Camera Pose Estimation Emerging in Video Diffusion Transformer

Anonymous authors
Paper under double-blind review

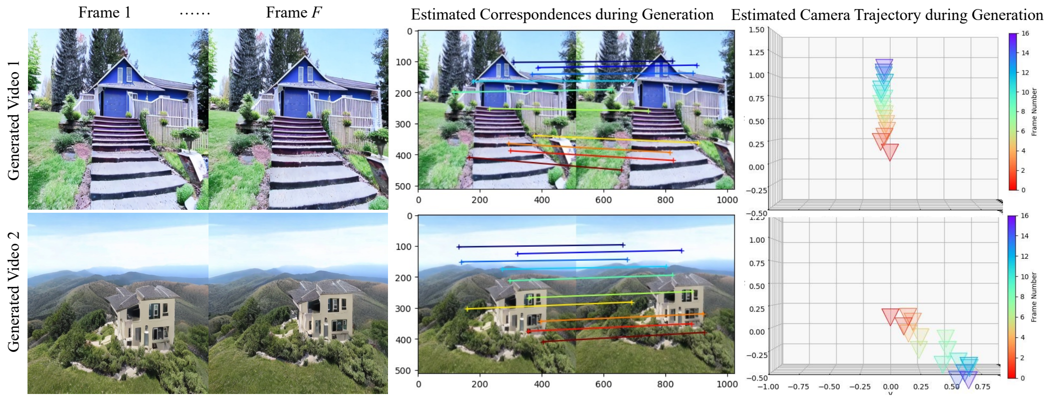


Figure 1: **JOG3R** creates realistic videos of stationary scenes while *simultaneously* generating the associated camera pose for each frame. Please refer to the supplementary page for video results.

Abstract

Diffusion-based video generators are now a reality. Being trained on a large corpus of real videos, such models can generate diverse yet realistic videos (Brooks et al., 2024; Zheng et al., 2024). Given that the videos appear visually coherent across camera changes, we ask, *do the underlying generators implicitly learn camera registrations?* Hence, we propose a novel adaptation to repurpose the intermediate features of the generator for camera pose estimation by linking them to the SoTA camera calibration decoder of DUSt3R (Wang et al., 2024a). This effectively unifies the video generation and camera estimation into a single framework. On top of unifying two different networks into one, our architecture can directly be trained on real video and simultaneously produces correspondence, with respect to the first frame, for all the video frames. Our final model, named JOG3R can be used in text-to-video mode, and additionally it produces camera pose estimates at a quality on par with the SoTA model DUSt3R, which was trained exclusively for camera pose estimation. We report that the synergy between video generation and 3D camera reconstruction tasks leads to around 25% better FVD scores with JOG3R against pretrained OpenSora.

1 Introduction

Video diffusion models have rapidly improved over the last two years, leading to the emergence of many commercial and open-sourced models (Guo et al., 2024; Zheng et al., 2024; Brooks et al., 2024; Menapace et al., 2024a; Blattmann et al., 2023a). They are trained on very large-scale datasets, *e.g.*, WebVid10M (Bain et al., 2021) or Panda-70M (Chen et al., 2024), and produce realistic, diverse, and temporally smooth videos, simply based on text or image prompts.

In another recent breakthrough, DUSt3R (Wang et al., 2024a) demonstrated that the long-standing optimization-based structure-from-motion framework for camera estimation can be directly replaced by the forward pass of a dedicated network that has been trained to establish correspondence between any given pair of video frames. This is in contrast to the current SoTA in optimization-based approach for structure-from-motion GLOMAP (Pan et al., 2024a).

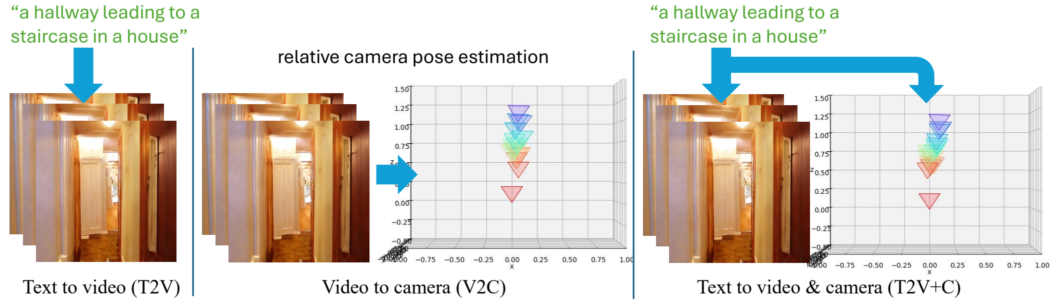

"a hallway leading to a staircase in a house"
relative camera pose estimation
"a hallway leading to a staircase in a house"
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
z
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
x
Text to video (T2V)
Video to camera (V2C)
Text to video & camera (T2V+C)

Figure 2: JOG3R is a versatile model that can (a) generate a video from text, (b) reconstruct 3D camera motion given a video, and (c) generate a video and the corresponding camera motion simultaneously. The camera trajectories obtained in (b) and (c) are consistent.

Inspired by the emergent behavior of intermediate features of large-scale image generators towards other tasks (*e.g*., correspondence, semantic segmentation, etc. Tang et al. (2023); Dutt et al. (2024)), we ask if the pretrained video generator features have similar emergent behavior. In particular, we investigate whether the pretrained features can be repurposed towards DUSt3R-like camera pose estimation. Surprisingly, we find that the video generator features, OpenSora in our setting, do not natively have emergency behavior and cannot be used directly for camera tracking.

Instead, we investigate whether the video generator features can be adapted towards camera pose estimation. In particular, we test if with a limited amount of fine-tuning, one can produce video generator features that also can be reused for camera tracking, without sacrificing video generation quality (see Figure 1). We present a *JOint Generation and 3d camera Reconstruction* network, in short JOG3R, that combines video generation with camera pose estimation into a single network, and can be supervised with generation and 3D reconstruction losses. We demonstrate that this does not lead to a loss in video quality while setting a new SoTA with respect to camera tracking on real video using a feedforward network (see Figure 2). In fact, we find that training with camera reconstruction leads to improved video generation, leading to a notable improved FVD score on the RealEstate10K-test.

In summary, the paper makes the following contributions:

- The first model that can both generate videos and estimate 3D cameras;
- Extensive experiment and study on how well the video features can be used for 3D camera estimation and ablating the various design choices; and
- Reporting SoTA video-based camera tracking results on both RealEstate10k-test and DL3DV10K datasets.

