source multi-view latent diffusion model (MV-LDM). It is inspired by the architecture of CAT3D [11], which is not publicly available. While CAT3D is trained on top of proprietary image/video diffusion models, we initialize our model with Stable Diffusion 2.1 [29]. For a detailed description of MV-LDM, we refer to the appendix Sec. A. Our code and model are publicly available for further research.

Architecture and Training. MV-LDM encodes images into a latent space using a pre-trained VAE encoder from Stable Diffusion. Then, ray maps are concatenated to the input latents, providing camera pose information. We take the 2D UNet architecture, add attention between views at each UNet block, and finetune the full model on RealEstate10k for 1.65M iterations.

Anchored Generation. We adopt the anchored generation strategy from CAT3D. When generating many views of a scene, the process starts by sampling four anchor images for widely distributed cameras, conditioned on a single input image. Then, in the second step, the remaining views are generated and conditioned on the closest anchor and the initial input image. The goal of the anchoring strategy is to prevent accumulating errors that often occur when generating target views autoregressively, conditioned on the previously generated views. When generating with anchors, the accumulation of errors can be effectively limited. We analyze the effect on consistency and image quality in Sec. 5.3.

5. Experiments

In this section, we evaluate MEt3R and existing generative models for multi-view and video generation. Specifically, we aim to answer the following questions:

Q1: Does MEt3R fulfill the requirements for a useful consistency metrics as stated in Sec. 2, and how does it fare against previous metrics?

Q2: How consistent are the outputs of existing generative models for multi-view and video generation of objects and scenes?

Q3: How do individual design choices in MEt3R influence the metric quality?

We begin by introducing the experimental setup in Sec. 5.1 before validating MEt3R (answering Q1) in Sec. 5.2. Then, we address Q2 in Sec. 5.3 and Q3 in Sec. 5.4.

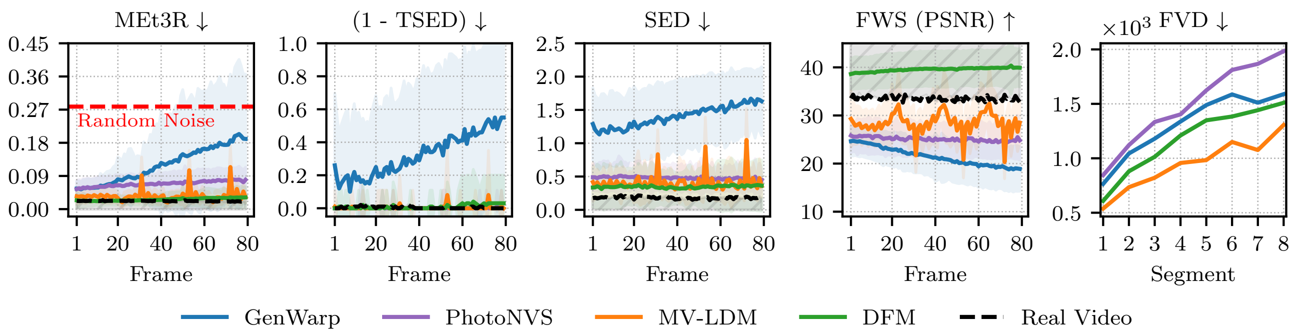


Figure 4. **Metric comparison.** We compare MEt3R against TSED, SED, FWS (PSNR), and FVD by computing average per-frame (/-segment for FVD) scores over many generated sequences. MEt3R can capture nuanced differences in the consistency of DFM, MV-LDM, and real videos, while TSED rates them all very similarly. Unlike MEt3R, SED does not capture increasing inconsistency for PhotoNVS and DFM. MEt3R also captures the influence of anchor views in MV-LDM (c.f. Sec. 4 and A.2) as structured high-frequency patterns. For MEt3R, the standard deviation gradually increases, starting from a small value. This behavior is expected due to the proximity of initial frames to the conditioning frame (c.f. Sec. 5.3.3) and is not the case for the other metrics.

5.1. Experimental Setup

To evaluate MEt3R, we consider three sets of baselines for multi-view, video, and object-level generation models. In addition, we categorize the multi-view generation methods into three general classes: 1) single-view, 2) multi-view, and 3) 3D diffusion models.

