OrbitArch: 360° Architectural Orbit Video Generation

I-Ting Tsai Bharath Hariharan
Cornell University
{itingtsai, bharathh}@cs.cornell.edu

Abstract

Generating a looping 360° orbit video around a building from a single image requires large-viewpoint camera motion that no existing video dataset captures at scale. We introduce OrbitArch, a framework that bootstraps this capability without any real orbit videos, relying instead on synthetic data and geometry-aware self-training. Our pipeline begins with orbit trajectories rendered around procedural cuboid buildings, whose backgrounds are filled in via video-diffusion inpainting. Because naive synthetic generation produces geometrically incoherent samples, we apply a two-stage filter: COLMAP-based 3D reconstruction verifies multi-view consistency, while a camera-coverage gate rejects partial or unstable orbits. A video diffusion model trained on the surviving data then generates new orbit videos that are passed through the same filters, creating a self-training loop that extends from simple box geometry to freeform facades. On automatic metrics, a user study, and downstream 3D Gaussian Splatting reconstruction, OrbitArch achieves the highest camera coverage, 3D point density, and novel-view fidelity among all compared methods, and its outputs are strongly preferred by human evaluators. Our results show that synthetic orbits around simple primitives, when coupled with geometry-aware filtering and iterative self-training, are sufficient to produce professional-quality 360° architectural visualizations of freeform designs—without a single real orbit video.

1. Introduction

Generative models are powerful interpolators of their training distribution, but what happens when the capability we want lies outside that distribution entirely? When no real data captures the behavior we need to teach, can synthetic data fill the gap?

We instantiate this question on a problem where the gap is unavoidable: generating 360° orbit videos of buildings from a single image. Orbit renderings—continuous camera trajectories circling a building—are essential for communicating a design across every viewpoint, yet producing them requires labor-intensive 3D modeling, material authoring, and cinematic lighting. Video diffusion models could bypass this pipeline, but models trained on natural video rarely encounter large-viewpoint orbital motion. When prompted for full orbits, they often produce limited camera travel, subject drift, or buildings that collapse into flattened geometry at side views [37, 44, 60]. The core obstacle is a data gap: no large-scale dataset of real architectural orbit videos exists, and existing multi-view datasets contain either isolated objects [7, 8, 36] lacking scene context, or SfM reconstructions [23, 50] biased toward landmarks and rarely spanning full orbits.

These observations motivate a different starting point: instead of waiting for real orbit data to emerge, we ask whether geometric supervision alone—extractable from purely synthetic sources—is enough to teach a pretrained video model the missing motion. Our key insight is that *synthetic data, despite its visual simplicity, can provide sufficient geometric supervision to bootstrap orbital camera motion*—and that self-training with geometry-aware filtering can progressively close the remaining domain gap without any real orbit video data. We present OrbitArch, a framework built on three synthetic-data components:

- **Hybrid synthetic data via inpainted backgrounds.** We render orbit trajectories around procedural cuboid buildings [49], producing geometrically consistent camera motion by construction, then replace static synthetic backgrounds with dynamic, photorealistic surroundings with natural parallax via video diffusion inpainting [16]. A split-and-condition strategy ensures loop closure, decoupling geometric supervision from visual realism.
- **Geometry-aware quality filtering.** A two-stage filter combining COLMAP-based [38, 39] 3D consistency verification with a camera-coverage threshold (>300°) retains only geometrically coherent, near-complete orbits—serving as an automated quality gate for synthetic data curation.

Non-archival workshop paper. An extended version of this work is currently under review at another venue.

