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
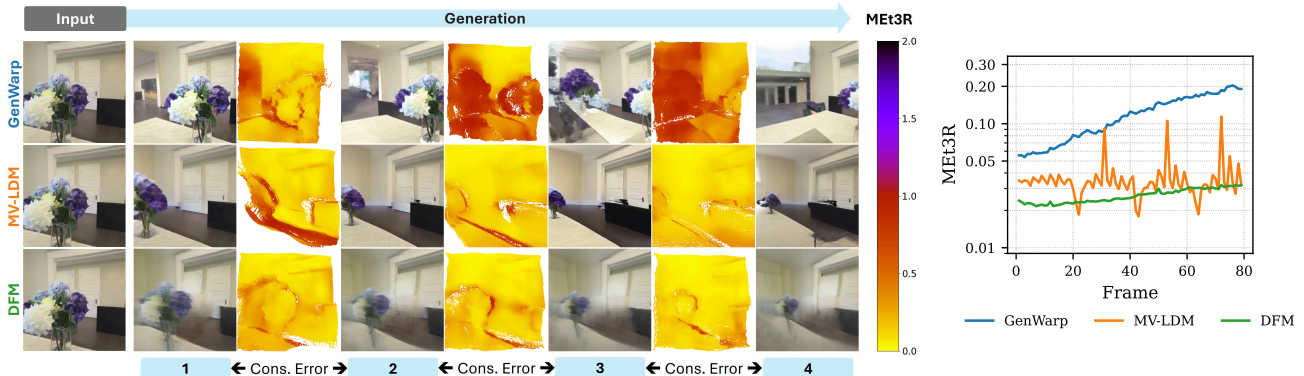
{masim, jlenssen}@mpi-inf.mpg.de

Figure 1. We introduce **MEt3R**, a metric for multi-view consistency between pairs of generated images, which is independent of image quality, image content, and does not require camera poses. **Left**: generated images from different generative models, conditioned on the first frame, with MEt3R score map 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 [13, 33] or flow-based [20] 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 [10, 36, 43]. However, they raise the inherent challenge that for individual samples no ground truth is available to measure the quality of generations with pair-wise distance metrics. As a result, metrics such as FID [12], KID [1] and CMMD [15] have been proposed to measure the quality of generated images without the need of a paired ground truth.

Recently, a trend is to repurpose video [3, 14] and image [28, 34] diffusion models for generation of 3D scenes and objects, by generating multiple views from different camera poses [10, 39, 50], with or without given images as conditioning. In comparison to direct generation of 3D representations [6, 24, 30], 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 results,
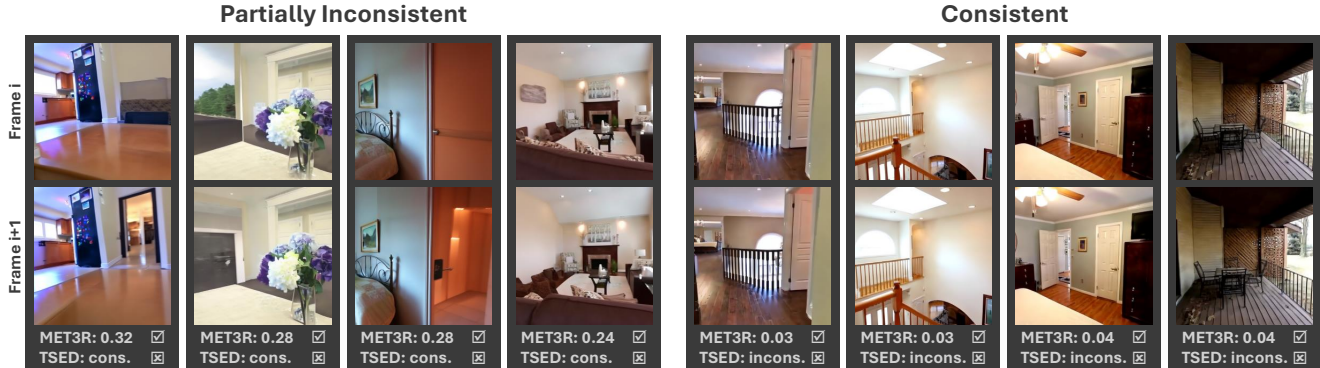
Figure 2. **Existing metrics.** A comparison between MEt3R and TSED [46] scores obtained from individual image pairs generated by GenWarp [31]. TSED misses obvious, partial multi-view inconsistencies and is biased to small violations of epipolar geometry. In contrast, MEt3R correctly captures clear 3D inconsistencies and is robust to insignificant artifacts almost invisible to the human eye.

which is of large importance for the subsequent lift into 3D. A reliable metric to evaluate the multi-view consistency of such generations is desperately needed to advance these models further. Luckily, similar to general image quality, 3D consistency between views can be evaluated without the existence of paired ground truth data. Existing metrics [46] 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 [40] 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 extractor, DINO [4] + FeatUp [9] are used to obtain features that are more robust to view-dependent effects, such as lighting, and which can be compared quantitatively. We further introduce an open-source multi-view latent diffusion model (MV-LDM) to be used in our studies, which is able to generate high-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 multi-view images of scenes, with and without intermediate 3D representation, as well as our MV-LDM. We can 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 expectations of measuring consistency. In contrast to previous metrics, it can distinguish perfectly consistent from almost consistent sequences and can capture fine-grained changes in consistency over time.

In summary, our **contributions** include:

- the first metric for measuring multi-view consistency of generated views without given camera poses,

- a comprehensive analysis of existing methods that generate multi-view images of scenes and videos, 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 [28] 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, 10, 11, 19, 21, 25, 27, 31, 32, 36, 39, 43, 46, 50].

3D-aware image generation methods can be grouped into methods for pose-conditioned single-view generation [19, 31, 41, 46, 50], simultaneous multi-view image generation [10, 27, 32] and methods that use an internal 3D representation of the scene as prior for generation [5, 21, 36, 42, 43]. Further distinction can be made between models that are trained on single-asset 3D datasets [19, 21, 32, 41], such as Objaverse [7], and models trained on full 3D scenes [5, 10, 27, 31, 36, 43, 46, 50]. 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 [31, 46], multiple output views [27], and methods that use an internal 3D representation [36] to enforce consistency.
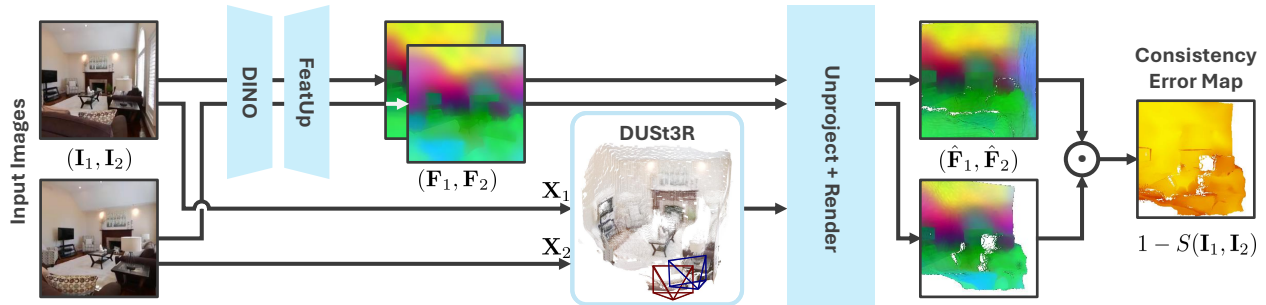
Figure 3. **Method overview.** Our metric evaluates the consistency between images $\mathbf{I}_1$ and $\mathbf{I}_2$. Given such a pair, we apply DUSt3R to obtain dense 3D point maps $\mathbf{X}_1$ and $\mathbf{X}_2$. These point maps are used to project upscaled DINO features $\mathbf{F}_1$, $\mathbf{F}_2$ 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 $S(\mathbf{I}_1, \mathbf{I}_2)$.

