Self Introduction

- I am a final-year MPhil. student at National Engineering Laboratory for Video and Vision Technology, **Peking University**, under the supervision of Prof. Ronggang Wang. Before that, I received my BEng. of Computer Science and Technology from **Shandong University.**
- I'm interested in **3D vision**, includes 3D representation, neural rendering (NeRF 3DGS), 3D AIGC, and multi-view stereo.

Homepage: https://gaohchen.github.io

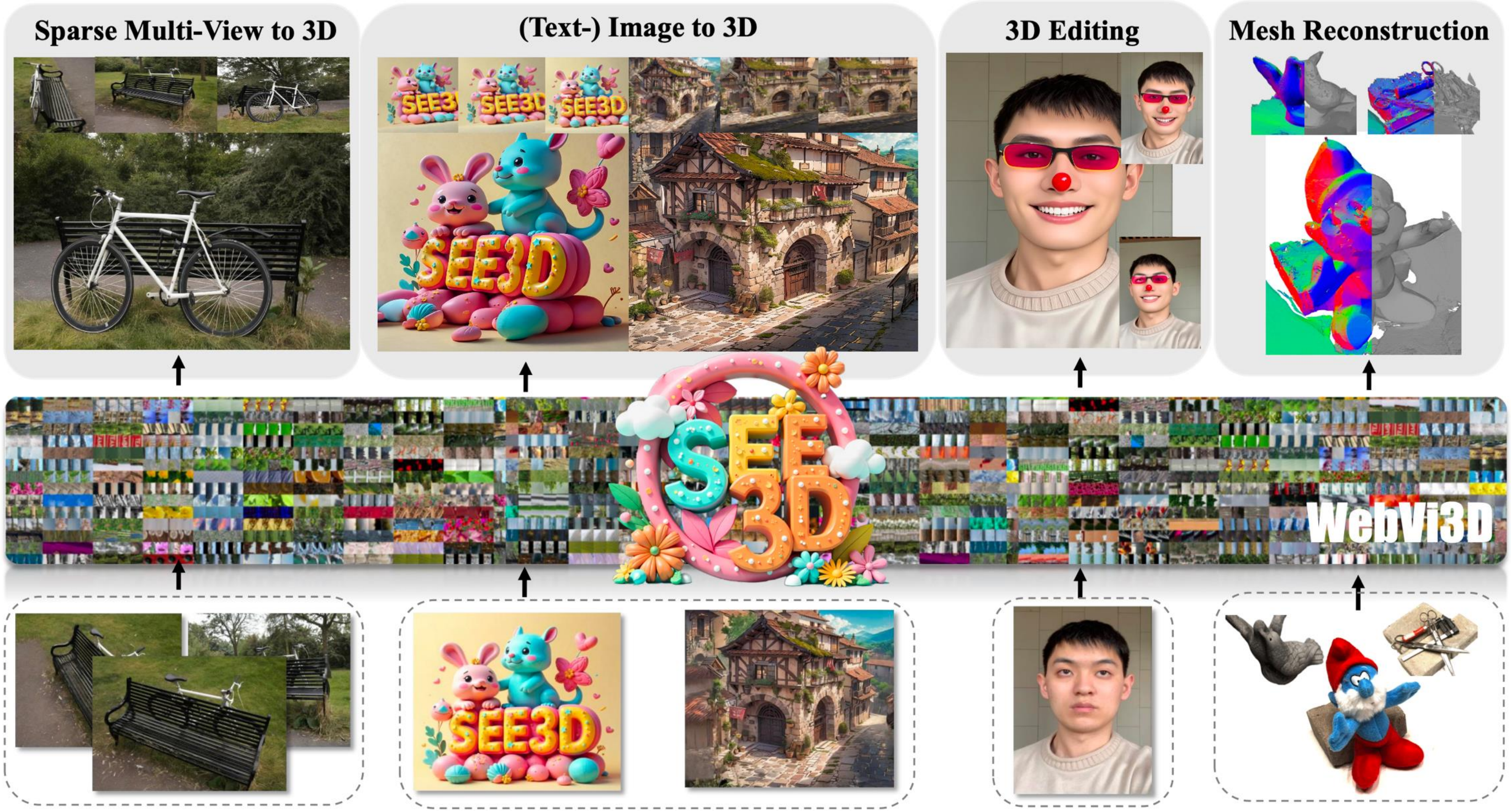
You See it, You Got it: Learning 3D Creation on Pose-Free Videos at Scale