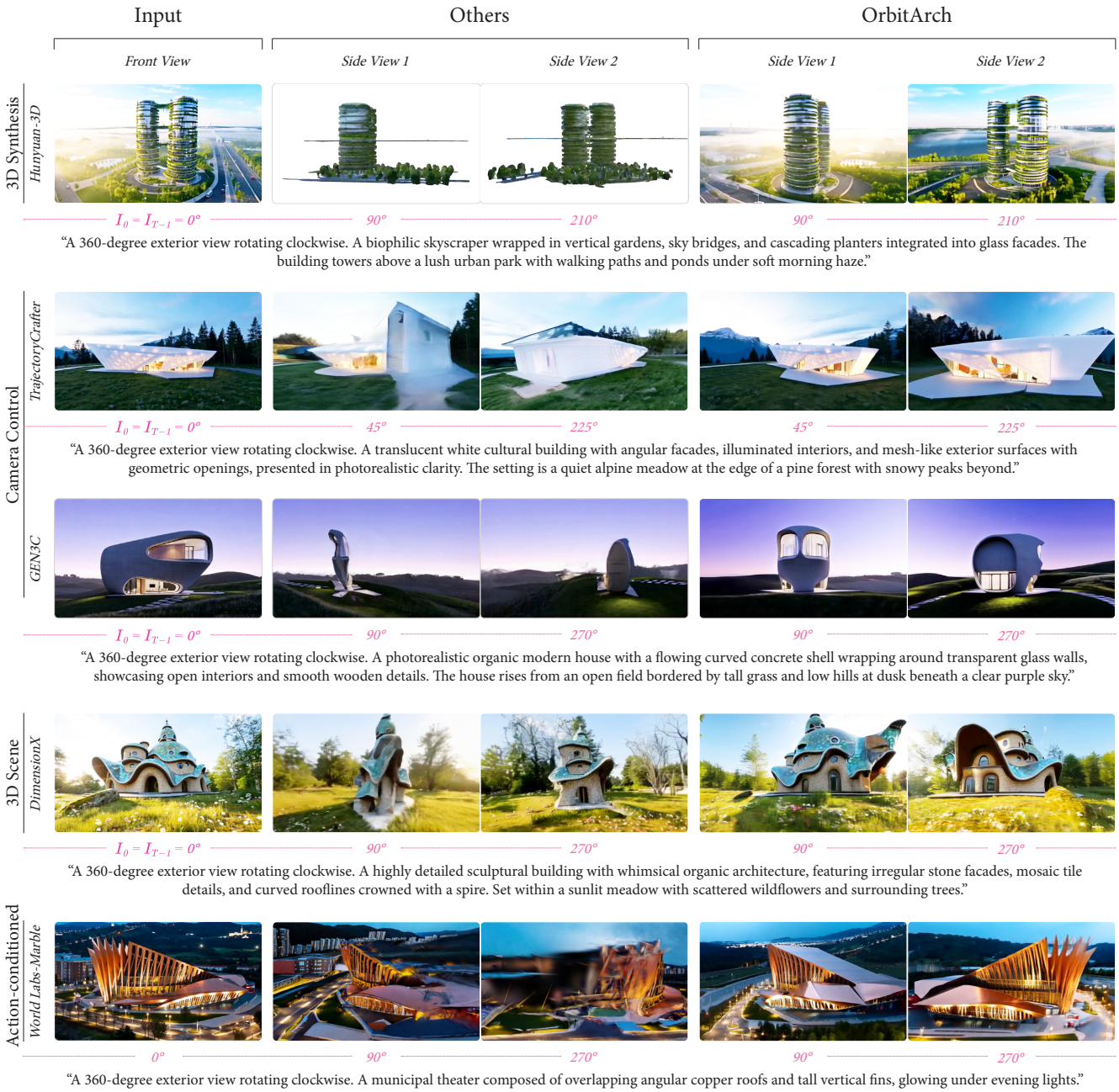

Figure 1. **Multi-view Comparisons.** Hunyuan3D [45, 46] (top, middle) recovers coarse geometry but lacks photorealistic quality. TrajectoryCrafter [60] (second row, middle) hallucinates spurious extruded volumes on facades. GEN3C [37] (third row, middle) and DimensionX [44] (fourth row, middle) both show geometric collapse at side views, where buildings exhibit flattened geometry rather than volumetric structure. Marble [54] (bottom, middle) degrades with visible artifacts at non-frontal views. OrbitArch (right) maintains geometric consistency, photorealism, and coherent detail across the full 360° orbit.

- **Self-training as data amplification.** The trained model generates orbit videos of diverse buildings; these pass through the same geometry-aware filter and rejoin the training pool, systematically expanding diversity and photorealism from simple cuboids to complex architecture—without manual annotation or real footage.

OrbitArch achieves the highest camera coverage (290.90°), 3D point density (14,705 registered points), and downstream 3DGS reconstruction quality (24.97 PSNR) among all compared methods, with significantly more convincing three-dimensional volume confirmed by reconstruction metrics, VQA-based evaluations, and a user study. Our

results show that geometry-grounded synthetic data, automated filtering, and iterative self-training can bootstrap orbital motion in a model that never observed it, and we see this as one instance of a broader recipe for tackling capabilities absent from existing training distributions.

2. Related Work

Existing approaches to novel view synthesis and 3D generation struggle with full 360° architectural orbits, and the root cause is a training data problem. SDS-based methods [19, 24, 25, 32, 34, 53] and feed-forward 3D models [15, 17, 28–30, 35, 40, 41, 43, 45–48, 55, 57] produce object-level geometry but cannot place a building in a coherent surrounding scene—a basic requirement for architectural visualization [49]. Video diffusion models [3–5, 9, 11–13, 20, 31, 33, 42, 51, 58, 62] generate high-quality clips but are trained on natural video that rarely contains large-viewpoint motion, so they default to small pans and zoom-ins when pushed toward full orbits [59]. Camera-controllable methods [1, 6, 10, 21, 51, 56, 61, 63] offer explicit viewpoint control but exhibit characteristic failures: limited rotation range [2, 61], hallucinated volumes [60] (Fig. 1 second row, middle), and flat-building collapse or unrealistic massing [37, 44] (Fig. 1 third row, middle; fourth row, middle). World models such as Marble [54] enable interactive scene navigation, but without multi-view supervision over full trajectories they exhibit geometric gaps and texture instability under novel viewpoints (Fig. 1 bottom, middle). The unifying limitation across all these families is the *absence of training data with large angular coverage*: existing datasets contain either isolated objects [7, 8, 36] or SfM reconstructions [23, 50] biased toward landmarks and rarely spanning complete orbits. OrbitArch addresses this gap by constructing synthetic orbit data from a procedural modeling method [49] and closing the domain gap via a generate-filter-retrain loop, using 3D geometric consistency as an automated quality signal.