Existing Metrics. Existing metrics used for quantifying image generation outputs include distribution-based metrics, such as the Fréchet Inception Distance (FID) [12], Kernel Inception Distance (KID) [1], Inception Score (IS) [29], or the CLIP Maximum Mean Discrepancy (CMMD) [16]. While these metrics are used to measure the alignment of generated samples with a target distribution using pre-trained feature extractors, they do not measure 3D consistency, which is of utmost importance for multi-view generative models. To this end, Xie et al. [44] proposed using the Fréchet Video Distance (FVD) [37] to measure the quality of generated sequences with moving camera.

To explicitly measure 3D consistency, Watson et al. [41] proposed to train a NeRF [23] 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. [46] 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) pe-

nalize 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 [12], 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 [9]. 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 1st 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 image pair $\mathbf{I}_1$, $\mathbf{I}_2$, the DUSt3R [40] 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 [40].

DUSt3R does not require camera poses, which is inherited by MEt3R. While MASt3R [18] 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 maps are of low resolution and do not represent detailed structures, we upsample them using FeatUp [9], 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. Then, we unproject both features into 3D space using the DUSt3R point maps and subsequently reproject them onto the camera frame of $\mathbf{I}_1$:

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

where $\mathcal{P}$ assigns each 3D point the feature vector from its corresponding pixel before rendering the feature point cloud using the PyTorch3D [17] 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 $\hat{\mathbf{F}}_2$:

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

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 [10], which is not publicly available. While CAT3D is trained on top of proprietary image/video diffusion models, we initialize our model with StableDiffusion and train it on the openly available dataset RealEstate10k [49]. 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 StableDiffusion 2.1 [28]. Then, camera ray encodings are concatenated to the latent images, providing camera pose information before being processed by the diffusion model. We take a pre-trained StableDiffusion UNet model, add attention between views in the latent space, and finetune it on RealEstate10k videos for 2M iterations. MV-LDM works with a total of 5 views at a time consisting of N conditioning and M target views.

Anchored Generation. We adopt the anchored generation strategy from CAT3D [10]. When generating many views of a scene, the generation process starts with sampling 4 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 images along with 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?

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.
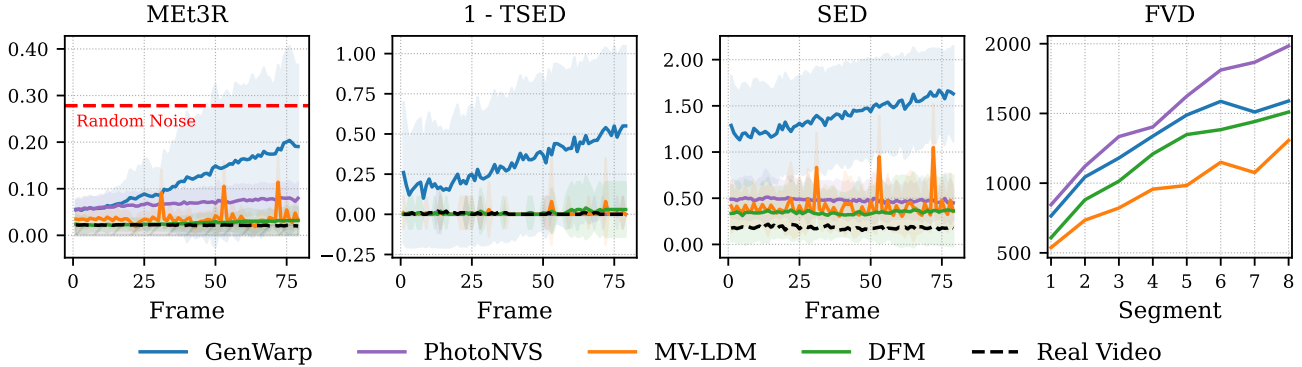
4

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

5.1. Experimental Setup

To evaluate MEt3R, we consider two sets of baselines for multi-view and video 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 [31], which is a single view image-to-image inpainting diffusion model, and PhotoNVS [46], which is an autoregressive multi-view generation model that generates a single view at a time conditioned on the previous. Moreover, we consider DFM [36], which is a 3D diffusion method that incorporates a neural radiance field into the architecture of an image diffusion model, forcing the rendered novel views to be 3D consistent by design. Finally, MV-LDM, our own open-source multi-view diffusion model, coupled with cross-view attention (c.f. Sec. 4), generates multiple novel views at a time, resulting in a stronger 3D prior than single-view methods, i.e., GenWarp and PhotoNVS. We refer to appendix Sec. C for further details on these baselines.

Video Generation Models   We take Stable Video Diffusion (SVD) [2], Ruyi-Mini-7B [35] and I2VGen-XL [47], which are standard open source video diffusion models that can generate videos with a single input image with an additional text prompt.

Dataset   To faithfully benchmark with MEt3R, we collect 100 image sequences from the RealEstate10K [49] test set. We take the first image for each sequence as the initial input, followed by 80 target poses, which the multi-view gen-

eration 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 case of DFM [36], we upsample from 128^2, and for GenWarp[31], we downsample from 512^2 bilinearly. Similarly, we use identical test sequences for video diffusion models but limit the number of generated frames to 48 due to memory restrictions. Note that we do not have explicit camera control over the generation and, therefore, are not equivalent in camera trajectories. In addition, the generated videos differ in resolution and aspect ratios, which we resize accordingly to the closest resolution of 256^2 while maintaining the aspect ratio.

5.2. Validating MEt3R

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

Comparison to other Metrics.   We compare MEt3R with existing metrics to measure 3D consistency. As baselines,

Sequence 1   **Sequence 2**

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.
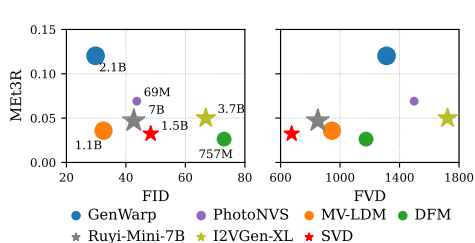
(a) Quality vs. Consistency, Number of Parameters

Methods	MEt3R ↓	TSED ↑	SED ↓	FVD ↓	FID ↓	KID ↓
GenWarp [31]	0.120	0.674	1.398	1312.7	**29.80**	**0.0033**
PhotoNVS [46]	0.069	0.996	0.479	1498.7	43.67	0.0081
MV-LDM (**Ours**)	0.036	**0.998**	0.405	**945.8**	37.29	0.0064
DFM [36]	**0.026**	0.990	**0.346**	1174.6	73.02	0.0324
I2VGen-XL [47]	0.051	-	-	1722.6	66.88	0.0161
Ruyi-Mini-7B [35]	0.047	-	-	850.5	**42.67**	**0.0071**
SVD [2]	**0.033**	-	-	674.6	48.33	0.0111

(b) Quantitative comparison with different metrics

Table 1. **Quantitative comparison.** Average MEt3R alongside FID [12], KID [1], FVD [38], SED [46] and TSED [46]. (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 and SED but the worst in FID and KID due to their sensitivity to blur artifacts, aligning with the visual impression in Figs. 5, 15 - 19. While GenWarp delivers the best image quality, it has the worst consistency. 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. Though a weak correlation between MEt3R and FVD is observed in video generation, this does not extend to multi-view generation. As FVD assesses the overall quality and temporal coherency across several videos, MEt3R evaluates pairwise 3D consistency within individual video sequences without relying on a ground truth dataset.
.

we consider SED [46], TSED [46], and FVD [37] for multi-view generation. 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, and FVD increase as we progress through the image-pair sequence, suggesting a decrease in consistency, which is also qualitatively visible in Fig. 5. Although TSED captures this trend for GenWarp [31], it does not report a meaningful separation for other baselines. Un-

like TSED and SED, MEt3R captures the gradual decrease in consistency for PhotoNVS [46] and MV-LDM. For Gen-Warp, 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 and SED, attributed to transition artifacts when we switch between anchors during sampling (c.f. Sec. 4 and A.2). Unlike all other metrics, DFM [36] is consistently worse than

MV-LDM in terms of FVD. Since DFM is supposed to be 3D consistent by design (c.f. Sec. 5.1 and C), this shows that FVD is sensitive to blurry samples and does not evaluate multi-view consistency. Ideally, a larger sample size is preferred to accurately capture and compare the underlying distribution of generated and ground-truth image sequences [37], to which FVD is sensitive. Therefore, it cannot be applied at the level of individual image pairs.

