ConsID-Gen: View-Consistent and Identity-Preserving Image-to-Video Generation

Mingyang Wu[1], Ashirbad Mishra[2], Soumik Dey[2], Shuo Xing[1], Naveen Ravipati[2], Hansi Wu[2], Binbin Li[2], Zhengzhong Tu[1†]
[1]Texas A&M University [2]eBay Inc.

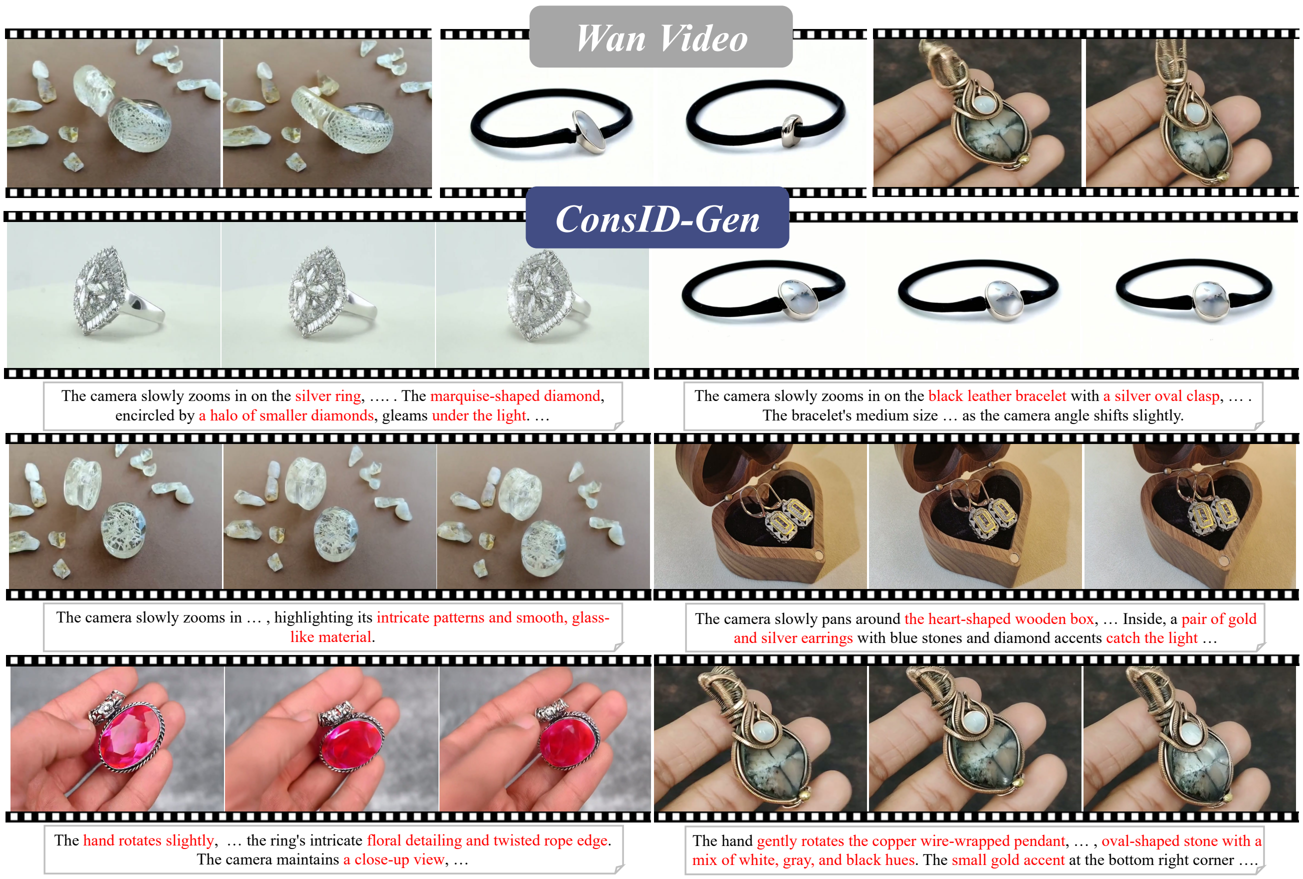

Figure 1. **Examples Synthesized by ConsID-Gen.** Given a textual instruction and reference image containing rigid objects (*i.e., rings, diamonds*), **ConsID-Gen** synthesizes realistic videos that faithfully preserve *object identity* and maintain *geometric consistency*. The initial row was generated by Wan [46] using the same prompt. Attributes highlighted in red denote object properties specified in the instruction.

Abstract

Image-to-Video generation (I2V) animates a static image into a temporally coherent video sequence following textual instructions, yet preserving fine-grained object identity under changing viewpoints remains a persistent challenge. Unlike text-to-video models, existing I2V pipelines often suffer from appearance drift and geometric distortion, artifacts we attribute to the sparsity of single-view 2D observations and weak cross-modal alignment. Here we address this problem from both data and model perspectives. First, we curate ***ConsIDVid****, a large-scale object-centric dataset built with a scalable pipeline for high-quality, temporally aligned videos, and establish* ***ConsIDVid-Bench****, where we present a novel benchmarking and evaluation framework for multi-view consistency using metrics sensitive to subtle geometric and appearance deviations. We further propose* ***ConsID-Gen****, a view-assisted I2V gener-*

† Corresponding author: tzz@tamu.edu

ation framework that augments the first frame with unposed auxiliary views and fuses semantic and structural cues via a dual-stream visual–geometric encoder as well as a text–visual connector, yielding unified conditioning for a Diffusion Transformer backbone. Experiments across ConsIDVid-Bench demonstrate that ConsID-Gen consistently outperforms in multiple metrics, with the best overall performance surpassing leading video generation models like Wan2.1 and HunyuanVideo, delivering superior identity fidelity and temporal coherence under challenging real-world scenarios. Our model and dataset are at https://myangwu.github.io/ConsID-Gen.