Multi-view Generation Models. We consider GenWarp [32], which is a single-view inpainting diffusion model, and PhotoNVS [48], which is a two-view generation model that generates a single view at a time conditioned on the previous. Moreover, we consider DFM [38], which is a 3D diffusion method, and MV-LDM, our own open-source multi-view latent diffusion model, coupled with cross-view attention (c.f. Sec. 4 and A). For more details on baselines, we refer to Sec. D in the appendix.

Video Generation Models. We take Stable Video Diffusion (SVD) [2], Ruyi-Mini-7B [36] and I2VGen-XL [49], which are standard open source video diffusion models that can generate a full video from a single input image.

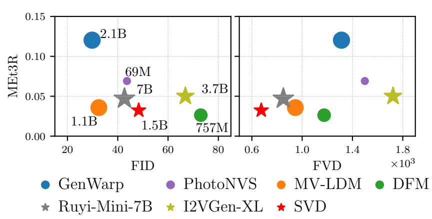
Object-Level Generation Models. From object-level methods, we compare EpiDiff [16], SyncDreamer [22], and VideoMV [53]. EpiDiff and SyncDreamer employ an underlying multi-view diffusion model, while VideoMV uses a video diffusion model to generate novel views of objects.

Dataset. To faithfully benchmark with MEt3R, we collect 100 image sequences from the RealEstate10K [51] test set. We take the first image for each sequence as the initial input, followed by 80 target poses, which the multi-view generation models generate. We perform consecutive pairwise evaluations on the generated images in a sliding-window fashion. In this way, we: 1) allow maximal projection area and more overlapping pixels to evaluate; 2) cover regions that are extrapolated and not visible in the input image; and 3) investigate the evolution of pairwise consistency as the camera pair moves further away from the input image. We set a standard resolution of 256^2 as input to MEt3R. In the case of DFM, we upsample from 128^2, and for GenWarp, we downsample from 512^2 bilinearly. Similarly, we use identical test sequences for video diffusion models but limit them to 48 frames to address memory constraints. Note that we do not have explicit camera control over the generation and, therefore, are not equivalent in camera trajectories. The generated videos also differ in resolution, which we resize accordingly to the closest resolution of 256^2 while preserving the aspect ratio. For object-level methods, we use Google Scanned Objects (GSO) [9] dataset, which consists of 360° views of objects. We subdivide the range $[0°, 360°]$ into 16 frames at 256^2 resolution, forming a closed loop with the frame at 0° as the conditioning. Both EpiDiff and SyncDreamer generate 16 frames, while VideoMV generates a fixed set of 32 frames, which we uniformly downsample to 16.

5.2. Validating MEt3R

Computing lower bound. We validate the efficacy of MEt3R by computing the lower bound that the baselines must follow. Intuitively, we can evaluate MEt3R on a dataset of real video sequences. Although they are assumed to be perfectly 3D consistent, a lower bound slightly above zero is observed, attributed to errors in point map alignment from DUSt3R [42] and small 3D inconsistencies in DINO [4] features. The results for both multi-view generation baselines and real videos are shown in Fig. 4.

Figure 5. **Qualitative comparison of generated novel views.** We compare generated views of the multi-view generation method for the same conditioning view. We can extract certain characteristics: DFM is almost perfectly consistent but has lower image quality. PhotoNVS and MV-LDM are reasonably consistent on a structural scale but fail to produce consistent details. GenWarp fails to keep the structural consistency over the sequence while producing high-quality images. These observations are confirmed by MEt3R in Tab. 1 and Fig. 4.



(a) Quality vs. Consistency, Number of Parameters

Methods	MEt3R ↓	TSED ↑	SED ↓	FVD ↓	FID ↓	FWS (PSNR) ↑
GenWarp [32]	0.120	0.674	1.398	1312.7	**29.80**	21.41
PhotoNVS [48]	0.069	<u>0.996</u>	0.479	1498.7	43.67	25.10
MV-LDM (**Ours**)	<u>0.036</u>	**0.998**	<u>0.405</u>	**945.8**	<u>37.29</u>	<u>28.46</u>
DFM [38]	**0.026**	0.990	**0.346**	<u>1174.6</u>	73.02	**39.56**
I2VGen-XL [49]	0.050	-	-	1722.6	66.88	<u>28.62</u>
Ruyi-Mini-7B [36]	<u>0.047</u>	-	-	<u>850.5</u>	**42.67**	28.01
SVD [2]	**0.032**	-	-	**674.6**	<u>48.33</u>	**29.93**