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 and KID along with the respective model size in terms of the number of parameters. We find that GenWarp 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, 15 - 19. This behavior is expected since GenWarp generates one image at a time. Meanwhile, PhotoNVS performs slightly better than GenWarp but produces low-quality results, which are captured quantitatively by FID and KID. GenWarp and PhotoNVS cannot learn an expressive multi-view prior since they are single-view generation models, hindering their ability to produce 3D consistent 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 uses an underlying 3D representation and, therefore, 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 with the drawback of blurry renderings far away from the ground-truth image distribution, as reflected by FID and KID. This highlights that MEt3R only focuses on 3D consistency irrespective of image content and can therefore complement standard image quality metrics well.

MEt3R on multiple scales. In Tab. 2, we investigate the effect of image resolutions on MEt3R compared to SED [46]. We find that SED is highly sensitive to variation in image resolution with a significant increase at 128^2. This is expected since SED computes the geometric distance of each correspondence from their epipolar line in the 2D-pixel space. Meanwhile, MEt3R is more robust, attributed to the measurement in the feature space (c.f. Sec. 3), thus maintaining only minor differences in the scores. Although
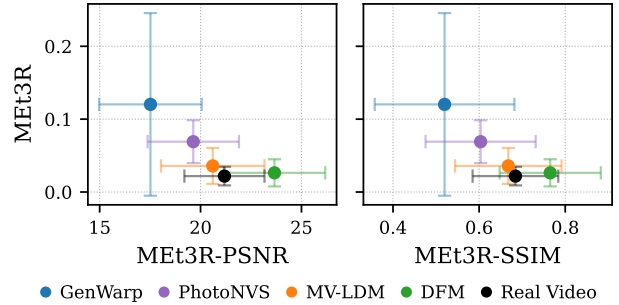

Figure 6. **Feature similarity ablation.** We compare MEt3R against versions of it that compare RGB projections via PSNR and SSIM. It can be seen that the PSNR versions give better scores to DFM than to real videos. We attribute this to their sensitivity to view-dependent effects, such as lighting. In contrast, MEt3R rates the real video best. Further, the standard deviation of PSNR and SSIM versions are much higher, also for real videos, indicating a lower signal-to-noise ratio.

	256^2		224^2		128^2	
	MEt3R	SED	$Diff_{MEt3R}$	$Diff_{SED}$	$Diff_{MEt3R}$	$Diff_{SED}$
GenWarp [31]	0.120	1.398	+2.58%	+74.70%	-6.95%	+80.27%
PhotoNVS [46]	0.069	0.479	+3.86%	+26.22%	-1.78%	+42.53%
MV-LDM (**Ours**)	0.036	0.405	+3.88%	-2.27%	-16.63%	+34.28%
DFM [36]	**0.026**	**0.346**	+9.39%	+52.73%	-5.97%	+70.15%
Real Video	0.022	0.181	+2.44%	+18.98%	-47.80%	+148.39%

Table 2. **MET3R vs. SED on multiple resolutions**. We show differences in SED [46] and MEt3R for the baseline multi-view generation models over changing image resolution against the base resolution of 256^2 in percentage. MEt3R is more robust to variations in the input resolution since it measures in feature space, unlike SED, which measures in pixel space (c.f. Sec. 3). Here, SEDs total scale is less than one order of magnitude larger than MEt3Rs, while its variations are more than one order of magnitude larger in most cases.

the differences are small, we still recommend using a similar resolution for all baselines for a fair comparison.

5.3.2 Video Generation

A particular advantage of MEt3R is that it does not require camera poses to measure consistency, unlike TSED [46] and SED [46]. Thus, it is the first quantitative multi-view consistency metric that can be used on videos generated by video diffusion models.

Table 1 shows the average MEt3R along with FID, KID, and FVD. Moreover, Fig. 7 shows the average MEt3R per image pair for I2VGen-XL [47], Ruyi-Mini-7B [35] and SVD[2] which clearly 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 mo-
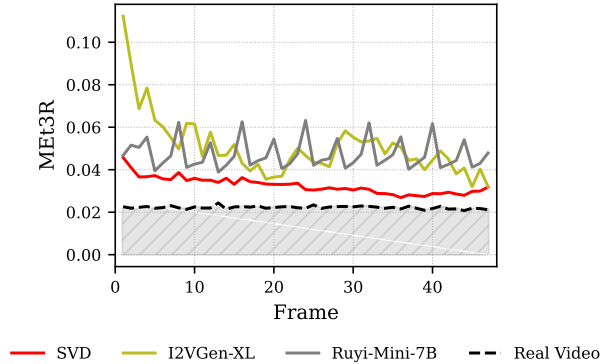
Figure 7. **MEt3R on generated videos.** Per-image-pair plot for MEt3R across 48 frames and averaged across 100 sequences of RealEstate10K. For I2VGen-XL, we observe large inconsistencies initially as the inputs are out-of-distribution, followed by gradual improvement, indicating the inputs get closer to being in-distribution, also visible qualitatively in Figs. 15 - 19. Meanwhile, Ruyi-Mini-7B shows several periodic spikes indicating abrupt inconsistencies throughout the video sequence, whereas MEt3R for SVD stays relatively low and smooth.
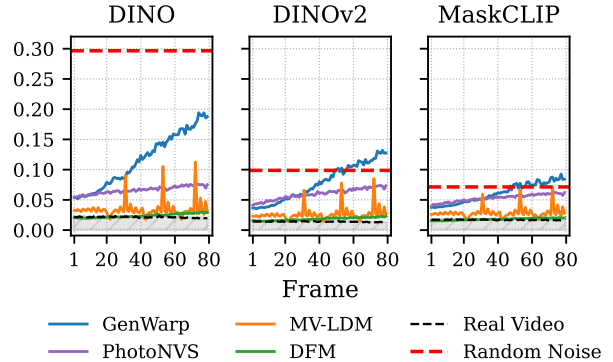

Figure 8. **Feature backbone ablation.** We analyze the effect of different feature backbones on MEt3R. While DINOv2 [26] and MaskCLIP [48] can be employed as well, we found DINO features to lead to a more informative separation of models.

tion at the expense of 3D consistency. Ruyi-Mini-7B is consistently worse, with periodic spikes, attributed to unstable camera motion and sudden 3D inconsistencies. 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. The resulting MEt3R scores correlate with visual judgment about the 3D consistency of the baseline video diffusion models, which can be observed in Figs. 15 - 19 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. 6 provides a comparison of such variants $\text{MEt3R}_{\text{PSNR}}$ and $\text{MEt3R}_{\text{SSIM}}$ with MEt3R. While a reasonable negative correlation can be observed, DFM [36] 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 [45] 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.

Feature Backbones. In Fig. 8, we evaluate MEt3R in combination with DINOv2 [26] and MaskCLIP [48] 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 be used to improve the overall metric and to further reduce the lower bound.