1. Introduction

Modern video generation models based on diffusion transformers (DiT) [19, 26, 38, 46, 59] can synthesize high-resolution, temporally coherent videos from text prompts, images, or both. This progress is beginning to reshape applications in advertising [13], entertainment [32], and digital content creation [9, 29, 44], where short, high-quality videos can now be synthesized rather than filmed [3]. Within this space, Image-to-Video (*I2V*) generation [20, 26, 46] is especially appealing: given a single reference image and a textual instruction, an *I2V* model animates a still frame into a temporally consistent, semantically rich video clip. This capability is particularly valuable for product-centric scenarios where a single catalog photo must be turned into multiple compelling videos or hand-held showcases while preserving the exact appearance [18, 23, 37].

Despite this promise, preserving fine-grained object identity under changing viewpoints remains challenging. Existing *I2V* systems [26, 41, 46] frequently exhibit **appearance drift** or **geometric distortion**: identity shifts, object shape warps, parts merge or disappear, and materials or textures subtly change across frames. As illustrated in Fig. 1, the glass gradually loses rigidity and appears to merge, violating the preservation of object-centric appearance. This failure to maintain instance-level consistency is a major roadblock for deploying *I2V* generation in real-world, high-stakes applications such as e-commerce, product advertising, and training videos.

Prior works [24] have explored explicit spatial supervision as a way to improve appearance drift; however, these methods are typically trained on small-scale curated datasets and evaluated on benchmarks [21] that emphasize semantic video quality rather than object identity. Consequently, they provide limited insight into preserving consistent object geometry and appearance, and they often fail to generalize to real-world product scenarios. More broadly, today's *I2V* ecosystem suffers from two systemic limitations: **1)** The available data is insufficient, where existing datasets rarely contain close-up, object-centric, multi-view videos that focus on identity continuity across space-time; **2)** Model architectures are not structurally equipped to preserve identity. For instance, we find that *T2V* models consistently outperform *I2V* models in identity-preserving generation (e.g., CogVideoX-1.5 [54]: $95.77\% \rightarrow 91.30\%$; Wan2.1 [46]: $96.72\% \rightarrow 91.84\%$; as shown in Table 1). We attribute this gap to a fundamental architectural issue where prevailing pipelines encode text and image inputs separately and only fuse them lately in the network (Fig. 2).

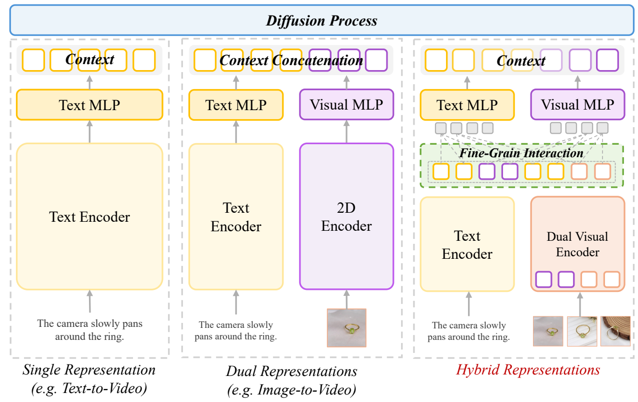

Figure 2. **Comparing Different Video Generation Paradigms.** Single-stream (*T2V*) uses only text tokens as context. Dual-stream (*I2V*) concatenates text and 2D visual tokens with limited interaction. Hybrid representations (*Ours*) pre-align text and visual tokens via fine-grained interaction before projection.

Motivated by these observations, here we systematically address the identity preservation issues in *I2V* generation from both the data and model perspectives. On the data side, we build **ConsIDVid**, a large-scale object-centric dataset constructed through a scalable pipeline (Sec. 3.2) that selects high-quality, temporally aligned videos of rigid objects, and establish **ConsIDVid-Bench**, a dedicated benchmark that reframes *I2V* evaluation as a multi-view consistency problem. Instead of relying on scene-level or frame-level scores, ConsIDVid-Bench incorporates geometry- and appearance-aware metrics explicitly designed to capture subtle distortions, shape inconsistencies, and within-object drift across viewpoints and time.

On the modeling side, we propose **ConsID-Gen**, a view-assisted *I2V* generation framework designed to explicitly encode appearance consistency and geometric stability. ConsID-Gen augments the single reference frame with unposed auxiliary views of the same object, allowing the model to recover richer structural cues to build a stable representation of object identity. These visual inputs are processed through a dual-stream visual–geometric encoder that captures both semantic appearance features and multi-view geometry. A multimodal text–visual connector then aligns these cues with textual motion instructions to produce unified conditioning for a diffusion-based video backbone. Our experimental results demonstrate that ConsID-Gen improves identity-preserving video generation. It delivers state-of-the-art identity fidelity on the proprietary subset and strong geometric preservation, achieving the lowest MEt3R (+30.2%) on the proprietary set and the lowest Chamfer Distance (+7.26%) on the public set.

Our main contributions are threefold: (i) A holistic I2V benchmark for identity preservation, with a diverse dataset and a novel multi-view evaluation suite; (ii) ConsID-Gen introduces unified representation before diffusion, with multi-view guidance and improved cross-modal alignment; (iii) Showing that ConsID-Gen outperforms open-source SOTAs in identity consistency and in human evaluation.

2. Related Works

2.1. Video Generation Models.

Text-Guided Video Generation. Driven by breakthroughs in visual generation and enhancement [36, 55, 60], the generation of videos from textual descriptions has garnered significant scholarly interest in the past year, spurred by advancements ranging from Sora [34] to MovieGen [38], Gen-4 [42], Sora2 [35], Kling [27, 44], Veo 3 [10], and others. In particular, Sora, which synthesizes a temporal Variational Autoencoder (VAE) with a DiT backbone, represents a critical achievement that has stimulated extensive architectural research within open-source communities. Prominent studies such as CogVideo [19] and CogVideoX [54] utilize a three-dimensional variational autoencoder coupled with an expert Transformer. HunyuanVideo [26] and Mochi 1 [15] implement asymmetric architectures and comprehensive attention mechanisms to improve the alignment between textual and video data. Wan2.1 [46] enhances the model capacity, while Wan2.2 [47] incorporates a sparse Mixture-of-Experts (MoE) approach, which delegates the diffusion process to specialized experts, thereby effectively capturing intricate motion dynamics.

