OmniVDiff: Omni Controllable Video Diffusion for Generation and Understanding

Dianbing Xi[1, 2, *], **Jiepeng Wang**[2, *, †], **Yuanzhi Liang**[2], **Xi Qiu**[2], **Yuchi Huo**[1], **Rui Wang**[1‡], **Chi Zhang**[2‡], **Xuelong Li**[2‡]

[1]State Key Laboratory of CAD&CG, Zhejiang University
[2]Institute of Artificial Intelligence, China Telecom

Abstract

In this paper, we propose a novel framework for controllable video diffusion, *OmniVDiff*, aiming to synthesize and comprehend multiple video visual content in a single diffusion model. To achieve this, *OmniVDiff* treats all video visual modalities in the color space to learn a joint distribution, while employing an adaptive control strategy that dynamically adjusts the role of each visual modality during the diffusion process, either as a generation modality or a conditioning modality. Our framework supports three key capabilities: (1) Text-conditioned video generation, where all modalities are jointly synthesized from a textual prompt; (2) Video understanding, where structural modalities are predicted from rgb inputs in a coherent manner; and (3) X-conditioned video generation, where video synthesis is guided by fine-grained inputs such as depth, canny and segmentation. Extensive experiments demonstrate that *OmniVDiff* achieves state-of-the-art performance in video generation tasks and competitive results in video understanding. Its flexibility and scalability make it well-suited for downstream applications such as video-to-video translation, modality adaptation for visual tasks, and scene reconstruction.

Project page — https://tele-ai.github.io/OmniVDiff

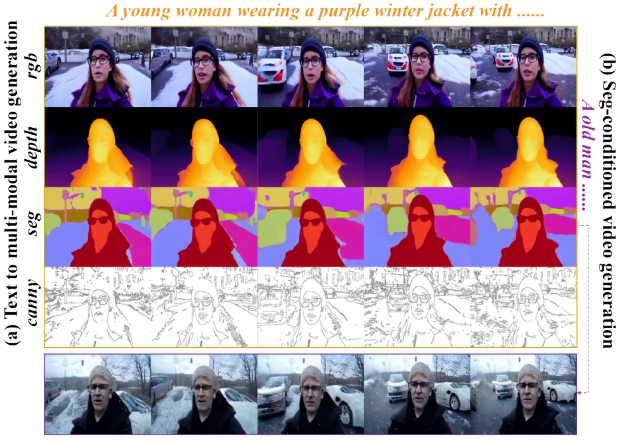

Figure 1: Omni controllable video generation and understanding. Given a text prompt, (a) *OmniVDiff* generates high-quality rgb videos while simultaneously producing aligned multi-modal visual understanding outputs (i.e., depth, segmentation and canny). Additionally, (b) *OmniVDiff* supports X-conditioned video generation within a unified framework, such as seg-conditioned video generation.

Introduction

Diffusion models have achieved remarkable progress in image (Rombach et al. 2022) and video generation (Blattmann et al. 2023; Kong et al. 2024; Yang et al. 2024b), demonstrating strong controllability and generalization through large-scale training. For controllable video generation, models typically employ conditions such as depth (Guo et al. 2024; Liu et al. 2024; Xing et al. 2024), segmentation (Zhao et al. 2023; Khachatryan et al. 2023; Hu et al. 2025), or canny edges (Lv et al. 2024) to guide the diffusion process. By fine-tuning pretrained text-to-video (T2V) models (Blattmann et al. 2023; Yang et al. 2024b), these approaches achieve high-quality controllable generation. However, most existing methods rely on task-specific fine-tuning and external expert models to obtain conditional modalities, which limits scalability and increases computational cost. Recent works further explore joint multi-modal generation (Zhai et al. 2024; Chefer et al. 2025; Byung-Ki et al. 2025; Wang et al. 2025; Jiang et al. 2025; Huang et al. 2025), yet they primarily focus on joint synthesis and lack support for generative understanding or conditional control. Overall, while video diffusion models show strong potential, their limited adaptability remains a key obstacle to developing a unified and efficient framework for diverse video-related tasks.