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

References

[1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 1, 3, 6

[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

[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, 4, 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. 4

[9] 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

[10] 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, 11

[11] 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

[12] 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

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS,*, 2020. 1, 11

[14] 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

[15] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. *CVPR,*, 2023. 1

[16] 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. 3

[17] 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

[18] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 4

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

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

[21] 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

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 11

[23] 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

[24] 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

[25] 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

[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. 8

[27] 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

[28] 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, 11, 12

[29] 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

[30] 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

[31] 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, 5, 6, 7, 13, 14

[32] 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, 11

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR,*, 2021. 1, 11

[34] 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

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

[36] 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, 5, 6, 7, 8, 13, 14

[37] 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

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

[39] 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

[40] 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, 4, 12

[41] 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

[42] 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

[43] 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

[44] 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

[45] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR,*, 2021. 8, 14

[46] 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

[47] 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

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

[49] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH,*, 2018. 4, 5, 11

[50] 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

MEt3R: Measuring Multi-View Consistency in Generated Images

Supplementary Material

The supplementary materials are structured as follows. First, we provide detailed information about our multi-view latent diffusion model (MV-LDM) in Sec. A. Then, we provide details about the MEt3R metric in Sec. B. Finally, we present additional details on the multi-view generation baselines in Sec. C and their corresponding runtime statistics in Sec. D. Please also note our **supplementary video**, showcasing evaluations in motion.

A. Multi-View Latent Diffusion Model

This section presents further details for MV-LDM, including the architectural components, training, and sampling details.

A.1. Architecture.

Like CAT3D [10], our architecture is based on a multi-view 2D UNet shared across multiple input views with 3D self-attention at each UNet block. We initialize the UNet weights with Stable Diffuson 2.1 [28] and replace each attention layer with a 3D self-attention layer from MVDream [32] where each token from one view attends to all tokens from the other views. This accounts for 1.1B parameters for the multi-view UNet and 83.7M for the VAE. Due to memory and resource limitations, we fix the total number of concurrent views to 5, including the target and the conditioning views. Figure 9 shows the architecture of MV-LDM. We apply a VAE encoder and map the input images $(H \times W \times 3)$ into latent representation $(\frac{H}{8} \times \frac{W}{8} \times 4)$. For the low-resolution latent maps, we generate the ray encodings of shape $(\frac{H}{8} \times \frac{W}{8} \times 6)$, which consists of a 3-dimensional origin and a 3-dimensional direction vector in relative camera space and concatenate it along the channel dimension.

A.2. Training and Evaluation with MEt3R

Dataset. We use RealEstate10K [49], which consists of 80K video sequences accounting for 10 million frames. During training, we randomly select a video sequence and the corresponding conditioning and target views that satisfy the following criteria:

- Sample 2 conditioning views (left and right) at frame number f_L and f_R with frame distance $d_c = f_R - f_L$ satisfying $50 \leq d_c \leq 180$.
- Sample 3 target views with distance d_t from the conditioning view that satisfies $f_L - 100 \leq d_t \leq f_R + 100$.

Afterward, we transform the absolute poses into relative poses with respect to the first conditioning view.

Training. The training procedure follows DDPM [13], sampling a noise level t, applying that to all given latent images and training the network to predict the noise present in the image. We randomly select the conditioning views N between 1 or 2 and the target views M between 3 and 4, respectively, to allow for single and few-view novel view generation. We linearly vary the beta schedule from 0.0001 to 0.02 for the forward diffusion process and train MV-LDM for a total of 1.65M iterations with an effective batch size of 24 at resolution 256^2. We use AdamW [22] optimizer with a constant learning rate of $2e^{-5}$. During sampling, the network can receive a combination of existing and pure noise images with camera ray encodings to perform conditional generation. The backward diffusion process is done with ϵ-parameterization defined as the output of the MV-LDM ϵ_θ as,

$$\epsilon_{pred} = \epsilon_\theta\left(\mathbf{z}_t, \mathbf{c}_t, t\right), \qquad (5)$$

where $\epsilon_{pred} = (\epsilon_{pred}^i)_{i=1}^M$ is the predicted noise latent, $\mathbf{z}_t = (\mathbf{z}_t^i)_{i=1}^M$ is the noisy latent, $\mathbf{c}_t = (\mathbf{c}_t^j)_{j=1}^N$ is the clean latent at the timestep t, whereas M and N are the number of target and conditioning views, respectively. The predicted noise is used to make a step in the direction of a sample in the target distribution under the DDIM [33] formulation. For classifier-free guidance, we randomly drop the clean conditioning views with a probability of 10%, and during sampling, we apply a guidance scale of 3 similar to CAT3D [10].

For training, we apply the standard diffusion loss on the predicted mean noise as the mean-squared error (MSE) against the ground truth noise:

$$\mathcal{L} = ||\epsilon - \epsilon_\theta\left(\mathbf{z}_t, \mathbf{c}_t, t\right)||_2^2, \qquad (6)$$

where $\epsilon = (\epsilon^i)_{i=1}^M$ and $\epsilon^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the ground truth noise for each target view.

Training evolution of MEt3R. Figure 10 shows the trend in 3D consistency in terms of MEt3R over training iterations, showing consistent improvements with longer training. There is a significant improvement in the initial 100k, and afterward, it saturates near 1M iterations.

Anchored vs. autoregressive sampling. We further test MEt3R with two sampling strategies, i.e., (1) autoregressively generating new target views and new anchors, conditioned on the previous anchor, and (2) using anchored sampling where we generate anchors first and then the rest as described in Sec. 4. Fig. 11 shows the average MEt3R
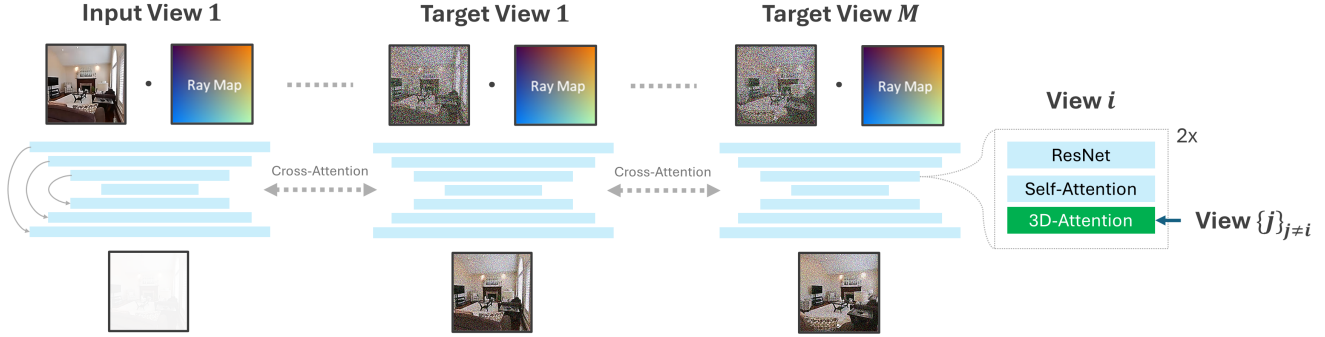
Figure 9. **MV-LDM.** Architecture overview of MV-LDM, which consists of a shared 2D UNet initialized from Stable Diffusion 2.1 [28] across multiple input views with cross-view attentions (3D attention) in between for modeling multi-view prior.
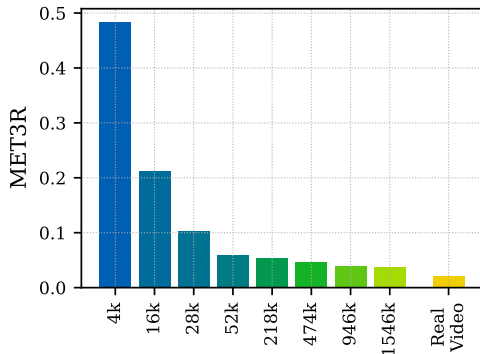

Figure 10. **MEt3R at different training iterations.** As we continue to train MV-LDM, we see a consistent improvement in 3D consistency, which is an expected behavior. Furthermore, in the beginning, the improvements are large, which slows down and saturates in the later iterations.