Text-Image-Guided Video Generation. Despite recent advances, text-only prompting in *T2V* affords limited control over content and appearance. A promising alternative is to extend pretrained video generators by modifying their architecture to incorporate image conditions. Within this paradigm, DynamiCrafter [52] and Moonshot [56] inject image embeddings via cross-attention layers. ConsistI2V [41] applies spatial–temporal attention to the first frame coupled with a frequency-aware noise initialization strategy to enhance temporal coherence. SVD [6] and CogVideoX [54] extend *T2V* to *I2V* by channel-wise concatenation of conditional latents with noise. Wan2.1 [46] adopts mask-guided conditioning and injects image embeddings via decoupled cross-attention. Furthermore, such conditioning techniques are adapted for subject-to-video generation [12, 28, 30] and video editing [14, 25] to ensure identity preservation and precise modification.

2.2. Video Generation Evaluations.

Evaluations Metrics for Video Generation. With advances in generation, systematic evaluation of video quality has become increasingly crucial. Early works relied on distribution-based metrics such as Fréchet Video Distance (FVD) [45] and its variants [33], which, despite widespread use, offer limited correspondence to human perception. Several *T2V* evaluation benchmarks like VBench [21] provide structured, multi-dimensional evaluations focusing on fundamental visual attributes and prompt adherence, but their dependence on generic similarity models restricts fine-grained assessment. More recently, VLM-driven evaluators [5, 17, 53, 58] leverage inherent vision–language understanding to score intrinsic faithfulness; UVE [31] further unifies this paradigm by prompting a VLM to perform both single-video rating and pairwise comparison under aspect-specific guidelines. Nevertheless, VLM-based approaches remain sensitive to prompt design and model bias.

3. ConsIDVid Dataset & Benchmark Curation

Prior methods [24] were trained on small, minimally curated appearance-preserving datasets ($\sim$ 600 videos), which limits identity-consistent modeling. In response, we present ConsIDVid, a large-scale object-centric, identity-preserving video dataset curated via a scalable pipeline, together with an object-preserving benchmark for standardized evaluation of *I2V* models. We illustrate the data curation pipeline in Fig. 3 and present its stats in Fig. 4.

3.1. Video Collection

To mitigate data scarcity, we curated a candidate dataset from three sources: (i) existing object-centric datasets (*Co3D* [40], *OmniObject3D* [51], *Objectron* [1]); (ii) proprietary monocular videos; and (iii) synthetic videos. *Co3D* provides in-the-wild, object-centric videos across 50 MS-COCO categories; *OmniObject3D* comprises 6,000 objects spanning 190 categories with accompanying real-world videos; *Objectron* offers $\sim$15,000 short clips across nine categories collected in 10 countries.

To further investigate the aspect of realism and practicality, we experiment on over 80 hours of object-centric monocular UGC from public e-commerce platforms, where each clip primarily showcases a single product; many entries include unposed, multi-view images of the same item for instance-level supervision. We also synthesize object-centric sequences using a video generator conditioned on first and last keyframes, yielding temporally coherent clips suitable for identity-preserving training.

3.2. Data Curation Pipeline

Video Preprocessing. In the initial stage, we convert image sequences into standardized video clips and perform valid-

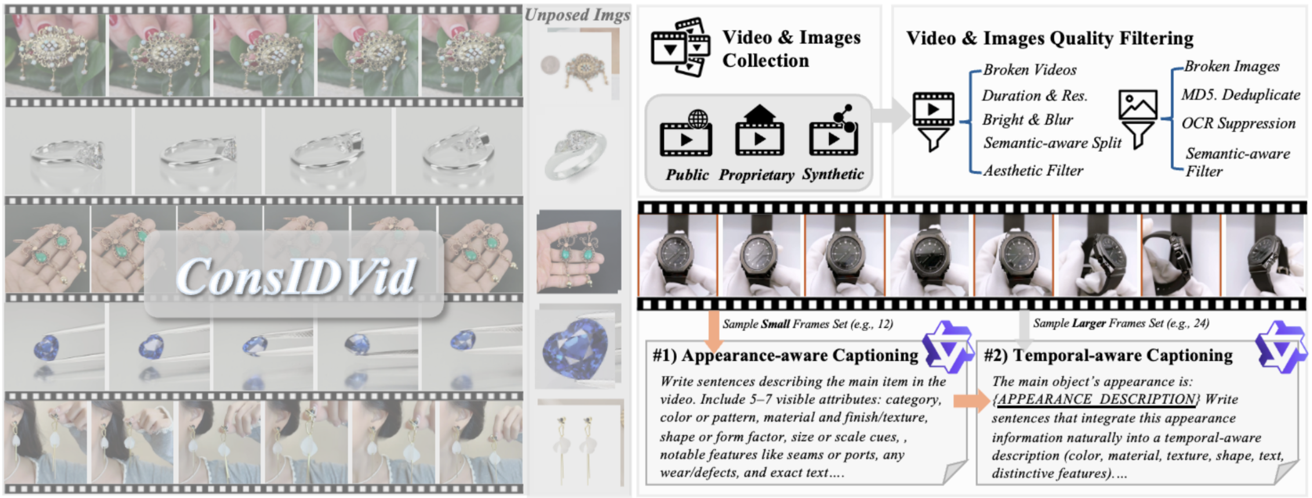

Figure 3. **Data Curation Pipeline.** We curate and synthesize videos from diverse sources, followed by an automated data curation pipeline to ensure visual and temporal quality. Video captions are produced by Qwen2.5-VL via a hierarchical captioning strategy.

ity checks using FFmpeg. These steps remove a substantial fraction of unusable media at the outset of the pipeline.