Recently, several concurrent studies in the image domain explored unifying multiple tasks within a single diffusion framework, by treating image-level tasks as a sequence of image views (Le et al. 2024; Chen et al. 2024b; Wang et al. 2025; Zhao et al. 2025) (analogous to video generation). For example, the depth-conditioned generation can be regarded as a two-view (depth and rgb) diffusion task. While this approach has been effective for image-based tasks, extending it to video generation presents significant challenges. Unlike images, videos introduce an additional temporal dimen-

[*]These authors contributed equally.
[†]These authors served as project leads.
[‡]These authors are the corresponding authors.

sion. Treating modalities as distinct video sequences would significantly increase the token length and computation cost in the transformer-based diffusion process, especially considering the quadratic computational complexity in the attention mechanism (Vaswani et al. 2017). The challenge of extending such approaches into a unified video diffusion framework that can handle both conditioned and unconditioned generation remains largely unexplored.

In this work, we propose *OmniVDiff*, a unified framework for controllable video generation. Our approach comprises two key components: (1) a multi-modal video diffusion architecture and (2) an adaptive modality control strategy, jointly enabling efficient handling of diverse visual modalities for both generation and understanding. (1) In the diffusion network, we extend the input noise dimensionality to match the number of modalities, allowing the model to process multiple visual inputs seamlessly. Distinct projection heads generate modality-specific outputs while preserving a unified framework. (2) To enhance adaptability, we introduce a flexible control strategy that dynamically assigns each modality as generative or conditional. For generative modalities, inputs are blended with noise, while conditional ones retain their original signals. This distinction is reinforced through learnable modality-specific embeddings. Through this design, our method achieves fine-grained control across modalities, providing a unified and adaptable framework for video generation and understanding tasks.

To this end, we focus on four representative visual modalities: rgb, depth, segmentation, and canny. To train our unified diffusion model, we construct a paired multi-modal dataset by filtering a subset of videos from Koala-36M (Wang et al. 2024a) and applying expert models to generate high-quality pseudo-labels for each modality.

We evaluate our approach on a broad range of tasks, including text-to-video generation, X-conditioned video generation, and multi-modal video understanding, and further assess its generalization to downstream tasks such as video-to-video style transfer and super-resolution. Extensive experiments demonstrate the robustness and versatility of our unified framework.

In summary, our main contributions are as follows:

- A unified controllable diffusion framework, supporting text-conditioned video generation, controllable generation with structural modalities (depth, canny, segmentation), and video understanding within a single model.
- An adaptive modality control strategy that dynamically determines the role of each modality (generation or conditioning), enabling fine-grained control and enhancing task adaptability.
- Comprehensive evaluation across generation and understanding tasks, demonstrating controllable video generation without expert dependency, and generalization to applications such as style transfer and super-resolution.

Related Works

Text-to-video Diffusion

Text-to-video (T2V) diffusion models have made significant progress in generating realistic and temporally consistent videos from text prompts (Kong et al. 2024; Polyak et al. 2025). SVD (Blattmann et al. 2023), VDM (Ho et al. 2022) and following works (Hong et al. 2022) explore extending image diffusion models (Rombach et al. 2022) for video synthesis with spatial and temporal attention (Chen et al. 2024a; Feng et al. 2024). Recent methods also introduce 3D Variational Autoencoder (VAE) to compress videos across spatial and temporal dimensions, improving compression efficiency and video quality(Yang et al. 2024b; Kong et al. 2024; Wan et al. 2025). However, these approaches primarily focus on text-conditioned video generation and lack fine-grained control over video attributes. Tasks such as depth-guided or segmentation-conditioned video generation remain challenging, as text-to-video diffusion models do not explicitly support these controls. Meanwhile, all these methods mainly focus on the rgb modality output, without considering the generative capability of other visual modalities.

Controllable Video Diffusion

To address controllable video generation, many methods try to introduce additional conditioning signals to guide the diffusion process. Depth maps can provide accurate geometric and structural information, ensuring realistic spatial consistency across frames (Xing et al. 2024; Chen et al. 2023; Zhang et al. 2023). Pose conditioning ensures accurate human motion synthesis by constraining body articulation and joint movements(Gan et al. 2025; Hu et al. 2025). Optical flow constrains motion trajectories by capturing temporal coherence and movement patterns, enhancing dynamic realism (Liu et al. 2024). However, these existing methods face two major challenges: (1) Fine-tuning for each task: incorporating new control signals typically requires task-specific fine-tuning on large-scale diffusion architectures, making these models computationally expensive and difficult to scale across diverse control modalities. (2) Dependency on external expert models: most approaches rely on pre-extracted conditioning signals from external expert models. For example, in depth-conditioned video generation, a separate depth estimation model is first applied to a reference video, and the estimated depth is then fed into a distinct video diffusion model for generation. This results in a multi-step, non-end-to-end pipeline where each component is trained separately, potentially causing inconsistencies across models and complex operations.