Figure 11. **Anchored vs. autoregressive.** Per-image-pair MEt3R on 2 different sampling strategies. For autoregressive sampling, we see significant and periodic spikes becoming larger as we progress and show the effect of compounding error, i.e., sequentially generating new frames and anchors conditioned on the previously generated ones. As illustrated in Fig. 12, autoregressive sampling produces several anchor-to-anchor transitions causing these periodic spikes. On the other hand, anchored generation limits the effect of compounding error by generating all anchors in parallel.

plot per image-pair, showing the improvements with anchored sampling. For autoregressive sampling, we observe many diverging peaks that refer to several anchor-to-anchor transitions and accumulating errors. In contrast, for anchored sampling, the anchors are generated together first, followed by generating the rest. This limits error accumulation and results in fewer anchor-to-anchor transitions. Refer to Fig. 12 for a visual illustration of anchored and autoregressive sampling schemes.

B. Additional MEt3R Architectural Details

This section presents additional details on the MEt3R pipeline, including the projection of both point maps to the first view and a description of the overlap mask used.

Projection matrix. Figure 13 shows a side-by-side comparison of different projections we obtain using 1): fixed focal length and 2): Adjusting focal length based on the
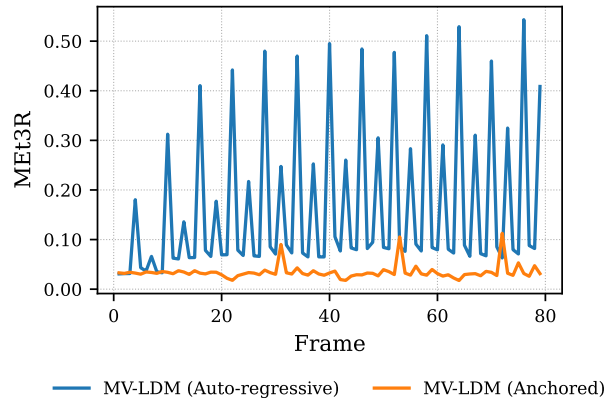
scale of canonical point map. We compute the canonical point map $\mathbf{X}_{canon}$ as the weighted sum of the point maps pair $\mathbf{X}_i$ and $\mathbf{X}_{i+1}$ using their corresponding confidences $\mathbf{C}_i$ and $\mathbf{C}_{i+1}$ from DUSt3R [40] as,

$$\mathbf{X}_{canon} = \frac{\mathbf{C}_i \odot \mathbf{X}_i + \mathbf{C}_{i+1} \odot \mathbf{X}_{i+1}}{\mathbf{C}_i + \mathbf{C}_{i+1}} \qquad (7)$$

Then, we extract the x, y, and z coordinate maps from $\mathbf{X}_{canon}$ as $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \in \mathbb{R}^{H \times W}$. Moreover, DUSt3R already implements this in their codebase, which we incorporate in MEt3R as shown in Alg. 1. The computed focal length f_x and f_y, along with the principle point offsets c_x and c_y, are used to form the projection matrix.
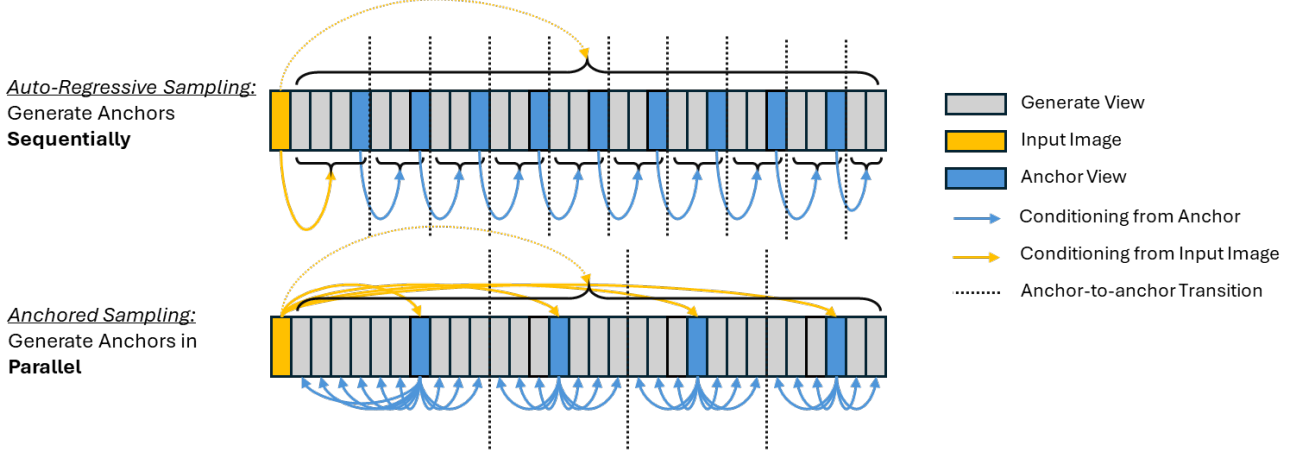
Figure 12. **Anchored vs. Autoregressive sampling schemes.** An illustration of the differences in the sampling schemes. In autoregressive sampling, we start from the initial input image and generate a set of target frames. The next set of frames is conditioned both on the input image and on the last frame (anchor) of the previously generated set. With this sampling strategy, we see several anchor-to-anchor transitions and results in large inconsistencies as visible in Fig. 11. Whereas using anchored generation i.e., generate anchors first and then sample the remaining conditioned on the closest anchor and the input image. With this strategy, we observe significantly fewer anchor-to-anchor transitions, limited error accumulation, and relatively stable and lower MEt3R across the image pairs.

Algorithm 1 Computing focal length given 2D grid of pixel positions and 3D canonical point maps

---

Input: 2D pixel position $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{H \times W}$, 3D position $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \in \mathbb{R}^{H \times W}$

Output: f_x, f_y

1: $\mathbf{Q}_x = \frac{\mathbf{U} \odot \mathbf{Z}}{\mathbf{X}}$ $\triangleright \odot$ is the Hadamard Product

2: $\mathbf{Q}_y = \frac{\mathbf{V} \odot \mathbf{Z}}{\mathbf{Y}}$

3: $f_x = median(\mathbf{Q}_x)$ $\triangleright$ Across spatial dimension

4: $f_y = median(\mathbf{Q}_y)$

Overlap mask. We normalize MEt3R with an overlap mask $\mathbf{M}$ as formulated in Eq. 4 which is a crucial step. During rasterization, we set the background values to a large negative value η for each channel and subsequently build the mask using the background values for each projected view, i.e,

$$m_{ij}^k = \begin{cases} 0 & \textbf{if } p_{ij}^k = \eta \\ 1 & \textbf{otherwise} \end{cases} \quad (8)$$

where $m_{ij}^k = [\mathbf{M}^k]_{ij}$ is the mask for k^{th} view, $p_{ij}^k = [\mathbf{P}^k]_{ij}$ are the pixel values after projection and rasterization. We set $\eta = -10000$, and we perform pixel-wise multiplication of both masks $\mathbf{M}^i$ and $\mathbf{M}^{i+1}$ to get the overlap mask $\mathbf{M}$:

$$\mathbf{M} = \mathbf{M}^i \odot \mathbf{M}^{i+1} \quad (9)$$

Figure. 14 shows MEt3R without normalizing against the overlap mask $\mathbf{M}$ in Eq. 4. Instead, we take the average of the similarity scores for all pixels. Compared to MEt3R (c.f. Fig. 4), the lower bound gets significantly larger with

Figure 13. **Fixed vs. adjusted projection matrix.** With fixed focal length, the projection area varies across different scales of DUSt3R point maps. We automatically adjust the focal length for each example pair to allow maximal projection and, therefore, more pixels for evaluating feature similarity.