2 Related Work

2.1 Diffusion-Based Video Generation

Building on the success of diffusion models (Ho et al., 2020; Song et al., 2020) in image synthesis (Dhariwal & Nichol, 2021; Rombach et al., 2021), the research community has extended diffusion-based methods to video generation. Early works (Ho et al., 2022a;b) adapted image diffusion architectures by incorporating a temporal dimension, enabling the model to be trained on both image and video data. Typically, U-Net-based architectures incorporate temporal attention blocks after spatial attention blocks and 2D convolution layers are expanded to 3D convolution layers by altering kernels (Ho et al., 2022b; Wu et al., 2023). Latent video diffusion models (Blattmann et al., 2023b; He et al., 2022; Wang et al., 2023b; Blattmann et al., 2023a) have been introduced to avoid excessive computing demands, implementing the diffusion process in a lower-dimensional latent space. Seeking to generate spatially and temporally high-resolution videos, another line of research adopts cascaded pipelines (Ho et al., 2022a; Singer et al., 2022; Zhang et al., 2023a; Wang et al., 2023c;

Bar-Tal et al., 2024), incorporating low-resolution keyframe generation, frame interpolation, and super-resolution modules. To maximize computational scalability, recent waves in video generation (Chen et al., 2023; Ma et al., 2024; Menapace et al., 2024b; Brooks et al., 2024; Zheng et al., 2024) diverge from U-Net-based architecture and employ Diffusion Transformer (DiT) (Peebles & Xie, 2023) backbone that processes space-time patches of video and image latent codes. Following this direction, we build our method on OpenSora (Zheng et al., 2024), a publicly available DiT-based latent video diffusion model.

2.2 3D Reconstruction

The fundamental principles of multiview geometry Wrobel (2001) including feature extraction Lowe (2004); Brown et al. (2011), matching Agarwal et al. (2009); Lou et al. (2012); Wu (2013a); Havlena & Schindler (2014), and triangulation with epipolar constraints are well known to produce highly accurate (yet spare) 3D point clouds with precise camera pose estimation from multiview images Schonberger & Frahm (2016). The efficiency of 3D reconstruction has been improved with linear-time incremental structure-from-motion Wu (2013b) and coarse-to-fine hybrid approaches Crandall et al. (2012); Cui et al. (2017). To improve robustness to outliers, researchers proposed global camera rotation averaging Cui et al. (2017), camera optimization techniques based on features of points vanishing with oriented planes Holynski et al. (2020) or from a learned neural network Lindenberger et al. (2021) to prevent rotation and scale drift issues in the process of the structure-from-motion. Global camera pose registration and approximation with geometric linearity Jiang et al. (2013); Cai et al. (2021) or joint 3D point position estimation Pan et al. (2024a) are designed to further push the scalability and efficiency of the 3D reconstruction as well as the robustness particularly to the image sequence with small baselines.

Given estimated camera poses and sparse 3D point clouds, multiview stereo can then produce a dense 3D surface using hand-created visual features Schönberger et al. (2016) or neural features with a cost volume Ma et al. (2022); Ummenhofer & Koltun (2021); Ma et al. (2022); Zhang et al. (2023c); Ye et al. (2023) to predict globally coherent depth estimates. Existing neural rendering methods reconstruct such a dense surface by modeling the implicit or explicit cost volume and differentiable rendering of the scene for photometric supervision from multiview images Li et al. (2023b); Sun et al. (2022); Peng et al. (2023); Guo et al. (2022); Yu et al. (2022); Wang et al. (2022); Oechsle et al. (2021); Wang et al. (2021); Murez et al. (2020) or monocular depth estimation Sayed et al. (2022). Some pose-free methods further erase the requirement of camera calibration: test time optimization produces globally consistent depth map under unknown scale and poses using frozen depth prediction model Xu et al. (2023); the unsupervised signals from dense correspondences such as optical flow is integrated to learn from unlabeled data Yin & Shi (2018); Teed & Deng (2018); Zhou et al. (2019). Recent works proposed a direct regression framework for dense surface reconstruction from pairwise images by learning to predict globally coherent depths and camera parameters Ummenhofer et al. (2016) or to directly predict per-pixel 3D point clouds from two views Wang et al. (2024b); Leroy et al. (2024) using a vision transformer with dense tokenization techniques Ranftl et al. (2021).

2.3 Diffusion Model as Features for 3D Reconstruction

A generative diffusion model is often trained on millions of paired image and text prompts and in the process develops a semantically meaningful visual prior. Naturally, researchers are interested in using this strong prior for many downstream 3D vision tasks. Injecting 3D awareness into the diffusion prior greatly improves the accuracy and generalizability of the monocular depth estimation and correspondence search tasks El Banani et al. (2024); Yue et al. (2024). The latent features from the frozen pretrained diffusion model are often used as a backbone, and a task-specific decoder with cross attention is newly trained for semantic correspondences Tang et al. (2023); Zhang et al. (2023b); Hedlin et al. (2024); Zhang et al. (2024); Hedlin et al. (2024); Jiang et al. (2024), 3D correspondences Dutt et al. (2024), semantic segmentation and monocular depth estimation Zhao et al. (2023), material and shadow prediction Zhan et al. (2023), general object 3D pose estimation Örnek et al. (2023); Cai et al. (2024). However, such image diffusion features do not inherently consider the temporal relation between the frames, leading to temporally unstable 3D prediction results from videos. In contrast, we propose to utilize the video diffusion features as a backbone for the multitasking prediction of video generation and 3D camera poses estimation.

3 Method

3.1 Model and preliminaries.

Video diffusion model. We consider OpenSora (Zheng et al., 2024) as our base video generation model, which is a DiT-based video generator inspired by the impressive success of Sora (Brooks et al., 2024). OpenSora performs the diffusion process in a lower-dimensional latent space defined by a pre-trained VAE encoder $\mathcal{E}$. Each frame x of the input video is first projected into this latent space, $z_0 = \mathcal{E}(x)$. Given a diffusion time step t, the *forward* process incrementally adds Gaussian noise to the latent code z_0 via a Markov chain and obtains noisy latent z_t. The denoising model ϵ_θ takes the noisy latents of all frames, the time step t, and the text prompt y as input to predict the added noise: $\epsilon_\theta(\{z_t^f\}_{f=1}^F, t, y)$, where F is the total number of frames and θ denotes the parameter of the DiT network (Peebles & Xie, 2023). The network consists of $m+1$ spatial-temporal diffusion transformer (STDiT) blocks $\{b^0, \ldots, b^m\}$, similar to Ma et al. (2024). The iterative process of noise prediction and noise removal is referred to as the *backward* process.