Unified Multi-modal Video Generation

Some efforts have attempted to unify multi-modal generation within a single diffusion model (Zhai et al. 2024; Wang et al. 2024b; Chefer et al. 2025; Byung-Ki et al. 2025; Wang et al. 2025; Jiang et al. 2025; Huang et al. 2025). VideoJAM (Chefer et al. 2025) jointly forecasts rgb frames and optical flow. However, such approaches primarily focus on joint modeling of two modalities, offering limited support for conditional generation and understanding. In addition, DiffusionRenderer (Liang et al. 2025) addresses both inverse and forward rendering, but relies on two separate models, where the forward rendering process is treated as conditional generation. Similarly, UDPDiff (Yang et al. 2025) supports joint generation of RGB with either depth or segmentation,

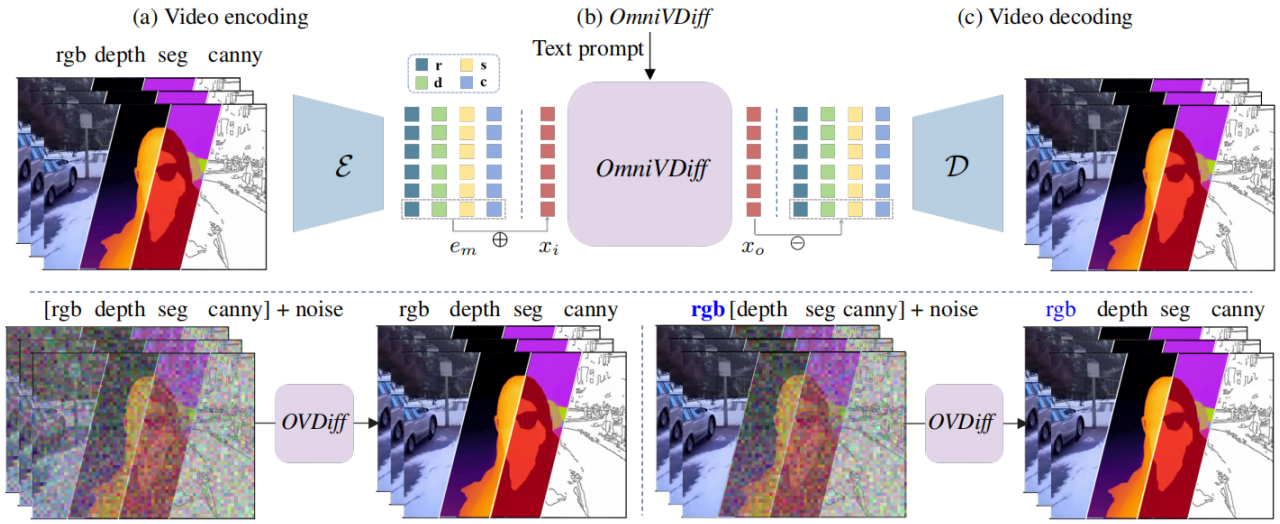

Figure 2: Method overview. (a) Given a video with four paired modalities, we first encode them into latents using a shared 3D-VAE encoder. (b) The latents are concatenated along the channel dimension and corrupted with noise for video diffusion, after which the denoised latents are decoded into their respective modalities via modality-specific decoding heads. (c) Each modality can then be reconstructed into color space by the 3D-VAE decoder. During inference, our model enables various tasks by dynamically assigning each modality either a generative or conditional role: (d) Text-to-video generation, where all modalities are denoised from pure noise; (e) X-conditioned generation, where the condition modality X is provided and the remaining modalities are denoised from noise. If X is the RGB modality, the model performs generative understanding.