a large offset, while DFM [36] gets worse than all other baselines. Meanwhile, PhotoNVS [46] gets almost similar to GenWarp [31]. This contradicts both the theoretical expectation and the visual judgment about the 3D consistency of the baselines. In addition, the standard deviations for all baselines are large and correspond to noisy scores for individual image pairs across the test sequences. However, some key features, such as spikes from anchor-to-anchor transitions in MV-LDM and the gradual increase in MEt3R due to decreasing 3D consistency, are still visible.
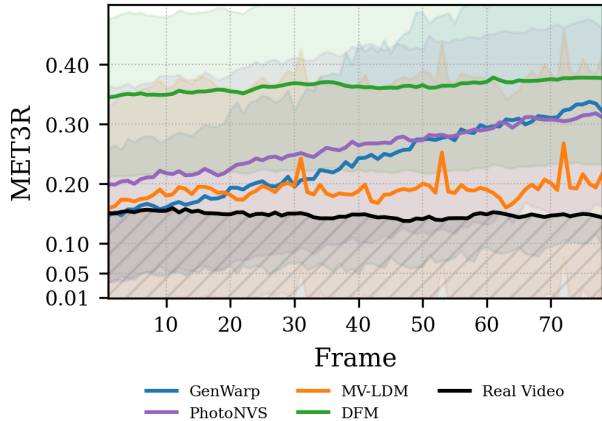
13

Figure 14. **MEt3R without overlap mask.** Per-image-pair MEt3R without normalizing against the overlap mask. Under this setting, DFM [36] is worse than all other baselines in 3D consistency, even though it has a strong inductive bias, which forces its results to be 3D consistent at the expense of blur. Whereas PhotoNVS [46] and GenWarp [31] are similar, both of which increase gradually, whereas MV-LDM stays relatively low with visible spikes due to anchor-to-anchor transitions.

C. Additional Details on Multi-View Generation Models

In the following, we present additional details on the multi-view generation baselines.

GenWarp. GenWarp [31] employs a two-step approach, i.e., project and in-paint. With a monocular depth estimator, it predicts depth maps for the input image and un-projects the RGB in 3D space. The 3D points are rendered onto a target view, followed by inpainting with an image-to-image diffusion model. GenWarp generates only one view at a time. For every novel view, we condition the model on the fixed input view for every novel view, as an autoregressive approach diverges very quickly due to error accumulation.

PhotoNVS. Just like GenWarp, PhotoNVS [46] also generates a single view at a time given a conditioning image. However, by employing a score-based diffusion UNet architecture for both views with cross-view attention in-between, it can always condition on the last generated frame in an autoregressive fashion, improving multi-view consistency across a full sequence.

DFM. DFM [36] incorporates a neural radiance field into the architecture of an image diffusion model such that novel views are 3D consistent by design. By employing pixel-NeRF [45], DFM generates the 3D representation given a set of conditioning views. Starting from a single view, it

	GenWarp	PhotoNVS	DFM	MV-LDM
Runtime (s)	**70**	7840	1020	<u>100</u>

Table 3. **Runtime comparison**. We report the runtime in seconds for all the baselines for generating a full video sequence comprising 80 frames. MV-LDM and GenWarp achieve the fastest sampling, followed by DFM and then PhotoNVS.

generates an extrapolated target view that acts as additional conditioning in all subsequent sampling steps.