Video Quality Filtering. We implement a multi-faceted video quality filter to remove unsuitable content: **(i)** *Duration and resolution*: each clip contains at least 81 frames and meets a minimum resolution of 320p; **(ii)** *Brightness and blur*: we prune the bottom/top 5% of the luminance and Laplacian-variance distributions to remove under-/over-exposed and excessively blurred clips; **(iii)** *Semantics-aware splitting*: a two-stage procedure first detects shot boundaries and then stitches adjacent segments using frame-embedding similarity to correct over-segmentation, handle fade-in/out transitions and long uncut sequences, and reduce redundancy (cf. Panda-70M [8]); **(iv)** *Aesthetics*: we apply the LAION-5B [43] aesthetics predictor on 10 uniformly sampled frames to discard low-quality videos whose mean score is below 3.0. To ensure scalability, proprietary videos are clustered and processed in batches under the same pipeline.

Image Filtering. We curate proprietary unposed multi-view object images with cascade image filters: **(i)** *validity & exact deduplication*: eliminate corrupt files and MD5 duplicates; **(ii)** *OCR suppression*: discard images containing more than 30 detected characters; **(iii)** *semantics-aware outlier removal*: apply CLIP-based [39] reference matching to a curated outlier gallery and per-item embeddings clustered by DBSCAN, retaining the dominant cluster.

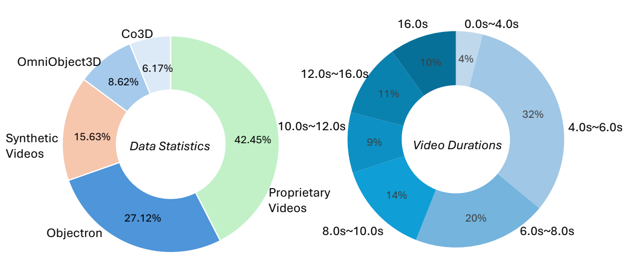

Figure 4. Statistics of video clips in ConsIDVid. The dataset includes diverse distributions of data source and video duration.

3.3. Hierarchical Video Captioning

The accuracy of captions is crucial for training video generation models. While Mixture-of-Multimodal-Experts captioning improves detail, its multi-model, multi-step inference is costly. By leveraging our object-centric dataset, we propose a two-stage hierarchical captioning protocol that produces fine-grained, temporally grounded video–text pairs with low computational overhead. We use Qwen2.5-VL [4] as the captioner and uniformly sample frames.

Stage 1: Appearance-aware Captioning. From a small frame subset (e.g., 12), produce a caption restricted to the primary object's visible attributes. The prompt restricts content to 5–7 concrete cues: category; color/pattern; material/finish/texture; shape/form factor; size/scale; notable parts; wear/defects; readable text/logos. Camera behavior, background context, and usage speculation are prohibited.

Stage 2: Temporal-aware Captioning. Conditioned on the Stage 1 caption and a larger frame set (e.g., 24), it generates a fluent caption that integrates *3–5* key appearance details with verified dynamics: camera motion, human–object interactions, and object motion.

3.4. Synthetic Video Generation

To enrich object- and viewpoint-level diversity, we synthesize videos from MVImgNet2.0 [16] multi-view imagery. For each object, we select two representative views as start and end frames and extend video generator [57] into an interpolation variant. This produces smooth temporal se-

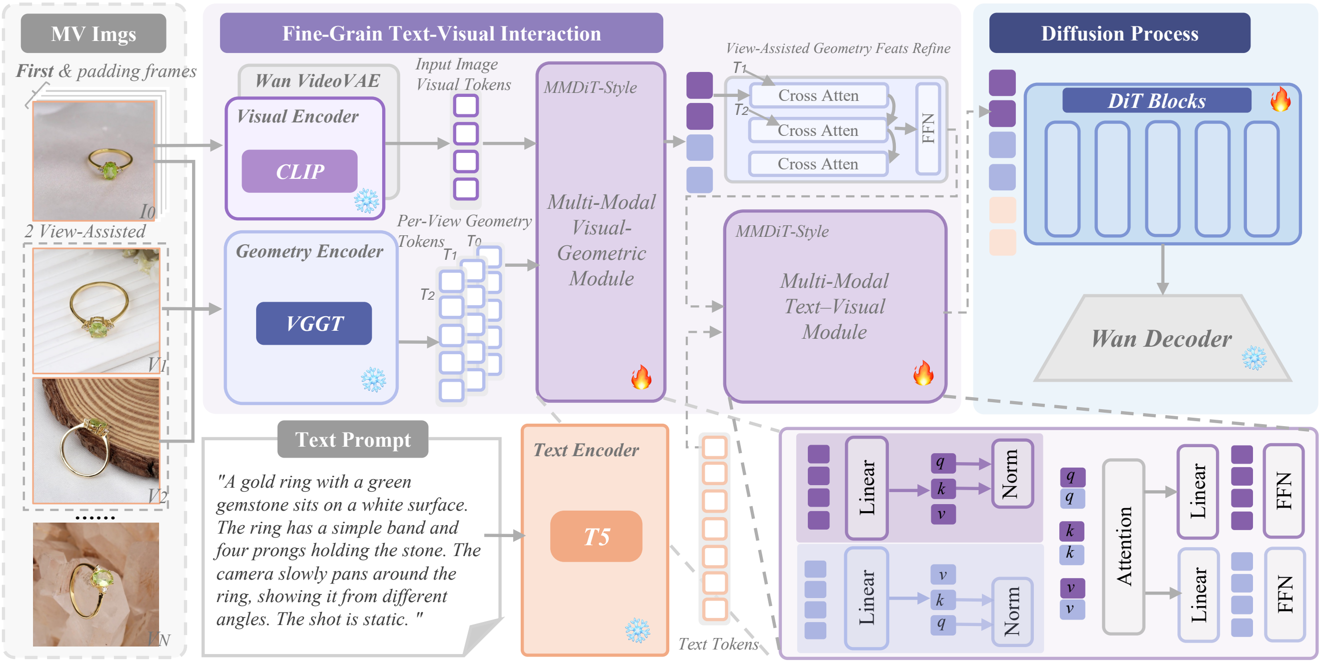

Figure 5. **Overview of ConsID-Gen.** The model takes as input the first frame, two uncalibrated images, and a text instruction. Our Dual-Visual Encoder combines a Visual Encoder and a Geometry Encoder to extract visual-appearance and geometric representations. A unified multimodal interaction projector then fuses these features with the prompt to generate conditioning tokens for the DiT backbone.

quences that preserve geometric consistency. Prompts are generated by Qwen2.5-VL [4], conditioned on the chosen start/end frames.