3. Method

Our goal is to train a video diffusion model to produce looping 360° orbit videos of buildings from a single image I_0, generating a sequence $(I_0, I_1, \ldots, I_{T-1})$ where $I_0 = I_{T-1}$. The central challenge is that *no real training data for this task exists*: web-scale video corpora contain virtually no complete orbital trajectories around buildings. Our pipeline addresses this entirely through synthetic data construction, automated quality filtering, and self-training-based data amplification (Fig. 3).

3.1. Constructing Synthetic Training Data

Our synthetic data strategy combines the geometric controllability of 3D rendering with the visual richness of generative models—leveraging the strengths of each while avoiding their individual limitations.

Procedural 3D rendering for geometric grounding. Following [49], we procedurally model cuboid buildings in 3D and render orbit trajectories with textured facades. These renderings are geometrically consistent by construction: the camera completes a full revolution and returns to its starting pose, guaranteeing $I_0 = I_{T-1}$. This synthetic geometry provides the orbital camera motion supervision that is entirely absent from natural video datasets. However, the rendered backgrounds are static and visually simplistic—directly training on such data produces a model that learns to rotate the building but generates frozen, unrealistic surroundings. The synthetic data is geometrically sufficient but visually insufficient.

Background Completion via Video Inpainting. To bridge the visual domain gap without sacrificing geometric grounding, we decouple the training signal: the building geometry is anchored by the synthetic render, while the background is regenerated using a video diffusion inpainting model [16]. We remove the synthetic background and inpaint new, photorealistic surroundings in the masked regions. This hybrid strategy produces training pairs that are simultaneously geometrically controlled (from 3D rendering) and visually realistic with background parallax (from the generative model)—a combination that neither source provides alone.

Split-and-condition inpainting for loop closure. A naive application of video inpainting fails to produce looping backgrounds: the model generates temporally smooth content but cannot ensure spatial consistency when the camera returns to its starting viewpoint. Directly constraining $I_0 = I_{T-1}$ causes degenerate behavior—a frozen background without parallax.

We resolve this with a two-pass strategy that splits the orbit at the midpoint $\lfloor \frac{T-1}{2} \rfloor$ (see Fig. 2). The *first half* $(I_0, \ldots, I_{\lfloor \frac{T-1}{2} \rfloor})$ is inpainted freely, conditioned only on a text prompt, producing dynamic surroundings with natural parallax. The *second half* $(I_{\lfloor \frac{T-1}{2} \rfloor}, \ldots, I_{T-1})$ is conditioned on two boundary frames: the last frame of the first-half video (for temporal continuity) and the first frame as the target for I_{T-1} (for loop closure), with masks $M'_{\lfloor \frac{T-1}{2} \rfloor}$ and M'_{T-1} protecting both boundaries. This constrains the inpainting model to synthesize a background that smoothly departs from the midpoint and returns to the starting view.

Limitations of the raw synthetic data. The resulting videos exhibit dynamic, parallax-consistent backgrounds

Figure 2. **Synthetic data construction and failure modes.** We render orbit videos of procedurally generated cuboid buildings [49] and inpaint photorealistic backgrounds [16]. Naive inpainting breaks loop closure (second row, left); constraining $I_0 = I_{T-1}$ yields a frozen background (second row, right). Our split-and-condition strategy produces looping backgrounds with parallax (middle), but artifacts persist: spurious structural additions (bottom row, left) and unrealistic surroundings (bottom row, right). The geometric artifacts motivate the geometry-aware filtering we describe in Sec. 3.2.

that loop seamlessly. However, the inpainting model introduces failure modes that make the raw data unsuitable for direct training: structural hallucinations (phantom massing attached to the building) and unrealistic backgrounds. These artifacts motivate the automated filtering pipeline described next.

3.2. Geometry-Aware Filtering as Automated Data Curation

A recurring challenge in synthetic data pipelines is quality control: generative models produce diverse outputs, but not all are suitable for training. Rather than relying on perceptual heuristics or manual inspection, we use *3D geometric consistency* as an automated, task-appropriate quality criterion. Our two-stage filter operates on the principle that a good orbit video must be both structurally coherent across views and orbitally complete.