Beijing Academy of Artificial Intelligence (BAAI)
Huachen Gao

2024/12/16

Generation Results (1/3 View Input)

Path to General 3D World Model



Motivations

- Existing 3D generation typically rely on **limited-scale 3D "gold-labels",** as 3D representations (mesh, GS, nerf, shape2vec...) or camera poses.
- We see our 3D world without relying on specific 3D representation, Instead, we shape this sense by **multi-view observations** throughout our lives.

Can models also learn universal 3D priors from large scale multi-view images?

Solution: Internet Videos + Multi-View Diffusion

Fossil Fuel: 3D Aware Data

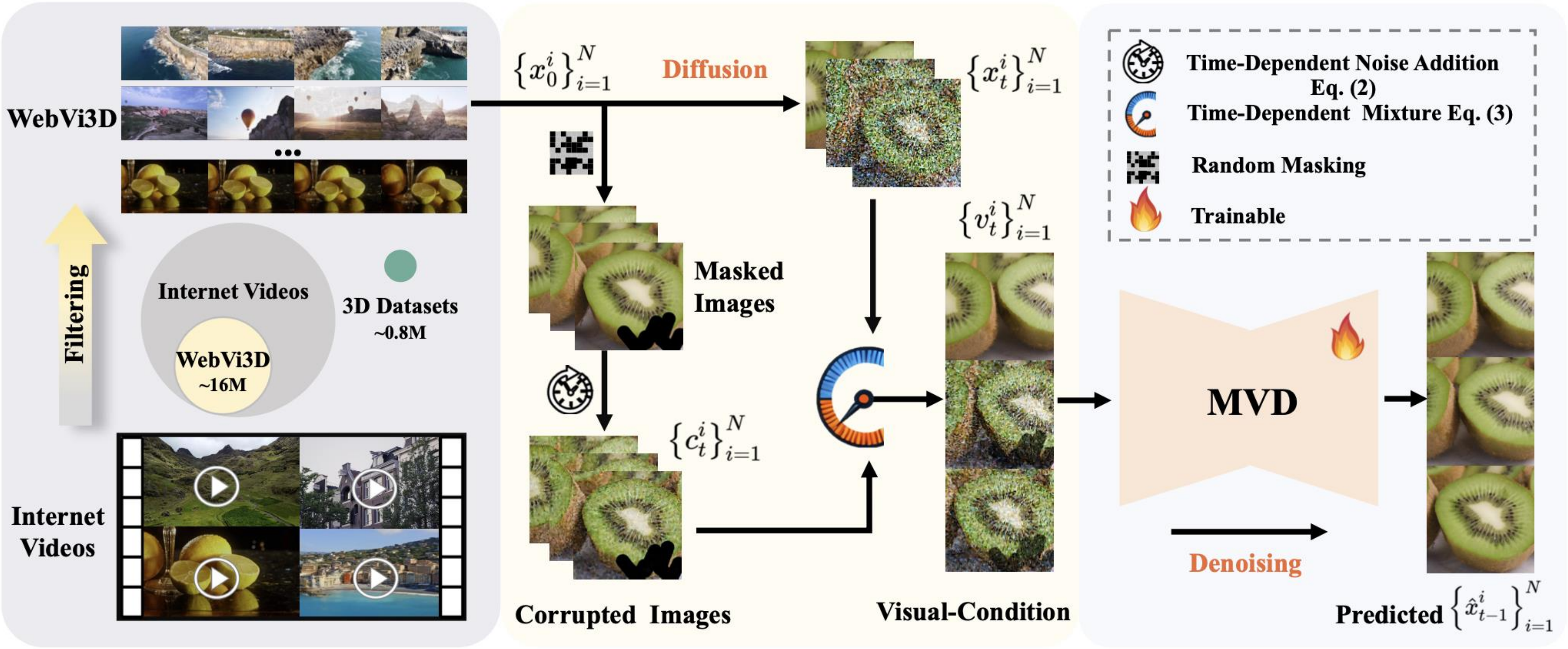
- **WebVi3D: Internet 3D Videos Curation**
  - Our curation pipeline consists of four core steps:
    1. Temporal-Spatial Downsampling,
    2. Semantic-Based Dynamic Recognition,
    3. Non-Rigid Dynamic Filtering
    4. Tracking-Based Small Viewpoint Filtering.
  - We collect approximately **25.48M** open-sourced raw videos, totaling **44.98 years** from the Internet, covering a wide range of categories, such as landscapes, drones, animals, plants, games, and actions.
  - Finally **16M** Video Clips, **320M** Multiview images (DLV3D (0.01M)、RealEstate10K (0.08M)、MVImgNet (0.22M) 和 Objaverse (0.8M))