Camera pose estimation module. We employ the state-of-the-art multi-view stereo reconstruction (MVS) framework DUSt3R (Wang et al., 2024a) as our downstream camera tracking branch. Given a pair of images, DUSt3R first encodes each one individually with a ViT encoder (Dosovitskiy et al., 2021; Weinzaepfel et al., 2022). A pair of decoders take both features as input for cross-view information sharing, followed by two separate heads estimating point maps $X \in \mathbb{R}^{H\times W\times 3}$ represented in the coordinate of the first view, denoted as $X^{1,1}$ and $X^{2,1}$, respectively. The relative camera pose is then estimated by aligning $X^{1,1}$ and $X^{1,2}$ (or, equivalently $X^{2,1}$ and $X^{2,2}$) using Procrustes alignment (Luo & Hancock, 1999) with PnP-RANSAC (Lepetit et al., 2009; Fischler & Bolles, 1981).

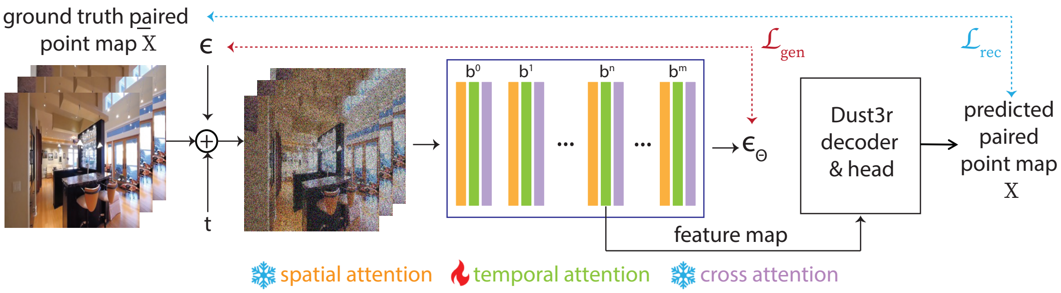


Figure 3: JOG3R repurposes the intermediate features from a video generation model for camera pose estimation by routing them to the SoTA camera calibration decoder of DUSt3R. We train both the temporal layers of the generation model as well as the DUSt3R decoders using a combination of generation and reconstruction losses.

3.2 Joint generation and reconstruction DiT network

We propose a unified network that is able to do both video denoising and camera tracking. We observe that ViT and DiT actually share many architectural designs in common since they both belong to the broad transformer family. Hence, our key insight is to replace the *image*-based ViT encoder in DUSt3R with the *video* DiT backbone in OpenSora. In other words, we provide the features of the denoising DiT network ϵ_θ to DUSt3R decoders and heads, see Figure 3 for illustration.

Specifically, we extract the output of the intermediate STDiT block b^n at a particular time step t during the backward process. Following Tang et al. (2023), we consider small t where the feature focuses more on low-level details, making it useful as a geometric feature descriptor to build correspondence across frames.

Our modification of DUSt3R. The features extracted from the video generator encode a sequence of F frames and are provided to the DUSt3R decoders. During training, we sample a pair of frames $\{(1, f)\}_{f=2}^F$ to predict the 3D point maps between the first frame and any other frame f in the sequence. At inference time, we first predict the point maps between all pairs $(1, f)$ and perform a global camera registration to obtain the camera pose estimation for the whole sequence.

Training objectives. During training, our model is supervised by two objectives: generation loss $\mathcal{L}_{\text{gen}}$ and reconstruction loss $\mathcal{L}_{\text{rec}}$. The generation loss $\mathcal{L}_{\text{gen}}$ is the common objective in training diffusion models that aims to match the added noise ϵ. The reconstruction loss $\mathcal{L}_{\text{rec}}$, following the definition in DUSt3R, is the sum of confidence-weighted Euclidean distance $L_2(f, i)$ between the regressed point maps X and the ground truth point maps $\bar{X}$ over all valid pixels i and all frames f. Formally,

$$\mathcal{L}_{\text{gen}} = \left\| \epsilon - \epsilon_\theta \left(\{z_t^f\}_{f=1}^F, t, y \right) \right\|_2^2 \tag{1}$$

$$\mathcal{L}_{\text{rec}} = \sum_{f \in \{2,..,F\}} \sum_i C_i^{f,1} L_2(f, i) - \alpha \log C_i^{f,1} \tag{2}$$

$$L_2(f, i) = \left\| \frac{1}{s} X_i^{f,1} - \frac{1}{\bar{s}} \bar{X}_i^{f,1} \right\|_2$$

where the scaling factors s and $\bar{s}$ handles the scale ambiguity between prediction and ground-truth by bringing them to a normalized scale, $C_i^{f,1}$ is the confidence score for pixel i which encourages network to extrapolate in harder areas, and α is a hyper-parameter controlling the regularization term (Wan et al., 2018). We refer interested readers to Wang et al. (2024a) for more details. The final loss is defined as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{gen}} + \lambda \mathcal{L}_{\text{rec}}$, and we empirically set $\lambda = 1$.

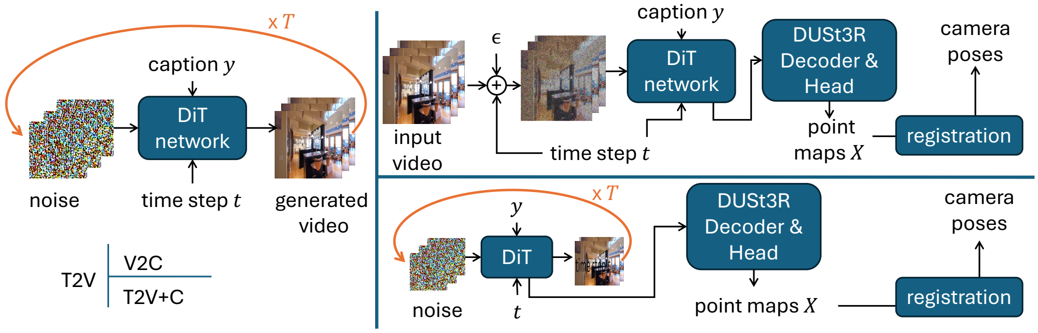


Figure 4: JOG3R supports text-to-video (T2V), video to camera estimation (V2C), and joint video generation and camera estimation (T2V+C) at inference time.