3D consistency filter. We run COLMAP [38, 39] structure-from-motion on the frames of each inpainted video. Videos in which all frames are successfully registered into a consistent 3D reconstruction are retained; those where COLMAP fails to register a substantial fraction of frames—indicating hallucinated geometry or structural breaks introduced by inpainting—are discarded. This leverages an off-the-shelf reconstruction tool as a binary quality gate: if the generated video is sufficiently 3D-consistent for SfM to succeed, it is sufficiently consistent for training.

Camera coverage filter. Even when COLMAP [38, 39] succeeds, the reconstructed trajectory may not correspond to a near-complete orbit. We extract the 3D camera cen-

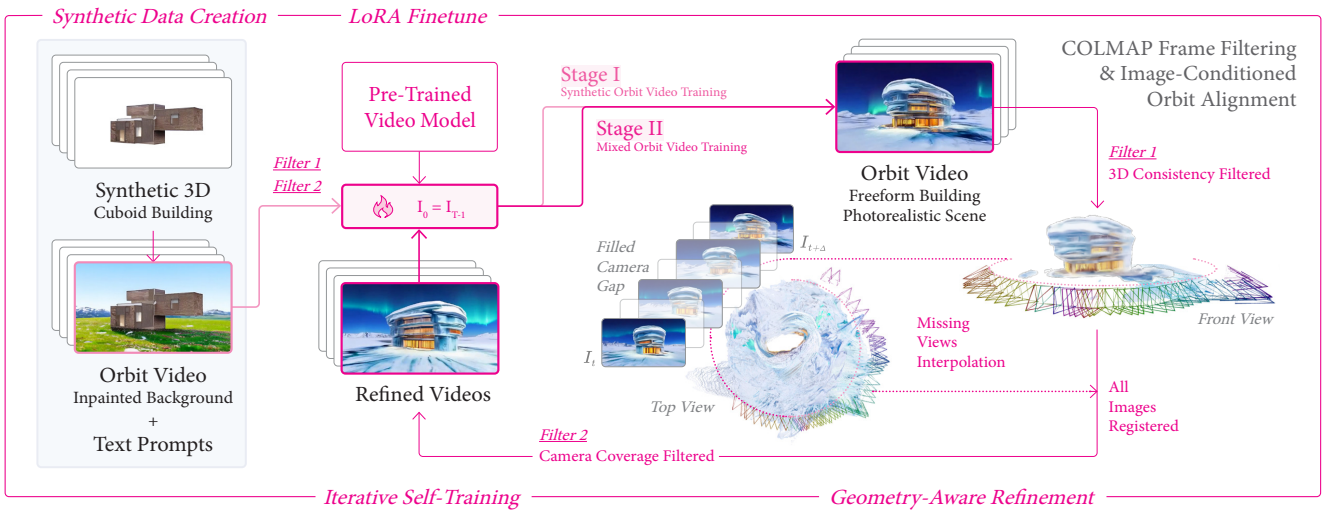

Figure 3. **Overall synthetic data pipeline. Stage I:** Synthetic cuboid buildings are rendered with inpainted backgrounds (Fig. 2), then filtered via COLMAP [38, 39] registration for 3D consistency and a camera-coverage threshold of $\geq 300°$. We fine-tune Wan Video [51] with LoRA [14] on the surviving videos, with loop-closure conditioning. **Stage II (data amplification):** The Stage I model generates orbit videos of diverse buildings; these are filtered identically, with partial registrations optionally recovered via orbit-aware interpolation. Stage II is trained on the union of model-generated and original synthetic data, extending the model from procedural cuboids to freeform architecture through iterative self-improvement.

ters, fit an orbital plane via SVD, project cameras onto this plane, and compute per-frame azimuth angles relative to the estimated orbit center. Angular coverage is defined as $360 - \Delta_{\max}$, where $\Delta_{\max}$ is the largest gap between consecutive cameras in sorted azimuth order. We retain only videos with coverage exceeding 300° and enforce smoothness constraints on camera radius variation, discarding trajectories with abrupt spatial jumps.

Together, these filters automatically curate the synthetic dataset, removing the failure modes introduced by inpainting while requiring no human annotation. The surviving samples are geometrically verified and orbitally complete, providing clean supervision for the downstream model.

3.3. Self-Training as Synthetic Data Amplification

The filtered synthetic data is geometrically clean but limited in diversity: buildings are procedural cuboids, and backgrounds, while dynamic, do not span the full range of real architectural scenes. We close this domain gap through a self-training procedure that uses the model's own outputs to *amplify* the synthetic training set—progressively expanding its visual diversity and architectural complexity without any real data.