3.5. ConsIDVid-Bench

To evaluate beyond scene-level semantics and frame-level fidelity, we introduce ConsIDVid-Bench, an object-centric benchmark for assessing identity preservation in *I2V* generation. It aims to measure whether a generator maintains consistent object geometry and appearance under dynamic object or camera motion. By reformulating video evaluation as a multi-view consistency problem, ConsIDVid-Bench provides metrics sensitive to fine-grained appearance drift and geometric distortions over time.

Task Definition. Given an object-centric reference image I_{ref} and a driving prompt y, the model is required to generate a temporally coherent video that maintains the object's geometric and textural consistency while incorporating plausible object or camera motion.

Evaluation Metrics. We assess identity preservation with a comprehensive metric suite: *Chamfer Distance (CD)*, computed between 3D point sets reconstructed from the input and synthesized views, capturing global shape alignment and geometric stability over time; *MEt3R* [2], which applies DUSt3R [49] to obtain dense pairwise reconstructions and measures cross-view feature similarity after projection; *Video Similarity* (CLIP-based) to measure global realism and content consistency; and *Object Similarity*, as illustrated in the supplementary, which uses DINO-based features on segmented objects to assess identity preservation.

4. Method

In this section, we introduce ConsID-Gen, a view-assisted model for identity-preserving video diffusion. Given a first frame I_0, two uncalibrated auxiliary images $\mathcal{V} = \{V_1, V_2\}$ of the same object, and a text instruction y, our goal is to synthesize a video $\mathcal{X} = \{X_t\}_{t=1}^{T}$ that maintains the object's identity throughout time.

4.1. Model Architecture

As illustrated in Fig. 5, the model consists of a dual-visual encoder, a unified text–visual interaction projector, and a DiT backbone. We build on Wan2.1 and explore strategies that strengthen identity preservation by jointly exploiting appearance and geometric cues. Before detailing components, we motivate our dual-encoder formulation and the pre-alignment of visual and textual features.

What hinders visual-conditioned I2V? Current *I2V* pipelines derive dynamics from a first frame and a driving text prompt. The first frame is encoded by a pre-trained 2D encoder (e.g., CLIP [39]) into semantic condition features that are fused with textual tokens via simple concatenation and a lightweight connector. While effective for high-level recognition, such 2D features under-represent fine-grained structure. During temporal synthesis the model tends to hallucinate missing spatial details, which leads to cumulative

Table 1. Comparison of T2V and I2V models on identity preservation using automatic VBench metrics.

Method	Auto. Metrics			
	T2V-S	T2V-B	I2V-S	I2V-B
CogVideoX1.5-T2V [54]	95.77	96.31	–	–
CogVideoX1.5-I2V [54]	91.30	93.01	91.52	96.29
Wan2.1-T2V [46]	96.72	97.10	–	–
Wan2.1-I2V [46]	91.84	93.56	96.60	97.23

appearance drift and geometric distortions, particularly for rigid objects and under viewpoint changes. The core bottleneck is twofold: 2D observations are sparse and cross-modal alignment is weak, which together under-constrain the object geometry and its identity over time. Consistent with this diagnosis, Table 1 shows that single-stream *T2V* models, which do not require alignment between sparse visual and textual representations, tend to achieve stronger identity consistency.

We stabilize identity-preserving *I2V* by (i) anchoring object shape and appearance with unposed multi-view reference imagery and (ii) introducing a dual-path visual representation that couples a semantic 2D encoder E_{2D} with a geometry-aware encoder E_{geo} pre-trained to recover structural cues. The first frame, augmented with $\mathcal{V}$, provides local constraints on appearance and geometry, while the text prompt y supplies global control over scene dynamics. A dedicated connector g_ϕ aligns and fuses semantic and geometric features with textual tokens, which mitigates modality misalignment and yields unified conditioning tokens for the DiT backbone f_θ. We next detail these core components of the design.

4.2. Dual-Visual Encoder

Our model employs a dual-visual encoder composed of a 2D encoder E_{2D} and a geometry encoder E_{geo}.

2D Encoder. We use a CLIP-style image encoder E_{2D} to extract semantic appearance tokens from the first frame:

$$F_{\text{2D}} = E_{\text{2D}}(I_0), \qquad F_{\text{2D}} \in \mathbb{R}^{\lfloor H/p_{\text{2D}} \rfloor \times \lfloor W/p_{\text{2D}} \rfloor \times d_{\text{2D}}},$$

where $H \times W$ is the image resolution, p_{2D} is the patch size, and d_{2D} is the feature dimension. These tokens provide high-level appearance priors for subsequent fusion.

Geometric Encoder. To complement semantic cues with geometry structure, we use VGGT [48] as the geometry backbone E_{geo}. Given unposed auxiliary views $\tilde{\mathcal{V}} = \{I_0, V_1, V_2\}$, each image is patchified and processed with alternating frame-wise and global self-attention to get dense geometry-aware tokens:

$$F_{\text{geo}} = E_{\text{geo}}(\tilde{\mathcal{V}}), \qquad F_{\text{geo}} \in \mathbb{R}^{3 \times \lfloor H/p_{\text{geo}} \rfloor \times \lfloor W/p_{\text{geo}} \rfloor \times d_{\text{geo}}},$$

where p_{geo} and d_{geo} denote the patch size and feature width of the geometry encoder, respectively. We retain the dense structural tokens for fusion with F_{2D}.