Inference. Once trained, JOG3R naturally supports three ways of inference (see Figure 2 and supplemental video): (i) *Text-to-video (T2V)*: the input is sampled Gaussian noise and we iteratively denoise it with the text guidance to generate a video. (ii) *Video-to-camera (V2C)*: we add noise to the input video based on a sampled time step t, denoise it for one time step, route the feature maps to DUSt3R decoders and heads, followed by registration of point maps X to obtain camera poses.

Given the two inference modes above, a straightforward combination is using the generated video of T2V as the input of V2C, which we denote as T2V→V2C, essentially chaining the two networks. However, thanks to our novel network design, we can provide the feature map directly to the reconstruction module at the desired time step, without the overhead of adding noise and passing it through the network again. As a result, cameras are generated alongside the video *in one go*. We term such a tightly coupled joint inference mode as (iii) *Text-to-Video+Camera (T2V+C)*. Fig. 4 illustrates the pipeline of these three inference modes.

Implementation Details. We adopt OpenSora 1.0 as our video generator, which uses 2D VAE (from Stability-AI) Rombach et al. (2022), T5 text encoder (Raffel et al., 2020), and an STDiT (ST stands for spatial-temporal) architecture similar to variant 3 in Ma et al. (2024) as the denoising network. Among the 28 STDiT blocks, we empirically set the first 4 frozen and update only the weights of the temporal attention layers for the remaining 24 blocks. We extract the output of the 26th block b^{25}

as feature maps for DUSt3R decoders. The final two blocks behave as a "generation" branch whose weights are only updated by the gradient of generation loss $\mathcal{L}_{\text{gen}}$.

We adopt the linear prediction head of DUSt3R for final pointmap estimation. DUSt3R originally uses a decoder with 12 transformer blocks that is duplicated for each of the pair of frames. However, information sharing is enabled between the two decoders. In our experiments, we find that a decoder structure with six transformer blocks provides similar performance and report our results accordingly. Furthermore, since the features we get from the generator encode all the frames in a video sequence, we also experiment with replacing the duplicate decoder architecture with a single decoder consisting of 6 transformer blocks that perform full 3D attention across all the frames. We empirically find that this performs on par with duplicate decoders (see Table 1), and hence we use the latter to provide a more fair comparison to DUSt3R.

During training, we sample the time step $t \in [0, 10]$ (corresponding to 10% of noise level) and consider the empty prompt for computing the reconstruction loss $\mathcal{L}_{\text{rec}}$, while for the generation loss $\mathcal{L}_{\text{gen}}$ we sample the full range of time steps and use the captions of the videos. At test time, we sample $t \in [0, 5]$ to add noise to the input video for camera estimation (V2C). To perform joint camera estimation and video generation (T2V+C), we run the standard T2V pipeline of OpenSora and when the time step hits the sampled $t \in [0, 5]$, we provide the output of block b^{25} to DUSt3R for camera estimation.

4 EXPERIMENTS

In this section, we evaluate the proposed method in three aspects. We follow standard approaches to assess the generated video quality (T2V). Since there is no ground truth camera trajectories for the videos generated from T2V+C, we focus on validating the accuracy of camera pose estimation on real videos (V2C) and report self-consistency for T2V+C.

4.1 SETUP

Data. We choose RealEstate10K Zhou et al. (2018) as the dataset, which has around 65K video clips paired with camera parameter annotations. We use the captions of RealEstate10K provided in He et al. (2024) and also follow their train/test split. As pre-processing, we pre-compute the VAE latents of the video frames and the T5 text embeddings of the captions. To obtain point map annotation $\bar{X}$, we first estimate metric depth with ZoeDepth (Bhat et al., 2023), un-project it to 3D and transform to the coordinate of the first frame using the camera parameters provided in RealEstate10K. All camera extrinsic parameters are expressed with respect to the first frame.

In addition, we consider DL3DV10K (Ling et al., 2024), which also provides camera annotations, as a failed test set. We choose a random set of 70 videos for testing and caption the first frame of each video using Li et al. (2023a). We prepare point map annotations using ZoeDepth (Bhat et al., 2023), similar to the RealEstate10K dataset.

Baselines. We compare with a pair-wise method DUSt3R Wang et al., 2024a with linear head and a video-based SfM method GLOMAP (Pan et al., 2024b). For DUSt3R we consider two variants: (i) off-the-shelf pretrained weights (DUSt3R†) and (ii) trained from scratch with the same data as ours (DUSt3R*). For GLOMAP we report the results before the global bundle adjustment part.

Metric. We validate the quality of camera tracking on real videos (V2C) by comparing the estimated camera poses $(\mathbf{R}, \mathbf{t})$ with the ground truth poses $(\bar{\mathbf{R}}, \bar{\mathbf{t}})$. For rotation, we compute the relative error angle between two rotation matrices. Since the estimated and ground truth translation can differ in scale, we follow Wang et al. (2023a) to compute the angle between the two normalized translation vectors, *i.e.*, $\arccos(\mathbf{t}^{\top}\bar{\mathbf{t}}/(\|\mathbf{t}\|\|\bar{\mathbf{t}}\|))$. Besides reporting the average of the two errors, we also follow Wang et al. (2024a) to report Relative Rotation Accuracy (RRA) and Relative Translation Accuracy (RTA), *i.e.*, the percentage of camera pairs with rotation/translation error below a threshold. Due to limit of the number of frames, hence small rotation variations, we select a threshold 5° to report RTA@5 and RRA@5. Additionally, we calculate the mean Average Accuracy (mAA@30), defined as the area under the curve accuracy of the angular differences at min(RRA@30, RTA@30). We also use FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2019) to measure image and video quality respectively, ensuring that our method maintains high generation quality.