Fossil Fuel: 3D Aware Data

🙅 (a) Source Videos with Dynamic Areas (Row 1) / Small Viewpoints Variation (Row 2).

😎 (b) WebVi3D Examples with Qualified Videos.

See3D: Pose-Free Visual-Conditional MVD



(a) Video Data Curation (b) Time-Dependent Visual Condition (c) Visual Conditional MVD Model

See3D: Pose-Free Visual-Conditional MVD

- We aim for multi-view prediction: generating novel views along **specified camera trajectories** from a **single or sparse** input while **ensuring consistency** with the input appearance.
- visual-condition can be derived from **pixel-space hints** within the original video implicitly guide the model to learn camera control.
- Moreover, it should be robust enough to handle domain gaps between **task-specific visual cues** and pixels extracted from video data.

See3D: Pose-Free Visual-Conditional MVD

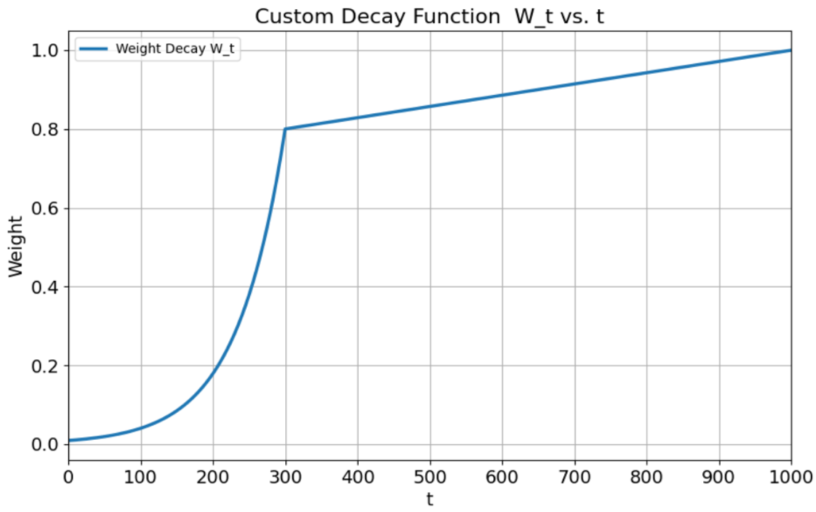
- **Time-dependent Visual Condition**
  - **Random Masking**
  - **Time-dependent Noise**

$$C_t = \sqrt{\bar{\alpha}_{t'}}(1 - M)\mathbf{X}_0 + \sqrt{1 - \bar{\alpha}_{t'}}\boldsymbol{\epsilon}. \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}).$$