4.3. Multi-visual-text interaction

After extracting semantic tokens F_{2D} and geometry-aware tokens F_{geo}, we introduce a connector g_ϕ bridging modality gaps. It comprises a *Multi-Modal Visual–Geometric Module* (MVGM) that injects structural cues from F_{geo} into appearance tokens F_{2D}, and a *Multi-Modal Text–Visual Module* (MTVM) that aligns the fused visual representation with text T for fine-grained interaction.

Multi-Modal Visual–Geometric Module. Motivated by the dual-stream architecture of the Multimodal Diffusion Transformer (MMDiT) [11, 50], which enables effective alignment between visual and textual modalities, we extend this paradigm to the visual–geometric domain to achieve joint modeling of semantic appearance and 3D structure. Specifically, the MVGM fuses appearance tokens F_{2D} with geometry-aware tokens F_{geo} extracted from the first frame I_0 through a dual-stream attention mechanism, enabling bidirectional interaction between semantic and structural cues. Furthermore, geometric features from the two auxiliary views $\mathcal{V}$ are integrated via cross-attention with the MVGM outputs, injecting multi-view structural priors that reinforce spatial and geometric consistency.

Multi-Modal Text–Visual Module. Building on the fused visual–geometric representation, the MTVM further aligns vision and language within a dual-stream attention mechanism. In this stage, textual features dynamically modulate the visual stream, while visual representations provide complementary cues to the text.

5. Experiments

In this section, we conduct comprehensive qualitative and quantitative evaluations of popular *I2V* generators and our proposed ConsID-Gen to assess their capability for identity-preserving video generation.

5.1. Experimental Settings

Implementation Details. We build our model upon Wan2.1-Fun-1.3B-InP [46], which generates 81-frame video clips at a 832×480 resolution. For training, we employ the AdamW optimizer with a learning rate of 10^{-4}. We use a per-GPU batch size of 1 with gradient accumulation over 4 steps (effective batch size 4). The model is trained for 33K steps. All experiments are conducted on NVIDIA A100 (80GB) GPUs. During inference, we utilize 50 sampling steps and a classifier-free guidance (CFG) scale of 5.

Evaluation Metrics. To evaluate identity consistency and temporally coherent dynamics, we adopt established metrics from the VBench-I2V [22] suite: Subject Consistency, Background Consistency, Motion Smoothness, and Temporal Flickering. To further assess the fidelity of identity preservation, we employ geometry-aware metrics, including MEt3R, Chamfer Distance, and Video Similarity.

Table 2. Quantitative comparison on the proprietary subset of ConsIDVid-Bench. We evaluate model performance using VBench-I2V suite, Video Similarity, Object Similarity, Chamfer Distance, and MEt3R metrics. Best and second-best scores are highlighted

Method	I2V Subject	I2V Background	Subject Consistency	Background Consistency	Motion Smoothness	Temporal Flickering	Video Similarity	Object Similarity	Chamfer Distance	MEt3R
Wan2.1-1.3B [46]	96.22	97.12	91.03	94.57	99.33	98.84	87.15	66.9	0.1064	0.1401
SkyReelv2 [7]	94.03	95.21	85.61	92.04	98.71	97.42	85.33	59.5	0.1107	0.2177
ConsistI2V [41]	94.93	93.42	91.41	94.07	98.25	96.72	82.48	62.0	0.1429	0.1614
Wan2.2-5B [47]	96.85	97.57	91.99	94.82	98.93	98.10	88.69	68.6	0.0921	0.1826
CogVideoX1.5-5B [54]	91.69	96.31	90.03	93.14	98.47	97.70	84.14	60.1	0.1194	0.1518
HunyuanVideo [26]	95.24	96.15	90.40	93.27	98.38	97.55	86.59	64.3	0.1017	0.2270
Wan2.1-14B [46]	96.14	96.86	90.37	94.14	98.89	98.05	87.33	67.9	0.0866	0.1572
ConsID-Gen	98.31	98.66	95.30	96.10	99.52	99.24	88.65	69.2	0.0996	0.0978

Table 3. Quantitative comparison on the public subset of ConsIDVid-Bench. We evaluate model performance using VBench-I2V suite, Video Similarity, Object Similarity, Chamfer Distance, and MEt3R metrics. Best and second-best scores are highlighted.

Method	I2V Subject	I2V Background	Subject Consistency	Background Consistency	Motion Smoothness	Temporal Flickering	Video Similarity	Object Similarity	Chamfer Distance	MEt3R
Wan2.1-1.3B [46]	97.34	97.71	92.86	94.09	99.32	98.48	83.37	69.1	0.1503	0.1324
SkyReelv2 [7]	96.59	97.00	91.67	93.23	99.16	97.92	84.80	68.0	0.1500	0.1526
ConsistI2V [41]	95.38	92.25	91.98	93.32	97.67	95.48	79.22	62.4	0.1700	0.1601
Wan2.2-5B [47]	98.47	98.64	94.02	94.39	98.85	97.47	84.81	71.6	0.1386	0.1591
CogVideoX1.5-5B [54]	91.26	96.14	90.58	92.05	98.91	97.84	80.26	61.5	0.1589	0.1409
HunyuanVideo [26]	96.66	96.88	92.16	93.20	98.36	97.20	83.00	67.4	0.1377	0.2126
Wan2.1-14B [46]	98.29	98.49	94.90	94.74	99.15	98.16	84.45	72.2	0.1322	0.0961
ConsID-Gen	98.14	98.49	94.81	95.19	99.22	98.33	84.95	71.8	0.1277	0.1321

Evaluation Datasets. We conduct quantitative evaluation using our proposed **ConsIDVid-Bench**. As detailed in Section 3.5, this benchmark is specifically designed to assess identity preservation and comprises two subsets: the proprietary subset (241 videos), which consists of the videos depicting the product of popular e-commerce listings, and the public subset (370 videos), built from existing object-centric datasets and synthetic videos.