Method	Rot. err. (°) ↓	Transl. err. (°) ↓	RRA@5 ↑	RTA@5 ↑	mAA@30 ↑
(0) ours w/ 3D attn	0.38	36.86	99.49%	9.17%	32.15%
(1a) ours w/o $\mathcal{L}_{\text{gen}}$	0.36	33.09	99.71%	12.18%	34.55%
(1b) JOG3R (ours)	0.37	32.66	99.77%	13.16%	35.62%
(2a) DUSt3R†	0.77	36.61	97.56%	7.54%	30.13%
(2b) DUSt3R*	0.33	30.51	99.71%	12.76%	37.88%
(3) GLOMAP	0.96	19.55	96.86%	25.92%	55.82%

Table 1: **V2C error comparison on RealEstate10K-test.** DUSt3R† indicates pretrained DUSt3R weights, whereas DUSt3R* is trained with the same training set as our method – RealEstate10K-train.

Method	Rot. err. (°) ↓	Transl. err. (°) ↓	RRA@5 ↑	RTA@5 ↑	mAA@30 ↑
(1a) ours w/o $\mathcal{L}_{\text{gen}}$	8.77	59.04	48.31%	0.37%	3.54%
(1b) JOG3R (ours)	9.01	58.82	47.73%	0.24%	3.86%
(2a) DUSt3R†	10.27	61.91	46.82%	0.33%	2.91%
(2b) DUSt3R*	8.38	58.70	49.78%	0.33%	3.93%
(3) GLOMAP	10.57	62.97	46.62%	0.21%	2.60%

Table 2: **V2C error comparison on DL3DV10K.**

4.2 Reconstruction Evaluation

In Table 1, we compare the camera pose estimation (V2C) errors on RealEstate10K-test and report the errors of withheld DL3DV10K in Table 2. Comparing (1a) and (1b) of two tables, we see that removing generation loss $\mathcal{L}_{\text{gen}}$ leads to overall worse results than our full model, confirming the hypothesis that *retaining generation ability helps reconstruction*.

Our full method – JOG3R, performs overall better than pretrained DUSt3R on both datasets, cf., (1b) and (2a). When trained with the same RealEstate10K-train, JOG3R still has on-par reconstruction quality compared with the DUSt3R counterpart DUSt3R*. When we replace the original DUSt3R decoders with full 3D attention blocks (0), we obtain on-par results with a marginal drop in accuracy.

We also report the results of GLOMAP (Pan et al., 2024b) before the final bundle adjustment step. It is the state-of-the-art method in a well studied SfM problem, which can be treated as a role of the upper bound to indicate how far we are. In Table 1 row (3), we observe it does surpass other methods in RealEstate10K, where videos often contain smaller motion and hence smaller baselines for each stereo pair. When the overlap between consecutive frames gets smaller, like in DL3DV10K, such a video-based method struggles and our method actually yields lower errors than GLOMAP.

Figure 5 shows the qualitative comparison of our method and baselines. Since camera poses are estimated through registration, which builds 3D correspondences along the way, we visualize the final camera trajectories as well as the correspondence between the first and the last frame. One can see that our method produces good camera trajectories similar to DUSt3R, which is a method tailored for reconstruction only, but no generation. In the last row we show a failure case where *both* our method and DUSt3R fail to estimate reasonable camera poses. We hypothesize this is due to the infinite depth in the sky region which could cause inconsistent scale normalization across each stereo pair.

4.3 Generation Evaluation

For each method, we generate 180 videos using the captions in RealEstate10K-test and report the FID/FVD against the real images/videos in RealEstate10K-test. Table 3 suggests that our full model generates more realistic images/videos than pretrained OpenSora ((1c) vs. (2)). When ablating the generation loss $\mathcal{L}_{\text{gen}}$, the quality slightly degrades compared to our full model ((1c) vs. (1a)). This is intuitive because without the generation loss, there is nothing to enforce the model to retain its

Method	FID ↓	FVD ↓
(1a) ours w/o $\mathcal{L}_{\text{gen}}$	110.40	1898.72
(1b) ours w/o $\mathcal{L}_{\text{rec}}$	88.02	1440.92
(1c) JOG3R (ours)	94.75	1339.74
(2) pretrained OpenSora	115.36	1872.41

Table 3: **Generation quality comparison.** We compute the FID and FVD with RealEstate10K-test.

full generation capability. See also supplemental videos. It is worth noting that (1b) corresponds to a baseline where $\mathcal{L}_{\text{rec}}$ is disabled by removing DUSt3R decoders/heads, *i.e.*, it is equivalent to standard video diffusion model finetuning except only the weights of the temporal attention layers are updated. We see removing $\mathcal{L}_{\text{rec}}$ leads to different impacts on FID and FVD. Since our method aims to generate videos, we argue FVD is a more important metric to measure the quality. As a result, the lower FVD of our full methdod (1c) suggests *learning camera pose estimation positively impacts the quality of video generation*. Figure 6 shows that our method generate realistic videos and the qualitative comparison also confirms the benefit of reconstruction loss $\mathcal{L}_{\text{rec}}$.