  - **Time-dependent Mixture**

$$V_t = [W_t * C_t + (1 - W_t) * X_t; M],$$

$$M = \left\{m^{0:S} \cup m^{S+1:N}\right\}$$



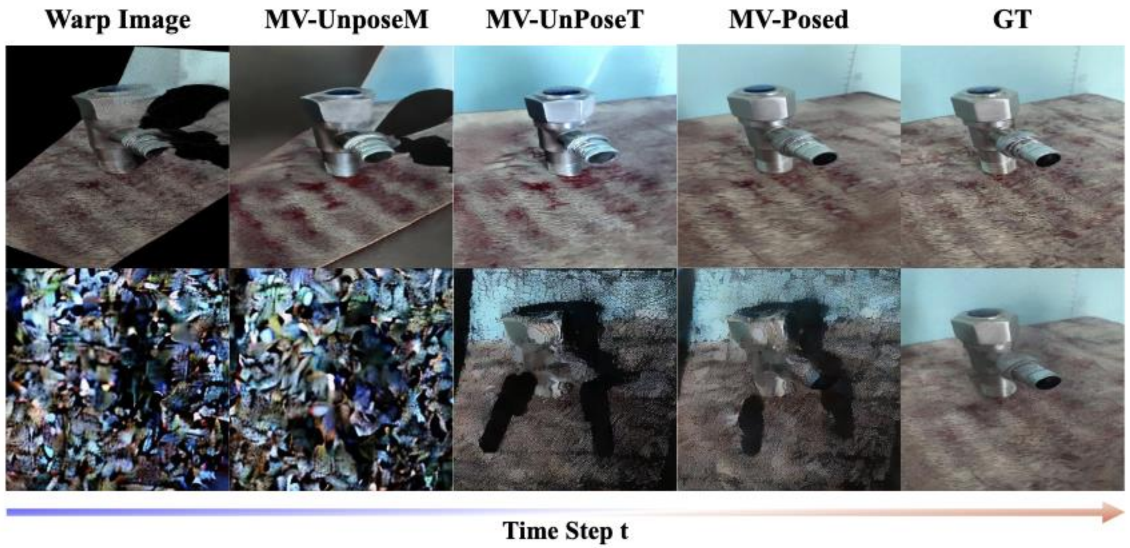
  - **Training Loss**

$$\mathbb{E}_{X_0, Y_0, \epsilon, t}\left[\|\epsilon_\theta(X_t, Y_0, V_t, t) - \epsilon\|_2^2\right]$$

See3D: Pose-Free Visual-Conditional MVD

Model	LPIPS ↓	PSNR ↑	SSIM ↑
MV-Posed	0.182	26.21	0.822
MV-UnPoseM	0.443	16.14	0.521
MV-UnposeT	0.194	25.56	0.811