Stage I: Bootstrapping from synthetic orbits. We fine-tune a pretrained video diffusion model (Wan [51]) using LoRA [14] on the filtered inpainted videos. We condition the model on the same reference image at both the first and last frame positions ($I_0 = I_{T-1}$), explicitly encoding loop closure into the generation process. At this stage, the model acquires the geometric prior for orbital camera motion from the synthetic data—it learns *what an orbit looks like*—even though its visual quality remains constrained by the cuboid training distribution.

Stage II: Expanding the synthetic distribution. The Stage I model generates orbit videos conditioned on diverse architectural images, moving well beyond the procedural cuboids. These model-generated videos are passed through the same geometry-aware filter. For videos where COLMAP registers only a subset of frames, we use the Stage I model for orbit-aware frame interpolation: registered frames serve as keyframes, and the model fills angular gaps to recover complete 360° sequences, which are re-verified through both filters.

Videos that pass filtering are combined with the original synthetic training data, and the model is retrained. This establishes a self-improving data loop: each iteration produces more diverse, more photorealistic training samples that remain geometrically verified. The geometry-aware filters are critical to preventing drift—they act as a consistent quality gate ensuring that only 3D-consistent samples re-enter the training pool, regardless of the visual complexity of the generated content.

The self-training strategy rests on a key observation about synthetic data: *visual simplicity does not preclude geometric sufficiency*. Procedural cuboids lack architectural richness, but they provide enough 3D structure to teach a

Table 1. **Quantitative Comparison of 3D Consistency and Camera Coverage.** We evaluate 3D consistency via COLMAP [38, 39] registration and orbital camera coverage (filtered: only temporally smooth, radius-consistent cameras counted). OrbitArch-StageII achieves the highest coverage (290.90°) and registered points (14,705), indicating the most complete and geometrically dense orbits. Low reprojection error alone can be misleading—Wan-FLF2V [51] achieves the lowest error by registering very few frames, reflecting minimal camera motion. Bottom group: pretrained base models without orbit-specific training. These metrics capture orbital geometry but not perceptual quality or volumetric fidelity—see Fig. 4 and Fig. 5 for evidence on flat-building and split-merge artifacts that registration cannot detect.

Method	Valid ↑	Reg. Frames ↑	Reg. Points ↑	Median RE ↓	Coverage ↑
OrbitArch-StageII	**100.00%**	98.56%	**14705**	1.054 px	**290.90°**
OrbitArch-StageI	96.00%	93.61%	13894	1.024 px	278.55°
DimensionX	100.00%	**98.59%**	8774	0.804 px	289.74°
GEN3C	100.00%	86.23%	12881	1.085 px	244.47°
ReCamMaster	100.00%	95.35%	9658	1.316 px	200.88°
ViewCrafter	97.97%	78.20%	4402	1.064 px	150.29°
GenXD	32.63%	14.83%	164	0.610 px	139.12°
TrajectoryCrafter	100.00%	48.00%	4054	0.800 px	114.27°
CameraCtrl	47.37%	11.17%	67	0.548 px	40.27°
Wan-T2V	99.50%	97.83%	9057	0.992 px	134.85°
Hunyuan-I2V	89.50%	88.93%	10845	0.881 px	132.74°
Wan-I2V	93.33%	93.52%	6883	0.969 px	132.65°
Wan-FLF2V	22.50%	9.45%	220	**0.173** px	27.81°

Table 2. **Text-Video Alignment Evaluation.** Semantic alignment measured via Video-VQAScore (querying "Does this video show {text prompt of the scene description}?") and Video-CLIPScore [22, 26, 27] (text-video similarity). OrbitArch-StageII leads on both metrics, with Stage I→II gains confirming that self-training also improves semantic fidelity.

Method	Video-VQAScore ↑	Video-CLIPScore ↑
Prompt: *Does this video show ... ?*	*scene description*	
OrbitArch-StageII	**0.962**	**0.314**
OrbitArch-StageI	0.952	0.311
DimensionX	0.954	0.306
GEN3C	0.911	0.308

pretrained video model what orbital motion is. Once this geometric prior is established, the model can generalize to complex buildings—and the geometry-aware filter can verify that generalization, creating a feedback loop that systematically closes the gap between simple synthetic data and complex real-world architecture.

4. Experiments

We compare against camera-controllable video generation (GEN3C [37], ReCamMaster [2], ViewCrafter [61], TrajectoryCrafter [60], CameraCtrl [10]), 3D/4D scene generation (DimensionX [44], GenXD [65]), and pretrained video models without orbit-specific training (Wan [51], Hunyuan-Video [20]). All methods are evaluated on the same inputs. The comparison tests whether our synthetic data recipe outperforms alternative strategies for the same data-scarce task.