yet it cannot synthesize all three modalities simultaneously or perform video understanding within a unified framework. Concurrently, Aether (Team et al. 2025) proposes a unified framework that supports both video understanding and joint multi-modal generation across rgb, depth, and camera pose. However, its primary focus lies in geometric world modeling, while generalization to a wider range of modalities like semantic masks and enabling flexible modality-conditioned controllable generation and understanding remains largely under-explored. In this paper, our method addresses these challenges by introducing a unified framework that allows fine-grained adaptive modality control. Unlike prior works, we do not require separate fine-tuning for each control modality and eliminate the reliance on external expert models by integrating multi-modal understanding and generation into a single pipeline. This enables efficient, end-to-end controllable video synthesis, improving scalability and coherence across video generation tasks.

In this work, we address these challenges by introducing a unified framework that enables fine-grained, adaptive modality control. Unlike prior approaches, our method eliminates the need for per-modality fine-tuning and external expert models, integrating multi-modal understanding and generation into a single end-to-end pipeline. This design facilitates efficient and coherent controllable video synthesis, improving both scalability and consistency across tasks.

Method

In this section, we introduce *OmniVDiff*, a unified framework for video generation and understanding, extending video diffusion models to support multi-modal video synthesis and analysis. We begin with a preliminary introduction to video diffusion models. Then, we detail our network design and adaptive control strategy, which enable seamless handling of text-to-video generation, modality-conditioned video generation, and multi-modal video understanding. Finally, we describe our training strategy. Figure 2 provides an overview of our framework.

Preliminary

Video diffusion models generate videos by progressively refining noisy inputs through a denoising process, following a learned data distribution. CogVideoX (Yang et al. 2024b), one of the state-of-the-art text-to-video diffusion models, incorporates a 3D Variational Autoencoder (3D-VAE) to efficiently compress video data along both spatial and temporal dimensions, significantly reducing computational costs while preserving motion consistency.

Given an input video $V \in \mathbb{R}^{f \times h \times w \times c}$, where f, h, w, c denote the number of frames, height, width, and channels, respectively, the 3D-VAE encoder downsamples it using a spatiotemporal downsampling factor of (8,8,4) along the height, width, and frame dimensions: $F = \frac{f}{4}$, $H = \frac{h}{8}$, $W = \frac{w}{8}$. This process captures both appearance and motion features while significantly reducing the memory and computational requirements of the diffusion process. The video diffusion model operates in this latent space, iteratively denoising $\mathbf{x}_t$ through a learned reverse process. The training objective minimizes the mean squared error (MSE) loss for noise prediction:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{\mathbf{x}_0, t, \epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2\right] \quad (1)$$

where ϵ_θ is the noise prediction model, $\mathbf{x}_t$ is the noisy latent at timestep t, and ϵ is the added noise.

Omni Video Diffusion

Multi-modal video diffusion architecture To achieve omni-controllable video diffusion, we design a novel video diffusion architecture that learns a joint distribution over multiple visual modalities. Building upon the pretrained text-to-video diffusion model CogVideoX, we extend the input space to accommodate multiple modalities. On the output side, we introduce **modality-specific projection heads(MSPH)** to recover each modality separately. This design enables our architecture to seamlessly support multi-modal inputs and outputs, ensuring flexible and controllable video generation.

Given a video sequence and its paired visual modalities $V = \{V_r, V_d, V_s, V_e\}$, where V_r, V_d, V_s, and V_e represent rgb, depth, segmentation, and canny, respectively, we first encode them into a latent space using a pretrained 3D-causal VAE encoder $\mathcal{E}$ (Yang et al. 2024b). Each modality is mapped to latent patches to get the noisy latents:

$$x_m = \mathcal{E}(V_m), \quad m \in \{r, d, s, c\}. \tag{2}$$

where $x_m \in \mathbb{R}^{F \times H \times W \times C}$ and F, H, W, C denote the number of frames, height, width, and latent channels.

Next, we blend the latent representations of each modality with noise:

$$x_m^t = (1-t) \cdot \epsilon + t \cdot x_m.$$

The noisy latents are then concatenated along the channel dimension to form a unified multi-modal representation: $x_i = \text{Concat}(x_r^t, x_d^t, x_s^t, x_c^t)$. This fused representation serves as the input to the diffusion transformer, enabling the video diffusion model to learn a joint distribution over the multiple modalities.