Table 3. **Ablation Study on Visual-condition.**

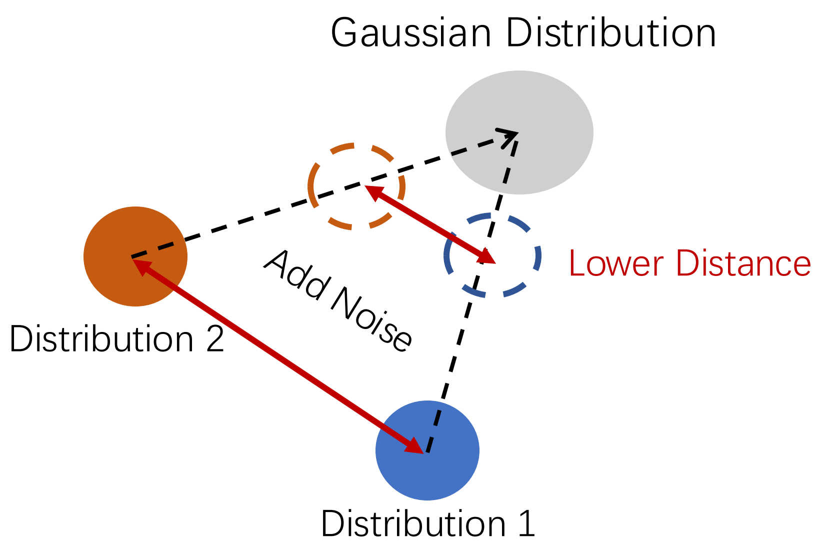


- We obtain warped images and form pairs with the **ground-truth multiviews** to train an MVD model, referred to as MV-Posed.
- We train an additional model **without any 3D annotations**, except for the modification of warp condition to the time dependent visual-condition Vt called MV-UnposeT.
- Meanwhile, we **employ randomly masked multiple views as condition** to train the model as an additional baseline, called MV-UnposeM.

See3D: Pose-Free Visual-Conditional MVD

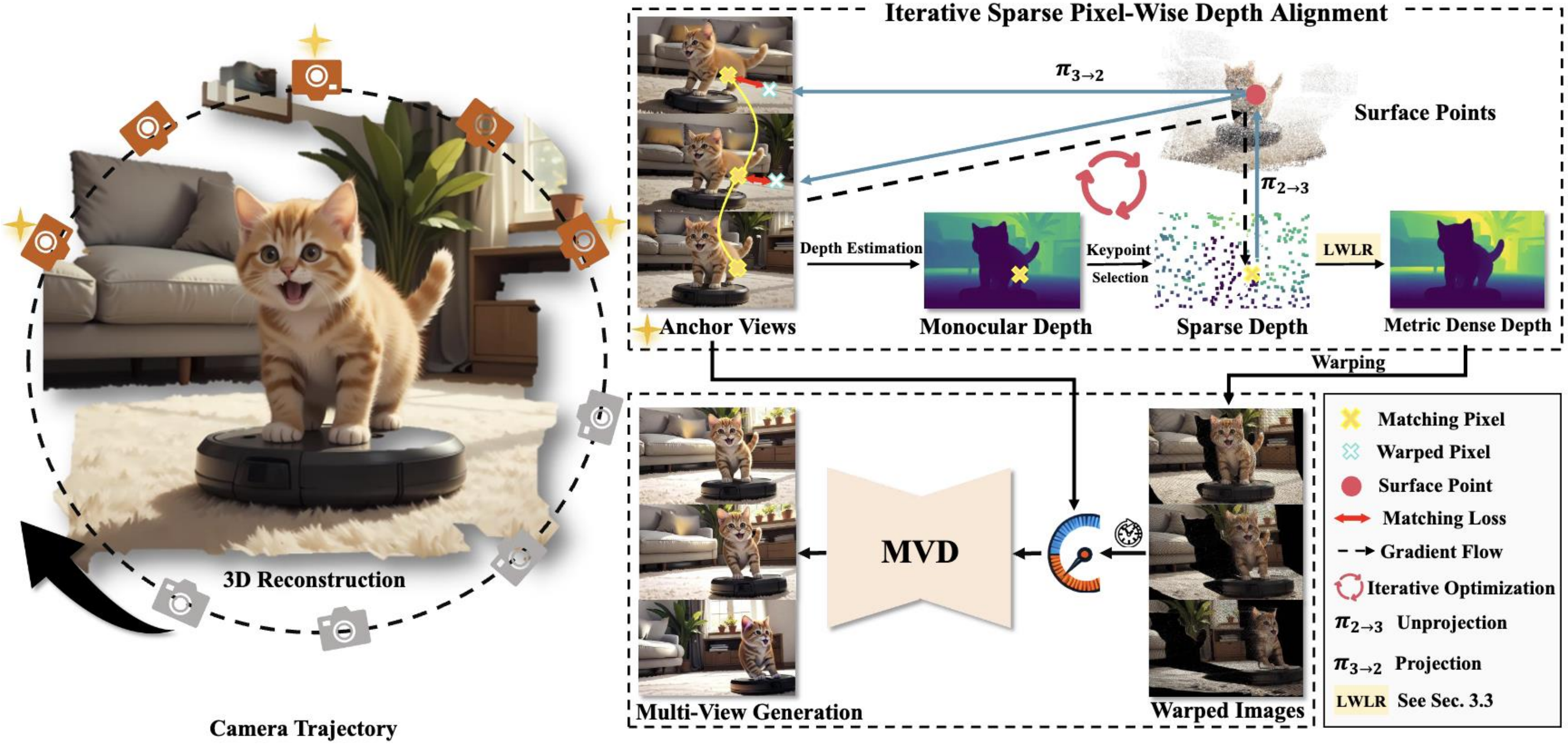
Model	LPIPS ↓	PSNR ↑	SSIM ↑
MV-UnposeT	0.194	25.56	0.811
MV-UnposeT-10%	0.187	25.95	0.817
MV-UnposeT-20%	0.183	26.19	0.820
MV-UnposeT-60%	0.181	26.14	0.819
MV-Posed	0.182	26.21	0.822