(b) Quantitative comparison with different metrics

Table 1. **Quantitative comparison.** Average MEt3R alongside TSED [48], SED [48], FVD [40], FID [13], and FWS (PSNR). (a) Plot comparing MEt3R with FID and FVD. (b) Quantitative comparison of multi-view and video generation baselines. Among multi-view methods, DFM achieves the best consistency in MEt3R, FWS (PSNR), and SED but the worst in FID. We attribute the low FID and high PSNR to blurry renderings, both of which are sensitive, whereas GenWarp delivers the best image quality with worse consistency but with a lower signal-to-noise ratio. In contrast, our MV-LDM achieves a favorable position in the image quality vs. consistency trade-off for multi-view generation. Unlike TSED and SED, MEt3R applies to generated video as it does not require camera poses.

Comparison to other Metrics. We compare MEt3R with existing metrics to measure 3D consistency. As baselines, we consider SED [48], TSED [48], and FVD [39] for multi-view generation. We also compare with several variants of flow warping score (FWS) using RAFT [37] to warp one frame to another and compute PSNR, SSIM, LPIPS, and RMSE, among which we show the results for PSNR in Fig. 4. For the remaining, we refer to Sec. B in the appendix. In Fig. 4, we plot per image-pair scores for all generated frames, averaged over 100 sequences. For FVD, we compare the distributions of image segments by splitting the sequences into chunks of 10 frames each. We find that MEt3R, SED, FWS (PSNR), and FVD increase as we progress through the image-pair sequence, suggesting a decrease in consistency, which is qualitatively visible in Fig. 5. Although TSED captures this trend for GenWarp [32], it does not report a meaningful separation for other baselines. Unlike TSED and SED, MEt3R captures the gradual decrease in consistency for PhotoNVS [48] and MV-LDM. For GenWarp, MEt3R captures this trend more accurately, starting with a lower score and standard deviation, as the first frame provides stronger conditioning for closer views with a larger overlap, resulting in better consistency. Furthermore, we observe sudden periodic spikes for MV-LDM in MEt3R, FWS (PSNR), and SED, attributed to transition artifacts when we switch between anchors during sampling (c.f. Sec. 4 and A.2). Unlike all other metrics, FVD cannot be applied to image pairs and requires a collection of frames. Ideally, a larger sample size is preferred to accurately capture and compare the underlying distribution of the generated and ground-truth image sequences [39], to which FVD is sensitive. Moreover, both FVD and FWS (PSNR) are sensitive to blur. Specifically, DFM achieves worse FVD, which is supposed to be 3D consistent by design (c.f. Sec. 5.1 and D), while the real video, which is perfectly 3D consistent, gets worse than DFM in PSNR, as shown in Fig. 4.

5.3. Evaluations of Models

5.3.1 Multi-View Generation

Following the validation of MEt3R in comparison to other metrics, we now benchmark our multi-view generation baselines on the test sequences (c.f. Sec. 5.1). In Tab. 1(a), we plot MEt3R against FID [13] and FVD [40] along with the respective model size in terms of the number of parameters. We find that GenWarp [32] achieves the worst consistency in terms of MEt3R, where the contents of the scene change drastically as we transition from one image to another, which can be qualitatively observed in Figs. 5, 17 - 21. This behavior is expected since GenWarp generates one image at a time. Meanwhile, PhotoNVS [48] performs slightly better than GenWarp but produces low-quality results, which the FID captures quantitatively. GenWarp and PhotoNVS cannot learn an expressive multi-view prior since they have a single-sized context window, hindering their ability to produce consistent 3D results.

Conversely, diffusing multiple views at a time induces a stronger prior towards 3D consistency, as in MV-LDM, where we see an overall improvement in MEt3R. Among all evaluated methods, MV-LDM achieves the best trade-off between 3D consistency and novel view quality, both qualitatively and quantitatively. Moving further towards 3D consistency, DFM [38] uses an underlying 3D representation and produces consistent novel views by design, which is captured quantitatively in the form of better MEt3R scores than MV-LDM. However, this strong inductive bias comes at the cost of blurry renderings pushing further from the ground-truth distribution, as reflected by the FID, and achieves a higher signal-to-noise ratio, as indicated by FWS (PSNR). This highlights that MEt3R only focuses on 3D consistency irrespective of image content and can, therefore, complement standard image quality metrics well.

5.3.2 Video Generation

A particular advantage of MEt3R is that it does not require camera poses, unlike TSED [48] and SED [48], where it can be used directly on generated videos to measure consistency similar to FWS (PSNR).