4.1. Comparison with Existing Methods

3D consistency and camera coverage. We run COLMAP registration on each output and report validity, frame registration rate, registered 3D points, reprojection error, and orbital coverage (Tab. 1). OrbitArch-StageII achieves the highest coverage (290.90°) and registered points (14,705) with 100% validity and 98.56% frame registration—despite being trained *entirely on synthetic data*. DimensionX [44] achieves comparable coverage (289.74°) but with far fewer points (8,774). Other camera-controllable methods peak at 244° (GEN3C [37]) or below, confirming their inability to complete full orbits. The pretrained base models all remain below 135° coverage, confirming that orbital motion must be explicitly taught and is absent from standard pretraining distributions.

Text-video alignment. OrbitArch-StageII achieves the highest Video-VQAScore (0.962) and Video-CLIPScore (0.314) among all methods (Tab. 2), indicating that synthetic-data training preserves semantic fidelity. The Stage I→II gain (0.952→0.962) suggests self-training enhances not only geometry but also the model's ability to realize diverse architectural styles.

Figure 4. **Flat-building Artifact.** DimensionX [44] generates buildings that appear plausible from front and back views but collapse into thin, depth-lacking profiles at side viewpoints.

Volumetric quality. We probe the *flat-building artifact*—buildings collapsing into thin structures at side views—using targeted Video-VQAScore prompts (Tab. 3). OrbitArch scores higher on all positive prompts (avg. 0.624 vs. 0.504) and lower on all negative prompts (avg. 0.074 vs. 0.243) compared to DimensionX [44]. The starkest contrast: OrbitArch scores 0.001 vs. 0.292 on "no visible depth from the side," confirming that synthetic cuboid supervision teaches volumetric structure that methods without orbit-level data cannot learn.

Qualitative results. Baseline failures include coarse geometry lacking presentation quality (Hunyuan3D [45, 46]), hallucinated volumes (TrajectoryCrafter [60]), unrealistic massing (GEN3C [37]), flat-building collapse (DimensionX [44]), and severe non-frontal artifacts (Marble [54]). OrbitArch maintains convincing 3D structure across the full trajectory (Figs. 1, 4 and 5). Orbit-level supervision from synthetic architectural data directly helps prevent these failure modes: geometric flattening and topological fragmentation.

Table 3. **Volumetric Quality Assessment via Video-VQAScore.** We probe perceived 3D quality using Video-VQAScore [22, 26, 27] with positive prompts (volumetric realism, architectural detail) and negative prompts (flat, thin, depth-lacking structures). OrbitArch scores higher on all positive and lower on all negative prompts than DimensionX [44], confirming significantly more convincing 3D structure and reduced flat-building artifacts.

Prompt	Video-VQAScore	
Prompt: *Does this video show ... ?*	OrbitArch	DimensionX
Positive (↑)		
"A building with realistic three-dimensional shape and volume"	**0.618**	0.516
"The building appears to have full three-dimensional volume when viewed from the side"	**0.721**	0.562
"The building has a visible side facade with architectural features"	**0.510**	0.389
"The side of the building shows windows, doors, and architectural depth"	**0.647**	0.547
Negative (↓)		
"A building that looks unnaturally thin or narrow when viewed from the side"	**0.111**	0.256
"The building has no visible front-to-back depth when seen from the side"	**0.001**	0.292
"The building looks two-dimensional and flat from the side view"	**0.109**	0.181
Average Positive (↑)	**0.624**	0.504
Average Negative (↓)	**0.074**	0.243

Table 4. **User Study.** 53 participants (including 18 architecture professionals) evaluated three methods on 18 video sets each, with identities hidden and order randomized. Orbit Extent is a 1–5 Likert scale; others are pairwise win rates (ties split 0.5). OrbitArch is preferred across all criteria.

Method	Orbit (1–5)↑	3D Con. (win)↑	Quality (win)↑	Prompt (win)↑
All participants (N=53)				
OrbitArch	**4.69**	**0.88**	**0.90**	**0.80**
DimensionX	4.64	0.19	0.15	0.35
GEN3C	3.41	0.04	0.05	0.05
Arch. professionals (N=18)				
OrbitArch	**4.70**	**0.87**	**0.91**	**0.86**
DimensionX	4.57	0.20	0.12	0.23
GEN3C	3.27	0.07	0.07	0.05

4.2. User Study

We conduct a user study with 53 participants, including architecture professionals (N=18) and the general public (N=35). Each participant views 18 sets of generated orbit videos with method identities hidden and presentation order randomized. Participants evaluate four criteria: *Orbit Extent* (how much of the 360° perimeter is visible, 1–5 Likert), *3D Consistency* (which method produces the most spatially coherent structure), *Perceived Quality* (most visually convincing result), and *Prompt Alignment* (best match

Figure 5. **Split-merge Artifact.** DimensionX [44] exhibits the split-merge artifact: the organic shell separates (top), rotates independently, and reattaches (bottom).