4.4 Discussion

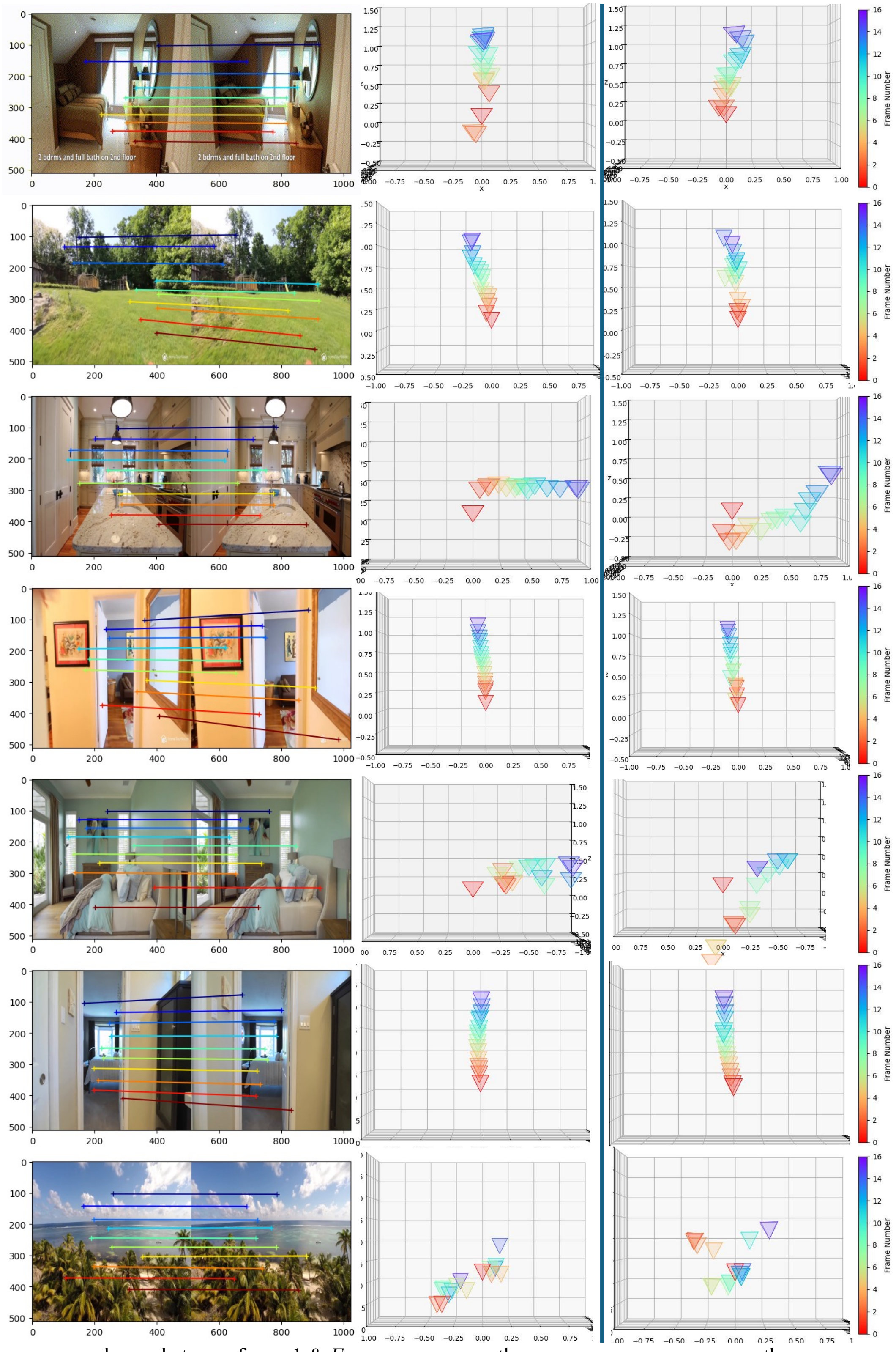
Synergy of two tasks. Our full model JOG3R is trained with two losses, generation loss $\mathcal{L}_{\text{gen}}$ and reconstruction loss $\mathcal{L}_{\text{rec}}$. In both Table 1 and 2, (1a) and (1b), we show that keeping the generation loss $\mathcal{L}_{\text{gen}}$ helps the reconstruction branch attain better camera poses estimation. On the other hand, Table 3 (1b) and (1c) also suggest that introducing the reconstruction task results in better video generation quality. Empirically, we demonstrate a synergy between two tasks – learning to generate helps reconstruction; learning to reconstruct also helps generation. It shares the same spirit with the known "analysis and synthesis" analogy, but our architectural design tightly couples them in one network and allow end-to-end training.

Self consistency of T2V→V2C and T2V+C. Since one can use JOG3R to *generate* camera trajectories in two ways: cascading T2V and V2C or the tightly coupled T2V+C pipeline, it is worth comparing how much the two results differ. We run the two pipelines with 100 prompts and report 0.45° average difference in rotation and 19.20° in translation, both of which are low errors compared with the corresponding numbers in Table 1 and 2, indicating that the camera poses from joint T2V+C pipeline is consistent with T2V→V2C. The qualitative results in Figure 7 also confirm this conclusion.

5 Conclusions and Future Work

We have presented the first framework to enable joint video generation and 3D camera reconstruction. Our method utilizes intermediate features of a pre-trained video generation model for predicting relative 3D point maps and hence enabling camera registration. Specifically, by providing the intermediate generation features to task specific decoders and prediction heads, we present a unified framework for text-to-video generation (T2V), joint generation and camera estimation (T2V+C), and camera estimation for real videos (V2C).

While being first of its kind, our method is not without limitations. First of all, since it is not trivial to obtain accurate camera annotations for dynamic scenes, our method is currently trained and applicable for videos of static scenes only. The length of the video sequences our method can handle is currently limited by the number of frames the generator can synthesize. Handling longer sequences may require extending our method to operate in a sliding window manner. As the video generators continue to improve to enable generation of longer sequences, our method will also naturally extend to handling longer videos with larger baseline.



correspondences between frame 1 & F from registration (ours)

camera paths (ours)

camera paths (pretrained Dust3r)

Figure 5: **Qualitative camera pose estimation (V2C) results.** The last row indicates a failure case. Please see suppmat. for videos and more analysis.