Table 6. **Ablation on Supplementary 3D Data.**



- We progressively introduce 3D pose annotations at levels of 10%, 20%, 60%, and 100% into the training set. When the training data is entirely composed of 3D annotations, the model configuration is equivalent to the MV-Posed model.

Multi-View Generation



Starting with one or a few input views, we iteratively generate warped images as visual hints, guided by predefined camera poses and estimated global depth. See3D is then utilized to generate novel views along the predefined camera trajectory, conditioned on the proposed visual-condition.

Multi-View Generation

- Pixel-wise Depth Scale Alignment

$$\alpha^{k*}, \beta^{k*} = \underset{\alpha^k, \beta^k}{argmin} \, ||\hat{d}_n^{k*} K_i T_i T_n^{-1} K_n^{-1} m_n^t - m_i^t||_2^2$$

- Global Metric Depth Recovery
- Novel View Generation

$$I_j = \textbf{See3D}(\hat{I}_j, M_j, \{I_0, I_k\})$$

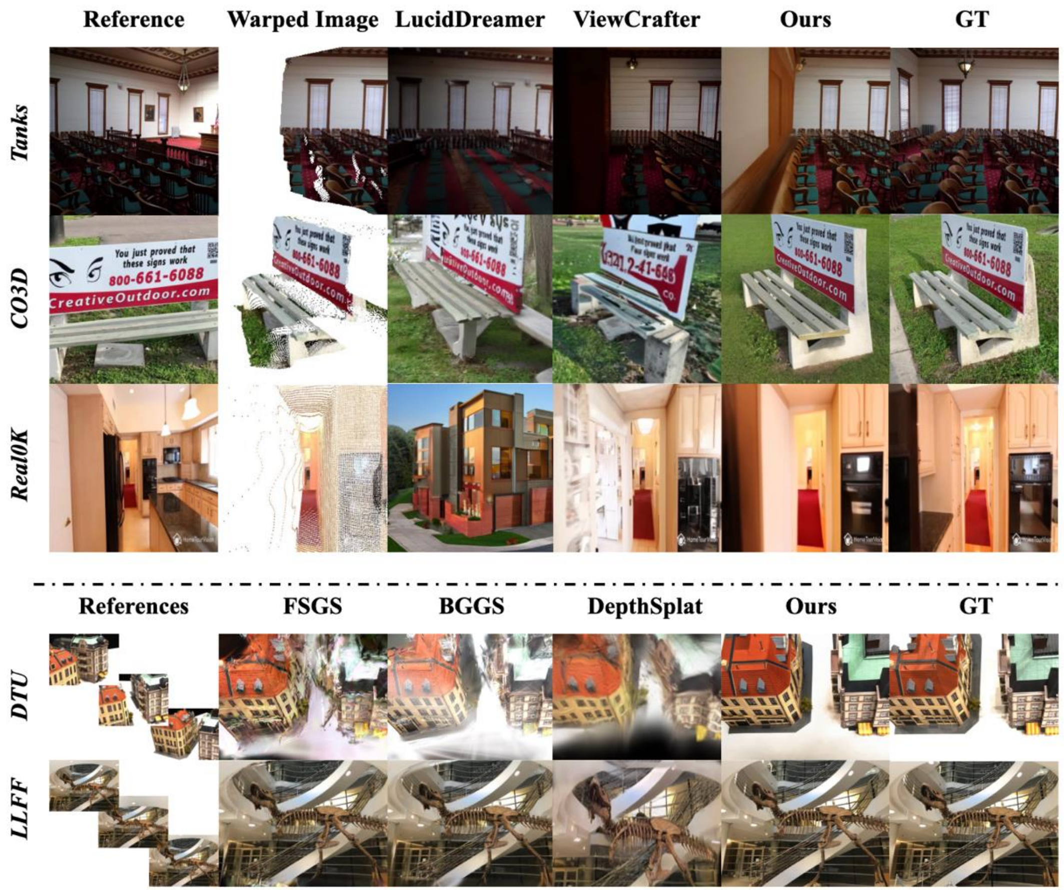
- 3D Reconstruction
  - Inter and inner frame diversity.
  - LPIPS loss + pose refinement.