D. Runtime

In Tab. 3, we compare the runtimes of the evaluated methods for generating 80 frames of a video sequence on an NVIDIA RTX4090 GPU with 24GB VRAM. GenWarp achieves the fastest sampling time, as high-quality but inconsistent novel views can already be obtained with 20 DDIM steps. Although MV-LDM generates multiple views at a time, which improves 3D consistency and uses 70 DDIM steps to achieve good image quality, it is only slightly slower than the single-view generation of GenWarp. Both DFM and PhotoNVS are an order of magnitude slower due to slow volumetric NeRF rendering and many denoising steps, respectively. Our proposed metric MEt3R can be evaluated in only $95ms$ per image pair.
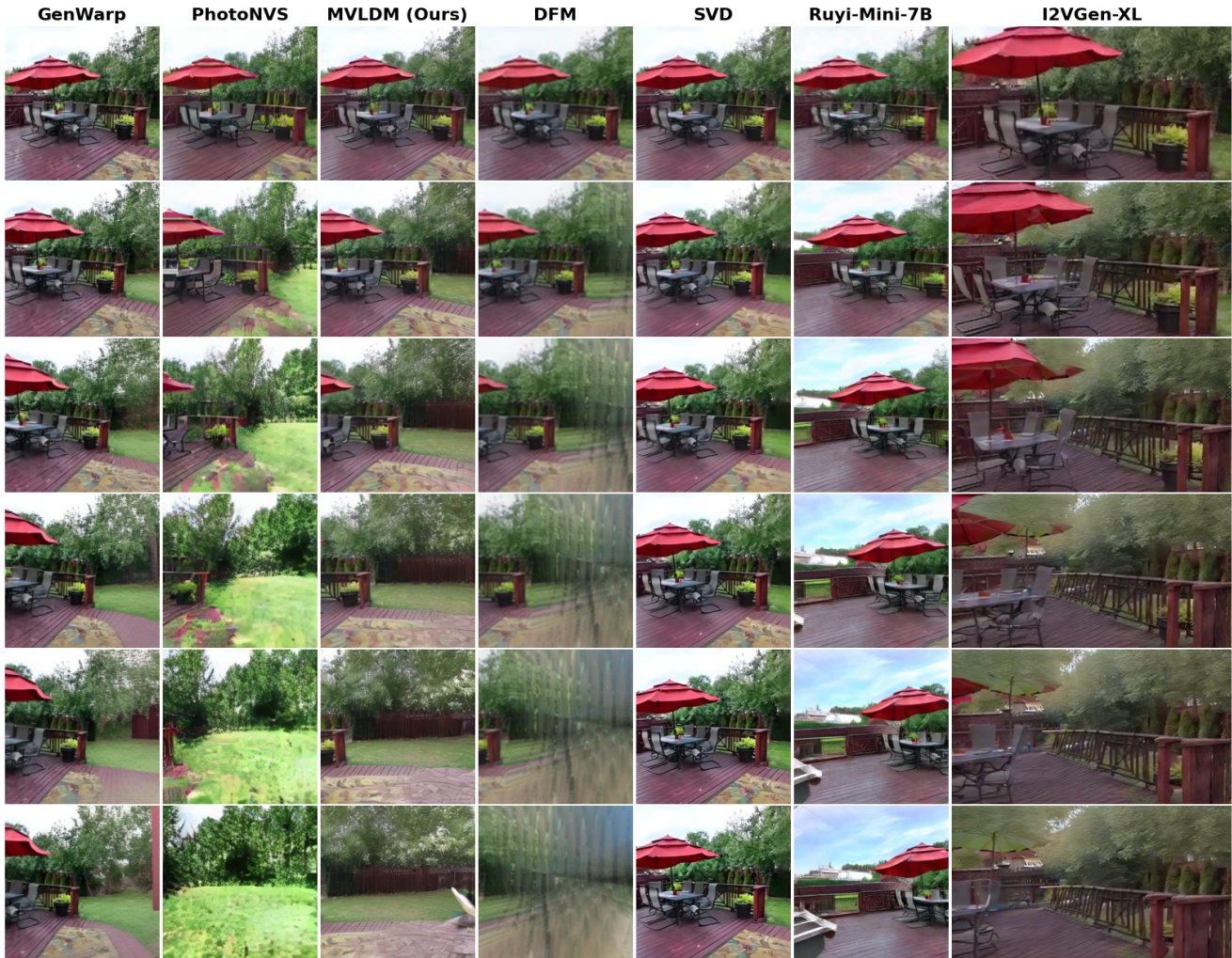
Figure 15. **Examples of generated multi-views and videos.** From Top → Down is the increasing frame number with columns for each method. Note that the first row is the input image, the first four columns are the results of multi-view generation models with explicit camera control, whereas the last three columns are generated videos from video diffusion models without any camera control.
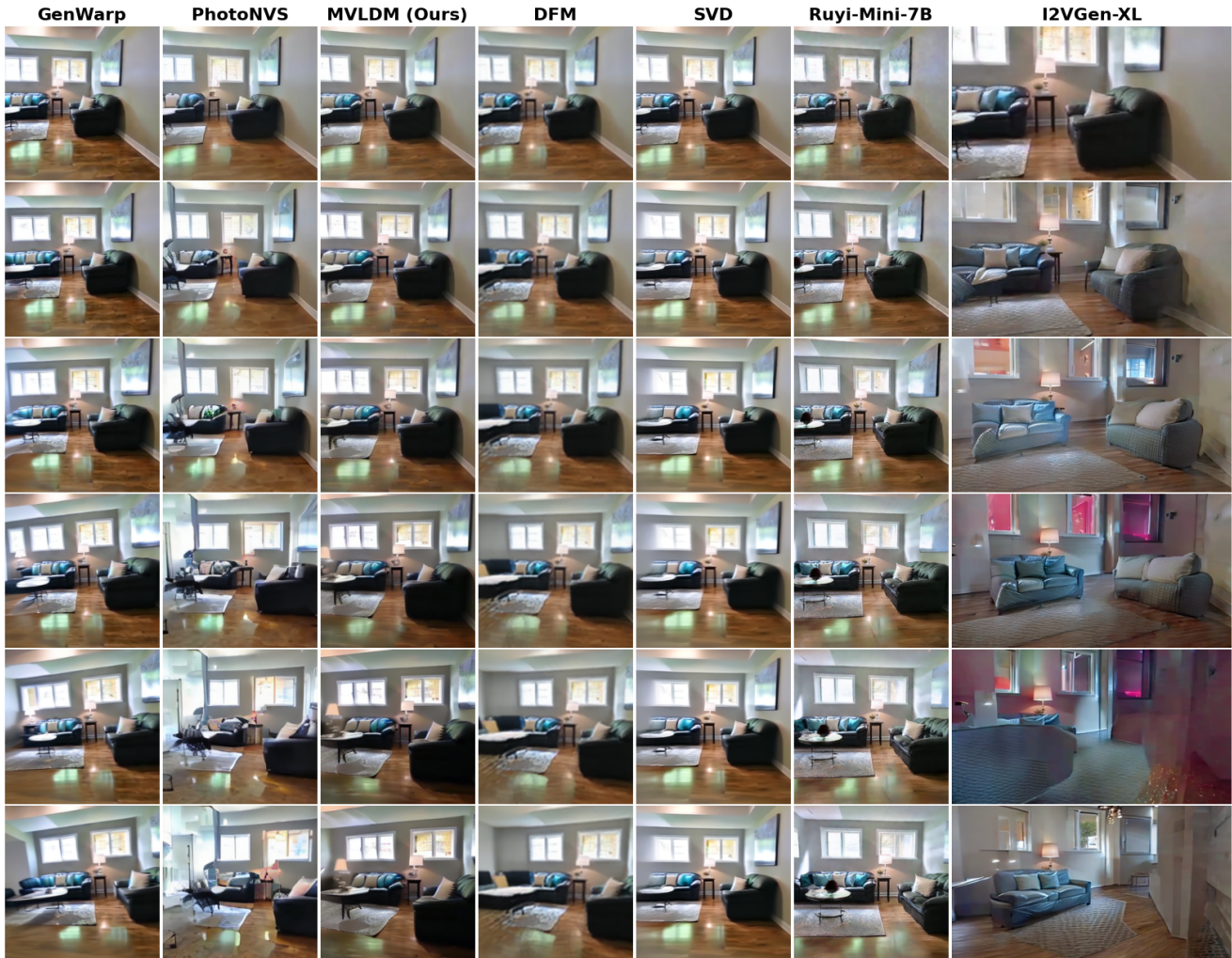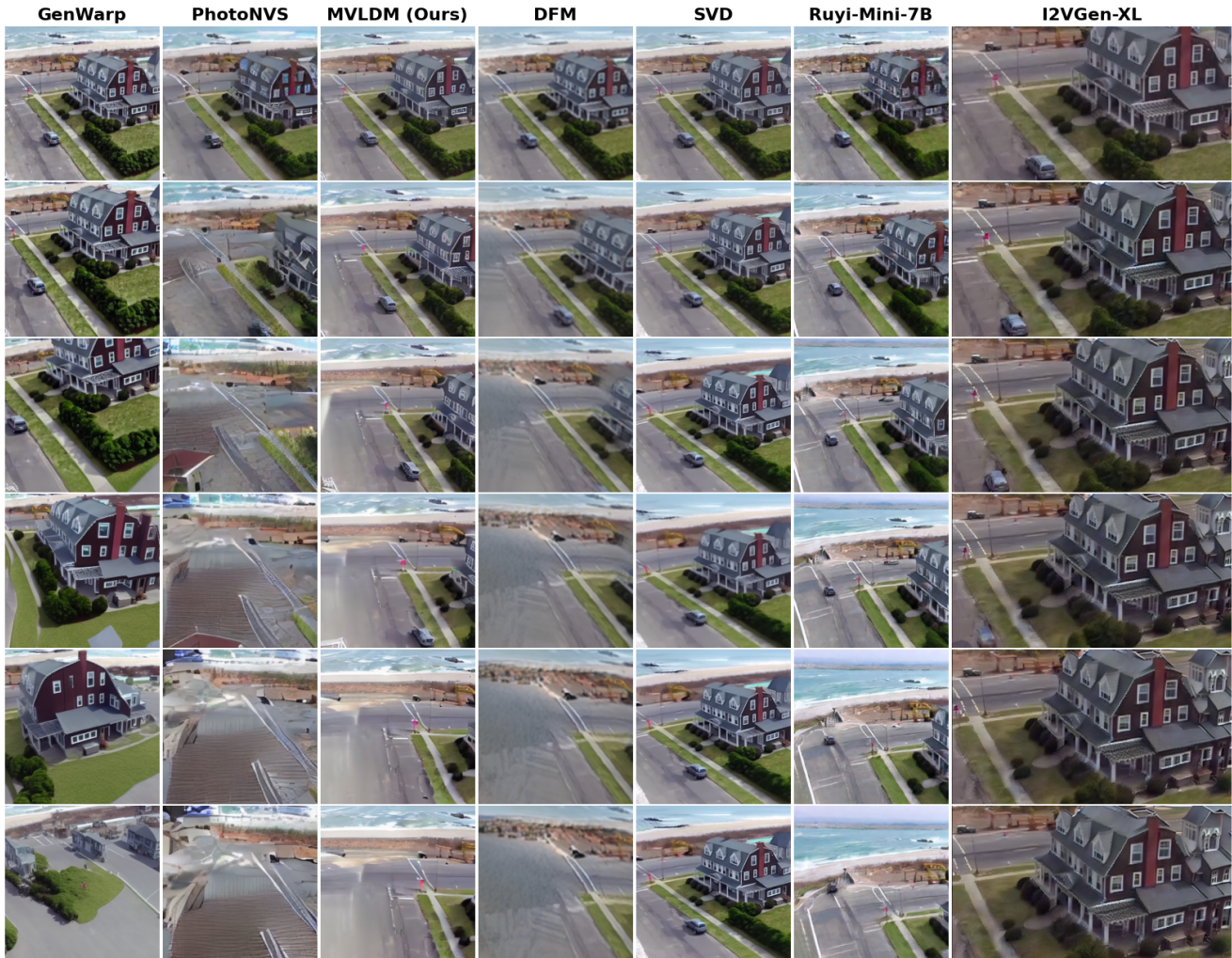
Figure 16. **Examples of generated multi-views and videos.** From Top → Down is the increasing frame number with columns for each method. Note that the first row is the input image, the first four columns are the results of multi-view generation models with explicit camera control, whereas the last three columns are generated videos from video diffusion models without any camera control.