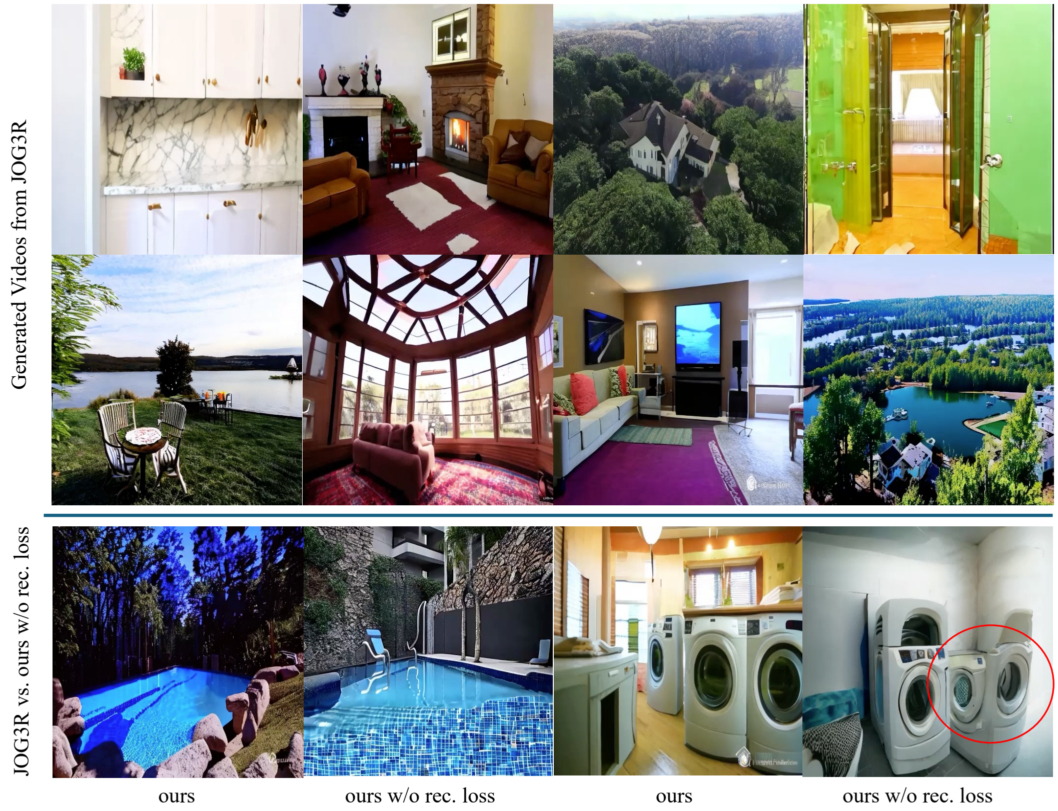


Figure 6: **Qualitative generation T2V results.** Please see suppmat. for videos.

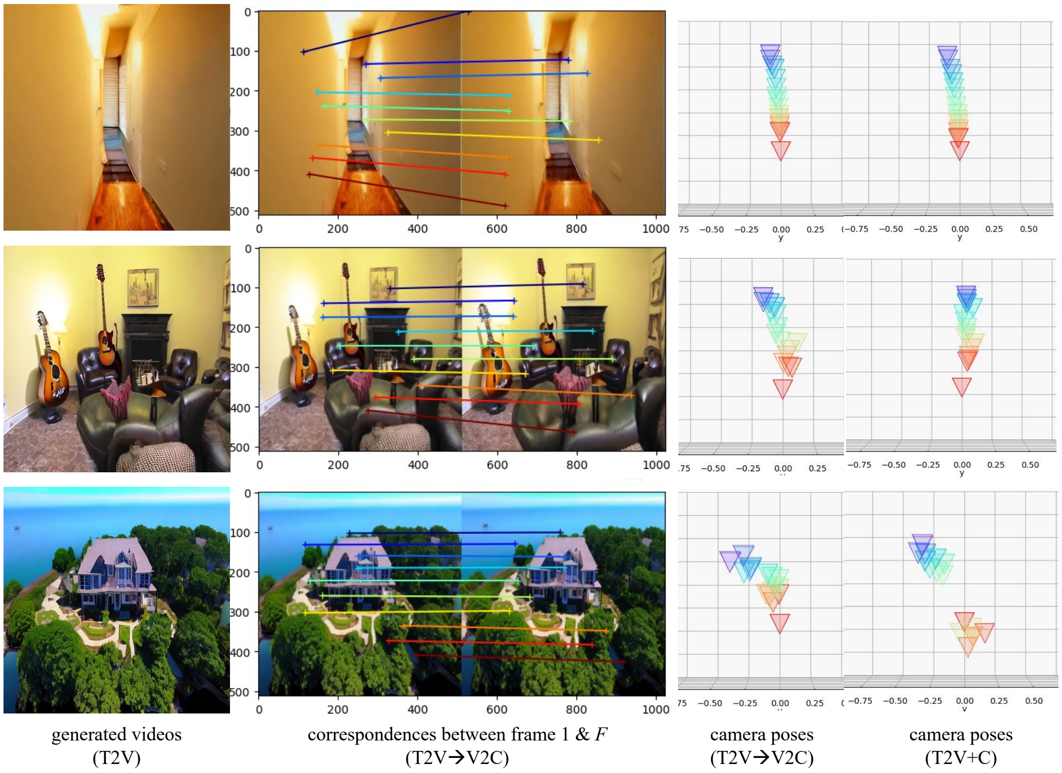


Figure 7: **Qualitative generation T2V+C results.** Please see suppmat. for videos.

REFERENCES

Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M. Seitz, and Richard Szeliski. Building rome in a day. *2009 IEEE 12th International Conference on Computer Vision*, pp. 72–79, 2009. URL `https://api.semanticscholar.org/CorpusID:7448214`.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv*, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023b.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Matthew A. Brown, Gang Hua, and Simon A. J. Winder. Discriminative learning of local image descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:43–57, 2011. URL `https://api.semanticscholar.org/CorpusID:12573831`.

Junhao Cai, Yisheng He, Weihao Yuan, Siyu Zhu, Zilong Dong, Liefeng Bo, and Qifeng Chen. Open-vocabulary category-level object pose and size estimation. *IEEE Robotics and Automation Letters*, 2024.

Qi Cai, Lilian Zhang, Yuanxin Wu, Wenxian Yu, and Dewen Hu. A pose-only solution to visual reconstruction and navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):73–86, 2021.

Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Delving deep into diffusion transformers for image and video generation. *arXiv preprint arXiv:2312.04557*, 2023.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70M: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, pp. 13320–13331, June 2024.

David J Crandall, Andrew Owens, Noah Snavely, and Daniel P Huttenlocher. Sfm with mrfs: Discrete-continuous optimization for large-scale structure from motion. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2841–2853, 2012.

Hainan Cui, Xiang Gao, Shuhan Shen, and Zhanyi Hu. Hsfm: Hybrid structure-from-motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1212–1221, 2017.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Niladri Shekhar Dutt, Sanjeev Muralikrishnan, and Niloy J Mitra. Diffusion 3d features (diff3f): Decorating untextured shapes with distilled semantic features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4494–4504, 2024.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21795–21806, June 2024.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5501–5510, 2022. URL `https://api.semanticscholar.org/CorpusID:248524713`.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024.

Michal Havlena and Konrad Schindler. Vocmatch: Efficient multiview correspondence for structure from motion. In *European Conference on Computer Vision*, 2014. URL `https://api.semanticscholar.org/CorpusID:15285158`.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 4 2024.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.

Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

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

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

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, 2022b.

Aleksander Holynski, David Geraghty, Jan-Michael Frahm, Chris Sweeney, and Richard Szeliski. Reducing drift in structure from motion using extended features. In *2020 International Conference on 3D Vision (3DV)*, pp. 51–60. IEEE, 2020.

Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, and André Araujo. Omniglue: Generalizable feature matching with foundation model guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19865–19875, 2024.