Table 5. **Novel View Synthesis via 3D Gaussian Splatting.** 3DGS [18] fit to raw generated orbit videos, with half the frames used for training and the other half held out for evaluation (no post-processing). OrbitArch leads on all three metrics, confirming that its orbit-level geometric coherence translates directly into high-fidelity 3D reconstruction.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
OrbitArch	**24.97**	**0.83**	**0.12**
DimensionX	20.40	0.68	0.16
GEN3C	18.20	0.60	0.22

to the input text). The latter three use pairwise comparisons with ties split 0.5.

Results are shown in Tab. 4. OrbitArch is strongly preferred on Perceived Quality (0.90 vs. 0.15) and 3D Consistency (0.88 vs. 0.19). DimensionX [44] achieves comparable Orbit Extent (4.64 vs. 4.69)—it completes near-full orbits, but the resulting buildings are not perceived as 3D-consistent or high-quality, consistent with the flat-building artifact (Fig. 4). Among architecture professionals, preferences are even more pronounced, with Prompt Alignment rising to 0.86, suggesting domain experts are particularly sensitive to whether the video faithfully reflects design intent. These results confirm that our synthetic-data-trained model produces outputs preferred by both general viewers and domain experts over methods using alternative training strategies.

4.3. 3D Reconstruction

To evaluate whether generated orbit videos support downstream 3D reconstruction, we fit 3D Gaussian Splatting [18] to the orbit videos generated by the top-performing methods, using alternating frames for training and held-out frames for evaluation. We report PSNR [64], SSIM [52], and LPIPS (Tab. 5). For fairness, all methods are evaluated on their raw video output—without any post-processing—through the same 3DGS pipeline.

OrbitArch achieves the best performance on all three metrics by substantial margins—24.97 PSNR, 0.83 SSIM, and 0.12 LPIPS—outperforming DimensionX [44] (20.40 / 0.68 / 0.16) by 4.57 dB PSNR and GEN3C [37] (18.20 / 0.60 / 0.22) by 6.77 dB. Despite using no explicit 3D representation during generation, OrbitArch produces orbit videos whose multi-view consistency is strong enough for off-the-shelf 3DGS optimization to recover high-fidelity reconstructions directly. The large gap over baselines confirms that the flat-building and split-merge artifacts they exhibit fundamentally degrade downstream reconstruction, whereas OrbitArch's orbit-level geometric coherence translates directly into superior novel-view synthesis.

5. Conclusion

We presented OrbitArch, a framework that generates full 360° orbit videos of buildings from a single image—trained entirely on synthetic data without real orbit footage, yet achieving the highest camera coverage and downstream reconstruction quality among all compared methods. Three lessons emerge from this result. First, visually simple synthetic data can supply sufficient geometric structure to bootstrap capabilities absent from a model's pretraining distribution. Second, automated geometry-aware filtering is essential for curating synthetic training sets: structure-from-motion serves as a free, task-appropriate quality verifier that catches the hallucinations generative inpainting inevitably introduces. Third, self-training creates a virtuous cycle in which filtered model outputs progressively diversify the training pool, closing the domain gap without manual intervention. Together, these form a general recipe—ground in controllable 3D geometry, filter with task-appropriate consistency checks, amplify through self-training—for tasks where the desired training distribution does not exist in the wild.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *Proc. CVPR*, 2025. 3

[2] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 3, 6

[3] 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. 3

[4] Andreas Blattmann, Robin Rombach, Patrick Esser, and Björn Ommer. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[5] Google DeepMind. Veo: Ai video generator. https://deepmind.google/models/veo/, 2025. Accessed: 2025-11-04. 3

[6] Google DeepMind. Flow: Ai-powered filmmaking with veo, imagen & gemini. https://blog.google/technology/ai/google-flow-veo-ai-filmmaking-tool/, 2025. Accessed: 2025-11-08. 3

[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. *arXiv preprint arXiv:2212.08051*, 2022. 1, 3

[8] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 1, 3

[9] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 3

[10] 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*, 2024. 3, 6

[11] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[12] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. Last revised 22 Jun 2022.

[13] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 5

[15] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. *arXiv preprint arXiv:2312.06725*, 2023. 3

[16] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17191–17202, 2025. 1, 3, 4

[17] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions, 2023. 3

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 8

[19] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Popa Tiberiu. Clip-mesh: Generating textured meshes from text using pretrained image-text models. *SIGGRAPH Asia 2022 Conference Papers*, 2022. 3

[20] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 3, 6

[21] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas Guibas, and Gordon. Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. In *arXiv*, 2024. 3

[22] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024. 6, 7

[23] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2041–2050, 2018. 1, 3

[24] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023. 3

[25] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler,

Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[26] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. 6, 7

[27] Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Yu Tong Tiffany Ling, Yuhan Huang, Sifan Liu, Mingyu Chen, Rushikesh Zawar, Xue Bai, Yilun Du, Chuang Gan, and Deva Ramanan. Towards understanding camera motions in any video. *arXiv preprint arXiv:2504.15376*, 2025. 6, 7

[28] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[29] 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, 2023.