Figure 11. **Ablation on Pixel-level Depth Alignment.**

Implementation

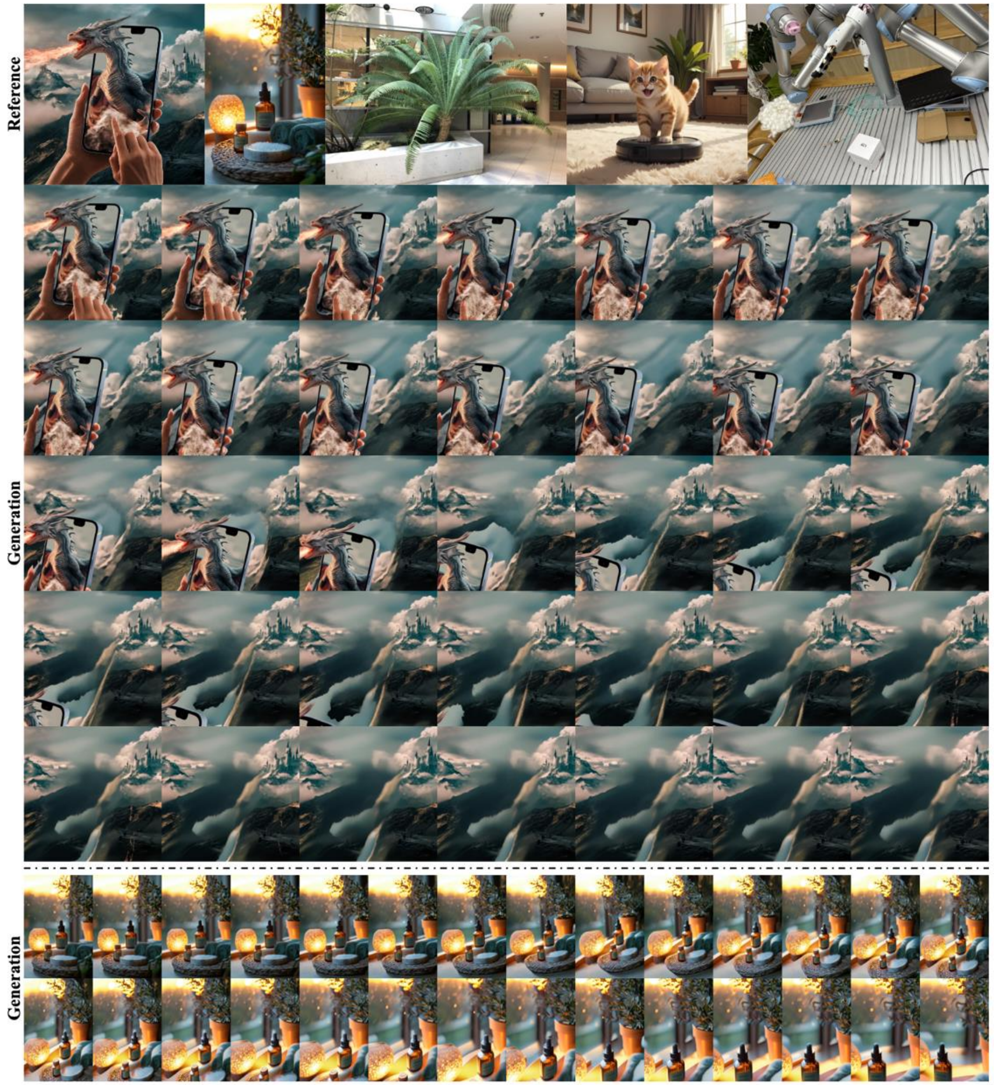
- The main backbone of See3D model is based on the structure of **2D diffusion models** but integrates **3D self-attention** to connect the latents of multiple images.
- We initialize the See3D model from **MVDream**, trained at a resolution of **512 × 512**.
- We render some extra multi-views or extract clips from **3D datasets** such as Objaverse, CO3D, RealEstate10k, MVImgNet, and DL3DV, forming a supplemental 3D dataset with fewer than 0.5M samples. During training, this supplemental data is randomly sampled and incorporated into our WebVi3D dataset (~16M).
- The See3D model is trained on 114 × NVIDIA-A100-SXM4-40GB GPUs over approximately 25 days.