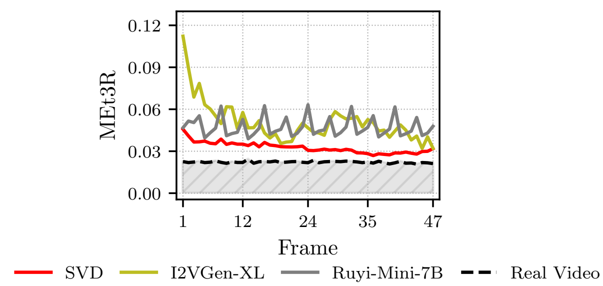


Figure 6. **Average pairwise MEt3R on generated videos.** Per-image-pair plot for MEt3R across 48 frames and averaged across 100 sequences of RealEstate10K [51]. We find that SVD [2] achieves the best MEt3R score, followed by Ruyi-Mini-7B [36] and I2VGenXL [49].

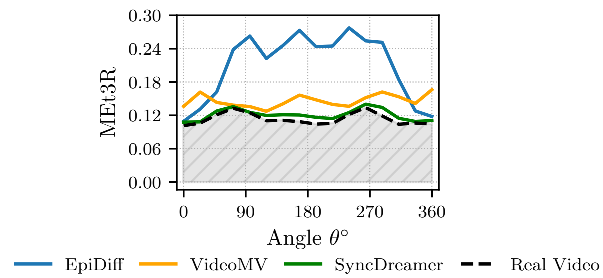


Figure 7. **Object-level evaluation on GSO [9].** Average pairwise MEt3R on 30 examples, each consisting of 360° rotation around the object with loop closure, i.e., at 360°, we evaluate the first and the last frame. We find that SyncDreamer [22] achieves the best in MEt3R, followed by VideoMV [53] and EpiDiff [16] while respecting the lower bound.

Table 1 shows the average MEt3R, FID, FWS (PSNR), and FVD. Moreover, Fig. 6 shows the average MEt3R per image pair for I2VGen-XL [49], Ruyi-Mini-7B [36] and SVD[2] which shows that SVD has better 3D consistency than Ruyi-Mini-7B and I2VGen-XL. However, SVD generates smoother and shorter camera trajectories, whereas Ruyi-Mini-7B and I2VGen-XL produce large motion at the expense of 3D consistency. For I2VGen-XL, as the inputs are out of distribution, MEt3R starts from a higher value followed by a gradual improvement as the model forces each progressing sample to be more in distribution while preserving similar global structures as in the initial input image. This behavior is also qualitatively visible in Figs. 17 - 21. Meanwhile, Ruyi-Mini-7B shows several spikes indicating abrupt inconsistencies throughout the video sequence, which are attributed to unstable camera motion. Furthermore, in Tab. 1, FWS (PSNR) gives the lowest score to Ruyi-Mini-7B, indicating a slightly noisy generation even though it is more 3D consistent than I2VGen-XL.

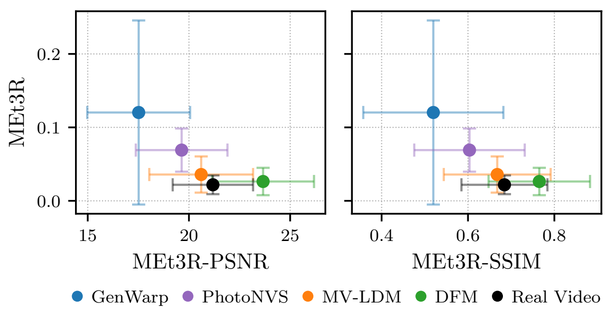


Figure 8. **Feature similarity ablation.** We compare MEt3R against versions of it that compare RGB projections via PSNR and SSIM. It can be seen that unlike MEt3R, PSNR and SSIM give better scores to DFM than to real videos. We attribute this to their sensitivity to view-dependent effects, such as lighting. Note that for real videos, the standard deviation of PSNR and SSIM are much higher, indicating a lower signal-to-noise ratio.