Nianjuan Jiang, Zhaopeng Cui, and Ping Tan. A global linear method for camera pose registration. In *Proceedings of the IEEE international conference on computer vision*, pp. 481–488, 2013.

Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *IJCV*, 81:155–166, 2009.

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

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 31–41, Toronto, Canada, July 2023a. Association for Computational Linguistics. URL `https://aclanthology.org/2023.acl-demo.3`.

Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8456–8465, 2023b.

Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5987–5997, 2021.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. DL3DV-10K: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, pp. 22160–22169, 2024.

Yin Lou, Noah Snavely, and Johannes Gehrke. Matchminer: Efficient spanning structure mining in large image collections. In *European Conference on Computer Vision*, 2012. URL `https://api.semanticscholar.org/CorpusID:17113492`.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004. URL `https://api.semanticscholar.org/CorpusID:174065`.

Bin Luo and Edwin R. Hancock. Procrustes alignment with the em algorithm. In Franc Solina and Alešs Leonardis (eds.), *Computer Analysis of Images and Patterns*, pp. 623–631, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg. ISBN 978-3-540-48375-5.

Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

Zeyu Ma, Zachary Teed, and Jia Deng. Multiview stereo with cascaded epipolar raft. In *European Conference on Computer Vision*, pp. 734–750. Springer, 2022.

Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, and Sergey Tulyakov. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. *arXiv preprint arXiv:2402.14797*, 2 2024a.

Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, et al. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7038–7048, 2024b.

Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *European Conference on Computer Vision*, 2020. URL `https://api.semanticscholar.org/CorpusID:214612128`.

Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5569–5579, 2021. URL `https://api.semanticscholar.org/CorpusID:233307004`.

Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomas Hodan. Foundpose: Unseen object pose estimation with foundation features. *arXiv preprint arXiv:2311.18809*, 2023.

Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global structure-from-motion revisited. *arXiv preprint arXiv:2407.20219*, 2024a.

Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes Lutz Schönberger. Global structure-from-motion revisited. In *European Conference on Computer Vision (ECCV)*, 2024b.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pp. 4195–4205, 2023.

Rui Peng, Xiaodong Gu, Luyang Tang, Shihe Shen, Fanqi Yu, and Ronggang Wang. Gens: Generalizable neural surface reconstruction from multi-view images. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168, 2021. URL `https://api.semanticscholar.org/CorpusID:232352612`.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12 2022.

Mohamed Sayed, John Gibson, Jamie Watson, Victor Adrian Prisacariu, Michael Firman, and Clément Godard. Simplerecon: 3d reconstruction without 3d convolutions. In *European Conference on Computer Vision*, 2022. URL `https://api.semanticscholar.org/CorpusID:251953231`.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pp. 501–518. Springer, 2016.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

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.

Jiaming Sun, Xi Chen, Qianqian Wang, Zhengqi Li, Hadar Averbuch-Elor, Xiaowei Zhou, and Noah Snavely. Neural 3d reconstruction in the wild. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–9, 2022.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *NeurIPS*, 2023. URL `https://openreview.net/forum?id=ypOiXjdfnU`.

Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. *ArXiv*, abs/1812.04605, 2018. URL `https://api.semanticscholar.org/CorpusID:54482591`.

Benjamin Ummenhofer and Vladlen Koltun. Adaptive surface reconstruction with multiscale convolutional kernels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5651–5660, 2021.

Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5622–5631, 2016. URL `https://api.semanticscholar.org/CorpusID:6159584`.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation. In *ICLR workshop*, 2019. URL `https://openreview.net/forum?id=rylgEULtdN`.

Sheng Wan, Tung-Yu Wu, Wing H. Wong, and Chen-Yi Lee. Confnet: Predict with confidence. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2921–2925, 2018.

Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9773–9783, 2023a.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023b.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *ArXiv*, abs/2106.10689, 2021. URL `https://api.semanticscholar.org/CorpusID:235490453`.

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

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024b.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023c.

Yiqun Wang, Ivan Skorokhodov, and Peter Wonka. Improved surface reconstruction using high-frequency details. *ArXiv*, abs/2206.07850, 2022. URL `https://api.semanticscholar.org/CorpusID:252438827`.

Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav ARORA, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jerome Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *NeurIPS*, 2022. URL `https://openreview.net/forum?id=wZEfHUM5ri`.

Bernhard P. Wrobel. Multiple view geometry in computer vision. *Künstliche Intell.*, 15:41, 2001. URL `https://api.semanticscholar.org/CorpusID:261497446`.

Changchang Wu. Towards linear-time incremental structure from motion. *2013 International Conference on 3D Vision*, pp. 127–134, 2013a. URL `https://api.semanticscholar.org/CorpusID:5296119`.

Changchang Wu. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision-3DV 2013*, pp. 127–134. IEEE, 2013b.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7623–7633, 2023.

Guangkai Xu, Wei Yin, Hao Chen, Chunhua Shen, Kai Cheng, and Feng Zhao. Frozenrecon: Pose-free 3d scene reconstruction with frozen depth models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9276–9286. IEEE, 2023.

Xinyi Ye, Weiyue Zhao, Tianqi Liu, Zihao Huang, Zhiguo Cao, and Xin Li. Constraining depth map geometry for multi-view stereo: A dual-depth approach with saddle-shaped depth cells. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17661–17670, 2023.

Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1983–1992, 2018. URL `https://api.semanticscholar.org/CorpusID:3714620`.

Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *ArXiv*, abs/2206.00665, 2022. URL `https://api.semanticscholar.org/CorpusID:249240205`.

Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2d feature representations by 3d-aware fine-tuning. *arXiv preprint arXiv:2407.20229*, 2024.

Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. What does stable diffusion know about the 3d scene? *arXiv preprint arXiv:2310.06836*, 2023.

David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023a.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements DINO for zero-shot semantic correspondence. In *NeurIPS*, 2023b. URL `https://openreview.net/forum?id=lds9D17HRd`.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3076–3085, 2024.

Zhe Zhang, Rui Peng, Yuxi Hu, and Ronggang Wang. Geomvsnet: Learning multi-view stereo with geometry perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21508–21518, 2023c.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *ICCV*, 2023.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024. URL `https://github.com/hpcaitech/Open-Sora`.

Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. Deeptam: Deep tracking and mapping with convolutional neural networks. *International Journal of Computer Vision*, 128:756 – 769, 2019. URL `https://api.semanticscholar.org/CorpusID:201815307`.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

A Appendix

You may include other additional sections here.