5.2. Quantitative Evaluations

Results on the proprietary Subset. Table 2 presents the evaluation results on our proprietary subset. ConsID-Gen achieves state-of-the-art (SOTA) performance across the VBench-I2V suite. Compared to the strong Wan2.2 [47], ConsID-Gen demonstrates higher identity fidelity, achieving a 3.6% higher score in Subject Consistency. Notably, ConsID-Gen yields a substantially lower score in the geometry-aware MEt3R [2] metric, demonstrating superior multi-view consistency. While Wan2.2 leads slightly in Video Similarity and Wan2.1-14B achieves the best Chamfer Distance, ConsID-Gen remains highly competitive.

Results on the public Subset of ConsIDVid-Bench. As shown in Table 3, ConsID-Gen demonstrates highly competitive performance on the public subset. Notably, ConsID-Gen achieves superior performance in geometric and fidelity metrics, achieving the top scores for both Chamfer Distance and Video Similarity. However, we observe that ConsID-Gen yields suboptimal results for I2V Subject and I2V Background compared to other methods [47]. We partially attribute this to a qualitative artifact: when the input contains distracting structures (*e.g., grid paper*), our generated videos occasionally suffer from degradation or collapse, a phenomenon that is analyzed in detail in the supplementary.

5.3. Qualitative Evaluations

Figure 6 presents the qualitative comparisons between ConsID-Gen and existing methods. As illustrated, ConsID-Gen generates videos with strong identity preservation, avoiding issues of appearance drift or geometric collapse. In contrast, videos produced by existing popular methods exhibit noticeable inconsistencies and temporal artifacts. For instance, in the ”gemstone” example (left), methods [7, 26] suffer from object jitter and severe scene changes. In the ”ring” example (right), other methods [54] fail to maintain geometric integrity, causing the subject to visibly deform.

5.4. Ablation Studies

Effect of Key Components. We conduct ablation studies to validate our key architectural components. Due to computational resource constraints, these ablated models were finetuned on 50% of the training data and evaluated on a randomly sampled 60-video subset. The results in Table 4 reveal a clear progression: we find that the geometry encoder (“+ Geo Enc.”) in isolation provides no significant benefits over the baseline. However, further adding multi-

Figure 6. **Qualitative comparison with popular *I2V* methods.** ConsID-Gen maintains the identity and geometry of objects in challenging scenarios. Compared to existing methods, our results demonstrate superior geometric fidelity and temporal coherence.

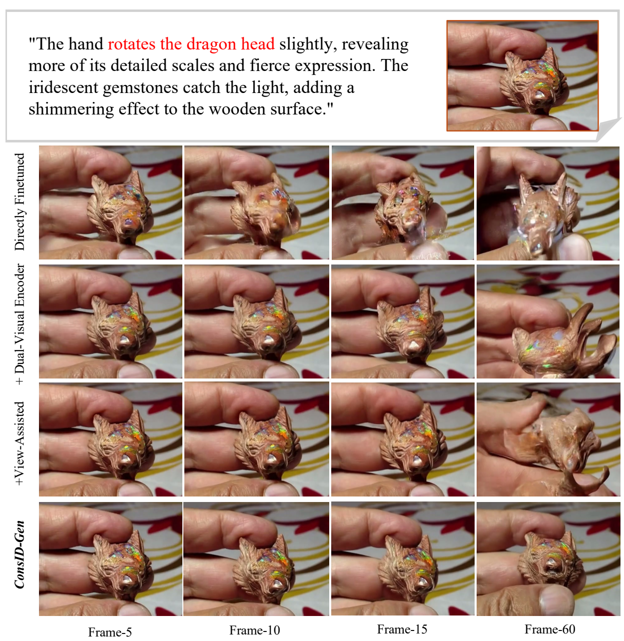

Figure 7. **Qualitative results of the ablation study.** ConsID-Gen maintains consistent identity across longer temporal spans.

ple unposed, view-assisted images ("+ View-Asst.") yields a clear improvement. These findings are further supported by what is illustrated in Figure 7, where we observe that direct finetuning Wan2.1 still leads to noticeable identity shift in the early frames of the generated videos. Incorporating the geometry encoder and view-assisted images is able to alleviate this issue to some extent. However, our full ConsID-Gen model that fuses text and visual cues ensures long-range identity stability.

Table 4. Quantitative ablation of key components. We evaluate model performance using VBench metrics and Video Similarity.

Method	I2V-Subj	I2V-Back	Subj-Cons.	Back-Cons.	Video-Sim.
Baseline	96.30	97.16	90.83	94.97	87.75
+ Geo Enc.	96.29	97.37	89.65	93.44	86.19
+ View-Asst.	96.97	97.85	91.87	94.33	87.35
ConsID-Gen	**98.48**	**98.85**	**95.13**	**96.20**	**88.25**

6. Conclusion

In this paper, we discuss the preservation of identity in *I2V* from both the data and the model perspectives. On the data side, we curate ConsIDVid, a large-scale object-centric dataset, and introduce ConsIDVid-Bench, which reframes evaluation as multi-view consistency to capture precise geometric and appearance drift. On the model side, we propose ConsID-Gen, a View-Assisted Video Generation framework that augments the first frame with unposed auxiliary views and performs fine-grained pre-alignment via dual-stream visual–geometric fusion and a text–visual connector. Across proprietary and public subsets of ConsIDVid-Bench, our model consistently exceeds baselines with stronger identity fidelity under challenging real-world scenes.

Acknowledgments

We thank Siyuan Yang for the discussions. We also thank Yixin Chen, Zhe Dong, Kuan-Ru Huang, Yanjia Huang, Zhaoming Xu, Jongze Yu, Zihao Zhu, and Yushen Zuo for their assistance with the user studies.