[30] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3

[31] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, Yu Zhou, Deshan Sun, Deyu Zhou, Jian Zhou, Kaijun Tan, Kang An, Mei Chen, Wei Ji, Qiling Wu, Wen Sun, Xin Han, Yanan Wei, Zheng Ge, Aojie Li, Bin Wang, Bizhu Huang, Bo Wang, Brian Li, Changxing Miao, Chen Xu, Chenfei Wu, Chenguang Yu, Dapeng Shi, Dingyuan Hu, Enle Liu, Gang Yu, Ge Yang, Guanzhe Huang, Gulin Yan, Haiyang Feng, Hao Nie, Haonan Jia, Hanpeng Hu, Hanqi Chen, Haolong Yan, Heng Wang, Hongcheng Guo, Huilin Xiong, Huixin Xiong, Jiahao Gong, Jianchang Wu, Jiaoren Wu, Jie Wu, Jie Yang, Jiashuai Liu, Jiashuo Li, Jingyang Zhang, Junjing Guo, Junzhe Lin, Kaixiang Li, Lei Liu, Lei Xia, Liang Zhao, Liguo Tan, Liwen Huang, Liying Shi, Ming Li, Mingliang Li, Muhua Cheng, Na Wang, Qiaohui Chen, Qinglin He, Qiuyan Liang, Quan Sun, Ran Sun, Rui Wang, Shaoliang Pang, Shiliang Yang, Sitong Liu, Siqi Liu, Shuli Gao, Tiancheng Cao, Tianyu Wang, Weipeng Ming, Wenqing He, Xu Zhao, Xuelin Zhang, Xianfang Zeng, Xiaojia Liu, Xuan Yang, Yaqi Dai, Yanbo Yu, Yang Li, Yineng Deng, Yingming Wang, Yilei Wang, Yuanwei Lu, Yu Chen, Yu Luo, Yuchu Luo, Yuhe Yin, Yuheng Feng, Yuxiang Yang, Zecheng Tang, Zekai Zhang, Zidong Yang, Binxing Jiao, Jiansheng Chen, Jing Li, Shuchang Zhou, Xiangyu Zhang, Xinhao Zhang, Yibo Zhu, Heung-Yeung Shum, and Daxin Jiang. Step-video-t2v technical report: The practice, challenges, and future of video foundation model, 2025. 3

[32] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360 reconstruction of any object from a single image. In *CVPR*, 2023. 3

[33] OpenAI. Sora: Generating videos from text prompts. https://openai.com/research/sora, 2024. Accessed: 2025-11-03. 3

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

[35] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. 3

[36] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Lorenzo Sbordone, Aljaz Bozic Labiano, Maximilian Seitzer, David Novotny, Torsten Sattler, Gabriel Brostow, Andrea Vedaldi, and Christian Rupprecht. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10901–10911, 2021. 1, 3

[37] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1, 2, 3, 6, 7, 8

[38] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 1, 4, 5, 6

[39] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 501–518, 2016. 1, 4, 5, 6

[40] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023. 3

[41] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 3

[42] Uriel Singer, Shelly Sheynin, Adam Polyak, Thomas Hayes, Oran Gafni, Eliya Nachmani, et al. Make-a-video: Text-to-video generation without text-video data, 2022. 3

[43] Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. *arXiv preprint arXiv:2310.16818*, 2023. 3

[44] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024. 1, 2, 3, 6, 7, 8

[45] Tencent Hunyuan3D Team. Hunyuan3d 1.0: A unified framework for text-to-3d and image-to-3d generation, 2024. 2, 3, 7

[46] Tencent Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025. 2, 7

[47] Tencent Hunyuan3D Team. Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details, 2025.

[48] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, , Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 3

[49] I-Ting Tsai and Bharath Hariharan. 3D synthesis for architectural design. In *WACV*, pages 4799–4809, 2025. 1, 3, 4

[50] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. In *ECCV*, 2024. 1, 3

[51] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 5, 6

[52] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 8

[53] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 3

[54] World Labs. Marble: A multimodal world model. https://www.worldlabs.ai/blog/marble-world-model, 2025. Published November 12, 2025. Accessed March 1, 2026. 2, 3, 7

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

[56] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 3

[57] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 3

[58] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3

[59] Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. In *International Conference on Learning Representations*, 2025. 3

[60] Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 100–111, 2025. 1, 2, 3, 6, 7

[61] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 3, 6

[62] 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, 2023. 3

[63] David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. *arXiv preprint arXiv:2411.05003*, 2024. 3

[64] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 8

[65] Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan, Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim Hee Lee, and Lijuan Wang. Genxd: Generating any 3d and 4d scenes. In *ICLR*, 2025. 6