5.3.3 Object-Level Generation

Lastly, we evaluate object-level diffusion models on GSO [9] dataset. We consider EpiDiff [16], VideoMV [53] and SyncDreamer [22] as the baselines and are evaluated on 360° camera rotation spread across 16 frames on 30 examples. We find that MEt3R can differentiate models with various levels of inconsistencies, which we report in Fig. 7. Although the baselines show good visual quality, consistency varies heavily. Notice how MEt3R captures the slightly increasing inconsistency for EpiDiff when moving further away from the condition, which suggests that the strength of conditioning plays an important role in producing better consistency (c.f. Fig. 4 and Sec. 5.2). Further qualitative results are provided in Figs. 22 - 25 in the appendix.

5.4. Analyzing Alternative Similarities

We evaluate alternatives to the cosine similarity between DINO features as described in Sec. 3.2.

Image Similarity. Instead of projecting features onto a shared view, staying in RGB space would enable the use of classical image quality metrics such as PSNR and SSIM. Fig. 8 provides a comparison of such variants $\text{MEt3R}_{\text{PSNR}}$ and $\text{MEt3R}_{\text{PSNR}}$ with MEt3R. While a reasonable negative correlation can be observed, DFM [38] outperforms the ground-truth video w.r.t. these metrics. We attribute this to the bias of PSNR and SSIM to blur, which is apparent in novel views generated by DFM due to its low resolution and reliance on pixelNeRF [47] acting as an architectural bottleneck. In contrast, real videos exhibit view-dependent effects, including brightness variations and reflections, to which PSNR and SSIM are highly sensitive. With MEt3R, we aim to abstract from these pixel-level inconsistencies and instead provide a metric that robustly measures the 3D consistency of generative approaches. Therefore, we opt for similarities in a suitable feature space.

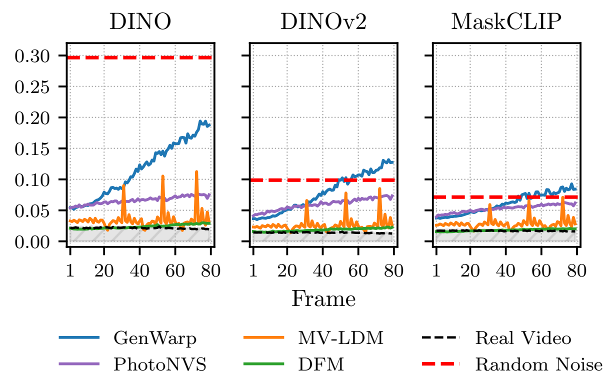


Figure 9. **Feature backbone ablation.** We analyze the effect of different feature backbones on MEt3R. While DINOv2 [27] and MaskCLIP [50] can be employed as well, we found DINO [4] features to lead to a more informative separation of models.

Feature Backbones. In Fig. 9, we evaluate MEt3R in combination with DINOv2 [27] and MaskCLIP [50] as alternatives to DINO [4] in the feature backbone. DINOv2 and MaskCLIP strongly compress the values in a tighter range, reducing the gap between extremely inconsistent and consistent generation. We find that DINO features provide a better separation of model performance and capture substantial inconsistencies more reliably, as seen from the random noise. Nevertheless, MEt3R is flexible with this design choice as better and more 3D consistent feature backbones can improve and reduce the lower bound further.

6. Conclusion

We presented MEt3R, a novel metric for 3D consistency of generated multi-view images. Given the huge success of large-scale image diffusion models and their applications as strong priors for the generation of multi-view images as a form of 3D representation, purely distribution-based metrics like FVD are insufficient to properly evaluate the 3D capabilites of such methods. First, MEt3R leverages DUSt3R to warp images robustly into a shared view without relying on ground truth camera poses as input. Secondly, by computing similarities in the feature space of DINO, MEt3R abstracts from view-dependent effects. As a result, we show that our proposed metric can be effectively employed for comparing the performance of multi-view generation approaches like our open-source multi-view latent diffusion model, which finds the best trade-off between novel view quality and consistency. Given the recent trend towards large video models, we see great potential for MEt3R to effectively evaluate their 3D consistency since no ground truth camera poses are required.

Acknowledgements

This project was partially funded by the Saarland/Intel Joint Program on the Future of Graphics and Media. Thomas Wimmer is supported through the Max Planck ETH Center for Learning Systems.