References

[1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7822–7831, 2021. 3

[2] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6034–6044, 2025. 5, 7

[3] 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. 2

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 4, 5

[5] Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025. 3

[6] 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*, 2023. 3

[7] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025. 7

[8] 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, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 4

[9] Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025. 2

[10] DeepMind. Veo 3. https://deepmind.google/models/veo, 2025. 3

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 6

[12] Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025. 3

[13] Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-s2v: Audio-driven cinematic video generation, 2025. 2

[14] Xiangbo Gao, Renjie Li, Xinghao Chen, Yuheng Wu, Suofei Feng, Qing Yin, and Zhengzhong Tu. Pisco: Precise video instance insertion with sparse control. *arXiv preprint arXiv:2602.08277*, 2026. 3

[15] Genmo Team. Mochi 1: A new sota in open text-to-video. https://www.genmo.ai/blog/mochi-1-a-new-sota-in-open-text-to-video, 2025. 3

[16] Xiaoguang Han, Yushuang Wu, Luyue Shi, Haolin Liu, Hongjie Liao, Lingteng Qiu, Weihao Yuan, Xiaodong Gu, Zilong Dong, and Shuguang Cui. Mvimgnet2. 0: A larger-scale dataset of multi-view images. *arXiv preprint arXiv:2412.01430*, 2024. 4

[17] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024. 3

[18] HeyGen Team. Heygen – ai spokesperson video creator. https://app.heygen.com/home, 2025. 2

[19] 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. 2, 3

[20] Haoyang Huang, Guoqing Ma, Nan Duan, Xing Chen, Changyi Wan, Ranchen Ming, Tianyu Wang, Bo Wang, Zhiying Lu, Aojie Li, et al. Step-video-ti2v technical report: A state-of-the-art text-driven image-to-video generation model. *arXiv preprint arXiv:2503.11251*, 2025. 2

[21] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 2, 3

[22] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si,

Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024. 6
[23] Invideo AI Team. Invideo ai. *Invideo*, 2025. 2
[24] Hyeonho Jeong, Chun-Hao P Huang, Jong Chul Ye, Niloy J Mitra, and Duygu Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7276–7287, 2025. 2, 3
[25] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 3
[26] 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. 2, 3, 7
[27] Kuaishou. Kling. `https://klingai.kuaishou.com`, 2025. 3
[28] Debang Li, Zhengcong Fei, Tuanhui Li, Yikun Dou, Zheng Chen, Jiangping Yang, Mingyuan Fan, Jingtao Xu, Jiahua Wang, Baoxuan Gu, et al. Skyreels-v3 technique report. *arXiv preprint arXiv:2601.17323*, 2026. 3
[29] Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, Zhengzhong Tu, et al. 4k4dgen: Panoramic 4d generation at 4k resolution. *arXiv preprint arXiv:2406.13527*, 2024. 2
[30] Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*, 2025. 3
[31] Yuanxin Liu, Rui Zhu, Shuhuai Ren, Jiacong Wang, Haoyuan Guo, Xu Sun, and Lu Jiang. Uve: Are mllms unified evaluators for ai-generated videos? *arXiv preprint arXiv:2503.09949*, 2025. 3
[32] Luma AI Team. Luma ai. `https://lumalabs.ai/`, 2025. 2
[33] Ge Ya Luo, Gian Mario Favero, Zhi Hao Luo, Alexia Jolicoeur-Martineau, and Christopher Pal. Beyond fvd: Enhanced evaluation metrics for video generation quality. *arXiv preprint arXiv:2410.05203*, 2024. 3
[34] OpenAI. Video generation models as world simulators. `https://openai.com/index/video-generation-models-as-world-simulators`, 2024. 3
[35] OpenAI. Sora 2 is here. `https://openai.com/index/sora-2/`, 2025. 3
[36] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 3
[37] Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, et al. Open-sora 2.0: Training a commercial-level video generation model in 200k. *arXiv preprint arXiv:2503.09642*, 2025. 2
[38] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2025. 2, 3
[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 4, 5
[40] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 3
[41] Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024. 2, 3, 7
[42] Runway. Runway gen-4. `https://runwayml.com/research/introducing-runway-gen-4`, 2025. 3
[43] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 4
[44] Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*, 2025. 2, 3
[45] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 3
[46] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 2, 3, 6, 7
[47] Wan Team. Wan2.2: More powerful, more beautiful. `https://wan.video/blog/wan2.2`, 2025. 3, 7
[48] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 6
[49] 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*, pages 20697–20709, 2024. 5
[50] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 6

[51] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 3

[52] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024. 3

[53] Shuo Xing, Soumik Dey, Mingyang Wu, Ashirbad Mishra, Naveen Ravipati, Binbin Li, Hansi Wu, and Zhengzhong Tu. Q-router: Agentic video quality assessment with expert model routing and artifact localization. *arXiv preprint arXiv:2510.08789*, 2025. 3

[54] 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. 2, 3, 6, 7

[55] Jiongze Yu, Xiangbo Gao, Pooja Verlani, Akshay Gadde, Yilin Wang, Balu Adsumilli, and Zhengzhong Tu. Sparkvsr: Interactive video super-resolution via sparse keyframe propagation, 2026. 3

[56] David Junhao Zhang, Dongxu Li, Hung Le, Mike Zheng Shou, Caiming Xiong, and Doyen Sahoo. Moonshot: Towards controllable video generation and editing with multimodal conditions. *arXiv preprint arXiv:2401.01827*, 2024. 3

[57] Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2(3):5, 2025. 4

[58] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 3

[59] Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. Allegro: Open the black box of commercial-level video generation model. *arXiv preprint arXiv:2410.15458*, 2024. 2

[60] Yushen Zuo, Qi Zheng, Mingyang Wu, Xinrui Jiang, Renjie Li, Jian Wang, Yide Zhang, Gengchen Mai, Lihong V Wang, James Zou, et al. 4kagent: agentic any image to 4k super-resolution. *arXiv preprint arXiv:2507.07105*, 2025. 3