Figure 17. **Examples of generated multi-views and videos.** From Top → Down is the increasing frame number with columns for each method. Note that the first row is the input image, the first four columns are the results of multi-view generation models with explicit camera control, whereas the last three columns are generated videos from video diffusion models without any camera control.
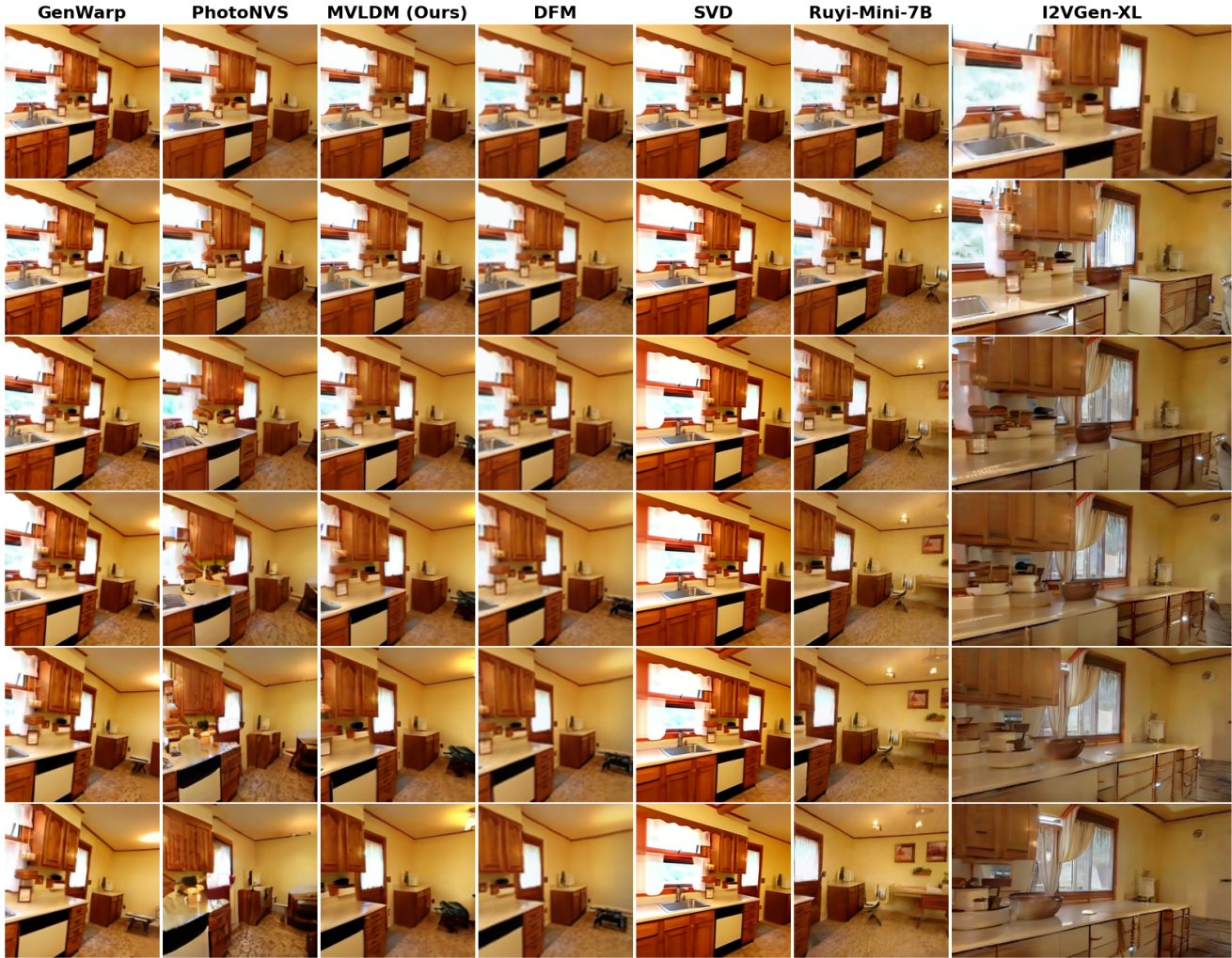
Figure 18. **Examples of generated multi-views and videos.** From Top → Down is the increasing frame number with columns for each method. Note that the first row is the input image, the first four columns are the results of multi-view generation models with explicit camera control, whereas the last three columns are generated videos from video diffusion models without any camera control.
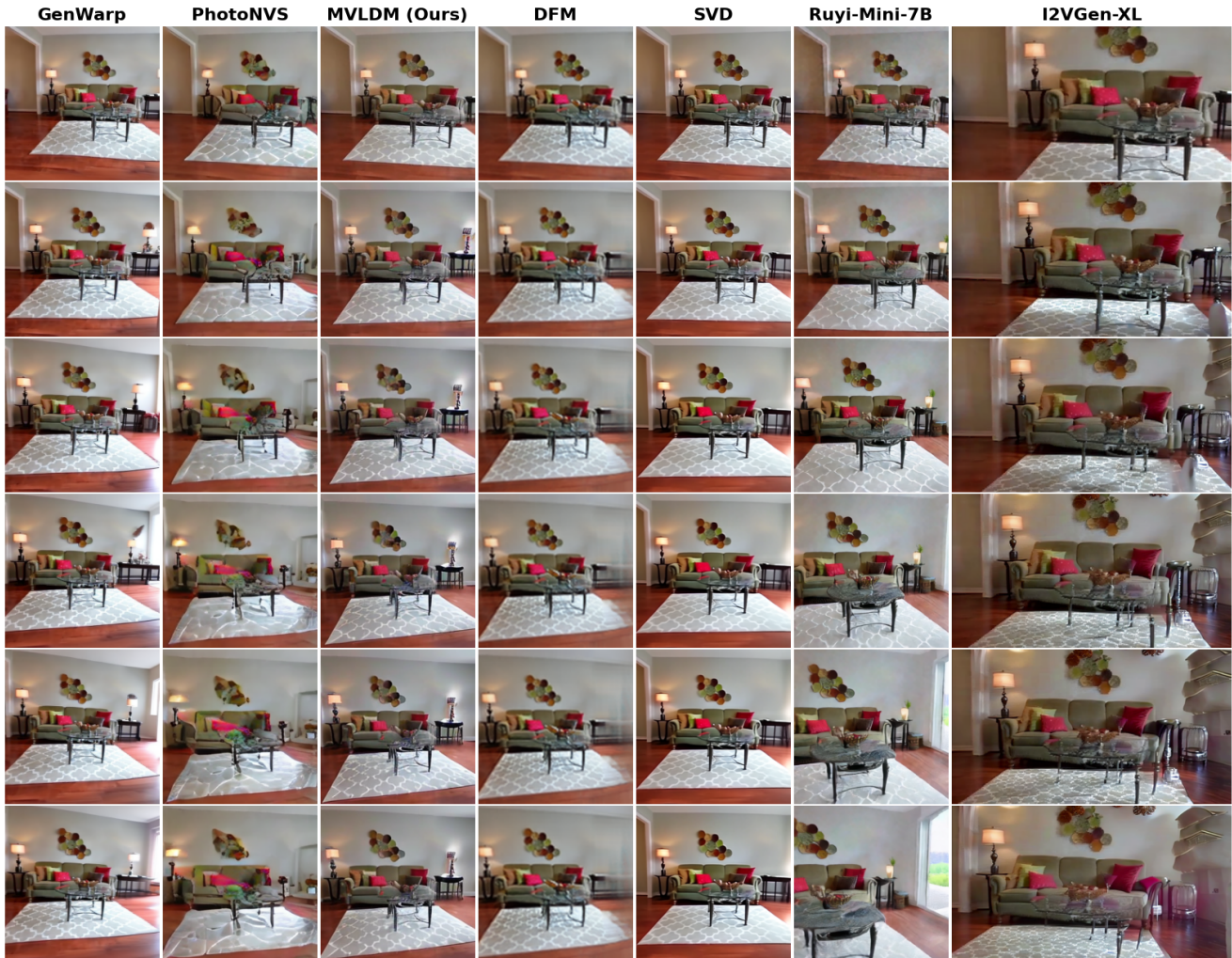
Figure 19. **Examples of generated multi-views and videos.** From Top → Down is the increasing frame number with columns for each method. Note that the first row is the input image, the first four columns are the results of multi-view generation models with explicit camera control, whereas the last three columns are generated videos from video diffusion models without any camera control.