References

[1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 1, 3

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 5, 6, 7, 14

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 22563–22575. IEEE, 2023. 1

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021. 2, 3, 5, 8

[5] 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. GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In *ICCV,*, 2023. 2

[6] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion NeRF: A unified approach to 3d generation and reconstruction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023. 1

[7] 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 *arxiv,*, 2022. 2

[8] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[9] 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, 2022. 5, 7, 8

[10] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T Freeman. Featup: A model-agnostic framework for features at any resolution. *arXiv preprint arXiv:2403.10516*, 2024. 2, 3, 4

[11] Ruiqi Gao*, Aleksander Holynski*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole*. Cat3d: Create anything in 3d with multi-view diffusion models. *NeurIPS,*, 2024. 1, 2, 4, 12

[12] Junlin Han, Filippos Kokkinos, and Philip Torr. Vfusion3d: Learning scalable 3d generative models from video diffusion models. In *European Conference on Computer Vision*, pages 333–350. Springer, 2025. 2

[13] 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. In *NeurIPS,*, 2017. 1, 3, 6, 7

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS,*, 2020. 1, 12

[15] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 1

[16] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9784–9794, 2024. 5, 7, 8

[17] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 9307–9315. IEEE, 2024. 1, 3

[18] Justin Johnson, Nikhila Ravi, Jeremy Reizenstein, David Novotny, Shubham Tulsiani, Christoph Lassner, and Steve Branson. Accelerating 3d deep learning with pytorch3d. In *SIGGRAPH Asia 2020 Courses*, page 1–1. ACM, 2020. 4

[19] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 3

[20] 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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 9264–9275. IEEE, 2023. 2

[21] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR,*, 2023. 1

[22] 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. In *ICLR,*, 2024. 2, 5, 7, 8

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 12

[24] 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*, page 405–421. Springer International Publishing, 2020. 3

[25] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, and Matthias Nießner. DiffRF: Rendering-guided 3d radiance field diffusion. In *2023 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR)*, page 4328–4338. IEEE, 2023. 1

[26] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. MultiDiff: Consistent novel view synthesis from a single image. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 10258–10268. IEEE, 2024. 2

[27] 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. 8

[28] Robin Rombach, Patrick Esser, and Bjorn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 14336–14346. IEEE, 2021. 2, 3

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 10674–10685. IEEE, 2022. 1, 2, 4, 12, 13

[30] 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

[31] Philipp Schröppel, Christopher Wewer, Jan Eric Lenssen, Eddy Ilg, and Thomas Brox. Neural point cloud diffusion for disentangled 3d shape and appearance generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 8785–8794. IEEE, 2024. 1

[32] Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. GenWarp: Single image to novel views with semantic-preserving generative warping. *arXiv preprint arXiv:2405.17251*, 2024. 2, 3, 5, 6, 7, 14, 15, 16

[33] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3d generation. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 12

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR,*, 2021. 1, 12

[35] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[36] CreateAI Team. Ruyi-mini-7b. `https://github.com/IamCreateAI/Ruyi-Models`, 2024. 5, 6, 7, 14

[37] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020. 6, 13

[38] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Joshua B. Tenenbaum, Frédo Durand, William T. Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. In *arXiv*, 2023. 1, 2, 3, 5, 6, 7, 8, 14, 15, 16

[39] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 3, 6, 7

[40] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation, 2019. 6, 7

[41] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2025. 1, 2

[42] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 20697–20709. IEEE, 2024. 2, 3, 5, 14, 15

[43] 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

[44] Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentSplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. In *arxiv,*, 2024. 2

[45] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Hołyński. ReconFusion: 3d reconstruction with diffusion priors. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 21551–21561. IEEE, 2024. 1, 2

[46] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*, 2024. 3

[47] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR,*, 2021. 8, 16

[48] Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 7071–7081. IEEE, 2023. 2, 3, 5, 6, 7, 13, 14, 15, 16

[49] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models, 2023. 5, 6, 7, 14

[50] Zhuowen Tu Zheng Ding, Jieke Wang. Open-vocabulary universal image segmentation with MaskCLIP. In *International Conference on Machine Learning*, 2023. 8

[51] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH,*, 2018. 5, 7, 12

[52] Zhizhuo Zhou and Shubham Tulsiani. SparseFusion: Distilling view-conditioned diffusion for 3d reconstruction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 12588–12597. IEEE, 2023. 1, 2

[53] Qi Zuo, Xiaodong Gu, Lingteng Qiu, Yuan Dong, Zhengyi Zhao, Weihao Yuan, Rui Peng, Siyu Zhu, Zilong Dong, Liefeng Bo, and Qixing Huang. Videomv: Consistent multi-view generation based on large video generative model, 2024. 5, 7, 8