Single/Sparse View Reconstruction



Methods	Tanks-and-Temples [43]			RealEstate10K [129]			CO3D [75]		
Single View	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
LucidDreamer [12]	13.11	0.314	0.485	15.24	0.545	0.357	13.90	0.412	0.473
ZeroNVS [77]	13.38	0.344	0.525	15.37	0.556	0.397	14.23	0.444	0.495
MotionCtrl [104]	14.31	0.405	0.436	16.30	0.596	0.363	16.16	0.515	0.418
ViewCrafter [121]	19.66	0.609	0.238	21.93	0.797	0.161	20.17	0.664	0.283
ViewCrafter* [121]	19.13	0.616	0.255	20.49	0.802	0.183	19.07	0.678	0.339
Ours	23.76	0.735	0.191	25.36	0.854	0.146	24.28	0.765	0.251
Sparse Views (3 Views)	**LLFF [64]**			**DTU [37]**			**MipNeRF-360 [3]**		
Zip-NeRF† [4]	17.23	0.574	0.373	9.18	0.601	0.383	12.77	0.271	0.705
MuRF [113]	21.34	0.722	0.245	21.31	0.885	0.127	-	-	-
FSGS [130]	20.31	0.652	0.288	17.34	0.818	0.169	-	-	-
BGGS [27]	21.44	0.751	0.168	20.71	0.862	0.111	-	-	-
ZeroNVS† [77]	15.91	0.359	0.512	16.71	0.716	0.223	14.44	0.316	0.680
DepthSplat [114]	17.64	0.521	0.321	15.59	0.525	0.373	13.85	0.254	0.621
ReconFusion [107]	21.34	0.724	0.203	20.74	0.875	0.124	15.50	0.358	0.585
CAT3D [23]	21.58	0.731	0.181	22.02	0.844	0.121	16.62	0.377	0.515
Ours	23.23	0.768	0.135	28.04	0.884	0.073	17.35	0.442	0.422

3D Editing & Open World Generation



Limitations & Future Works

- Our model facilitates open-world 3D content creation from large-scale video data, **eliminating the need for costly 3D annotations**.
- By leveraging visual data from the **rapidly growing Internet videos**, it accelerates 3D creation in real-world applications.
- Future research could extend the model to simultaneously generate **3D and 4D content** for dynamic scenes.
- While the data scaling approach is effective, the **scalability of the model** itself has not been explored.

You See it, You Got it: Learning 3D Creation on Pose-Free Videos at Scale

Thanks for Listening!

Huachen Gao

2024/12/16