On the output side, we employ modality-specific projection heads H_m, where each head is responsible for reconstructing the noise output ϵ_m of a specific modality from the diffusion transformer output x_o:

$$\epsilon_m = H_m(x_o) \tag{3}$$

Specifically, we adopt the original rgb projection head from CogVideoX and replicate it for each modality, rather than simply extending the output channels of a shared rgb head. This design better accommodates the distinct characteristics of different modalities. Finally, the denoised latents are decoded back into the color space using the pretrained 3D-VAE decoder $\mathcal{D}$ (Yang et al. 2024b), producing high-fidelity multi-modal video outputs.

Adaptive modality control strategy A key challenge in unified video generation is determining the role of each modality—whether it serves as a generation signal or a conditioning input. To address this, we introduce an **adaptive modality control strategy(AMCS)** that dynamically assigns roles to different modalities based on the task.

During training, generation modalities are blended with noise before being fed into the diffusion model, while conditioning modalities remain unchanged and are concatenated with the noisy inputs of other modalities to serve as conditioning signals. This mechanism ensures flexible and adaptive control over different modalities, allowing the model to seamlessly handle diverse tasks within a unified framework. Specifically, in a text-to-video generation task, all modalities are generated from pure noise, meaning they act as generation signals. In an X-conditioned generation task, where X represents depth, segmentation, or canny, the conditioning modality X is provided as input directly without blending with noise and concatenated with the noisy latent representations of other modalities. Notably, if X represents the rgb modality, the model instead performs a video understanding task and predicts corresponding multi-modal outputs.

$$\mathbf{x}_m^t = \begin{cases} (1-t) \cdot \epsilon + t \cdot x_m, & \text{if } m \text{ is for generation} \\ x_m, & \text{if } m \text{ is for conditioning} \end{cases} \tag{4}$$

To further enhance the diffusion model's ability to distinguish modality roles, we introduce a modality embedding $\mathbf{e}_m$ that differentiates between generation ($\mathbf{e}_g$) and conditioning ($\mathbf{e}_c$) roles, which can be directly added to the diffusion model input $\mathbf{x}_m^t$.

$$\mathbf{e}_m = \begin{cases} \mathbf{e}_g, & \text{if } m \text{ is for generation} \\ \mathbf{e}_c, & \text{if } m \text{ is for conditioning} \end{cases} \tag{5}$$

$$\mathbf{x}_m^{t,'} = \mathbf{x}_m^t + \mathbf{e}_m \tag{6}$$

This strategy enables flexible and efficient control, allowing the model to seamlessly adapt to different tasks without requiring separate architectures for each modality.

Training

Training data Training a unified multi-modal model requires a large amount of paired data across modalities such as segmentation and depth. However, high-quality labeled video datasets are inherently scarce, posing a significant bottleneck. To address this, we employ expert models to generate pseudo labels for unlabeled videos, allowing us to efficiently construct a large-scale multi-modal dataset without manual annotation. Benefiting from the rapid advancements of 2D foundation models (Ravi et al. 2024; Chen et al. 2025), these expert models can provide high-quality annotations at scale, enabling us to leverage large volumes of raw video data for effective training. Specifically, for video depth, we use Video Depth Anything (Chen et al. 2025) to generate temporally consistent depth maps across video sequences. For segmentation, we apply Semantic-SAM (Li et al. 2023) on the first frame for instance segmentation, then propagate the results to subsequent frames using SAM2 (Ravi et al. 2024) to maintain semantic consistency. For canny edges, we adopt the OpenCV implementation of the Canny algorithm (Canny 1986) for edge detection.

In total, we processed 400K video samples, randomly sampled from the Koala-36M (Wang et al. 2024a) dataset. The inference of the video depth estimation model took approximately 3 days, while the video segmentation model required around 5 days, both conducted using 8 NVIDIA H100 GPUs in parallel.

Training loss We optimize our unified video generation and understanding framework using a multi-modality diffusion loss, ensuring high-quality generation while maintaining flexibility across different modalities. For each modality, we apply an independent denoising loss. If a modality

	subject consistency	b.g. cons.	motion smoothness	dynamic degree	aesthetic quality	imaging quality	weighted average
CogVideoX(Yang et al. 2024b)	95.68	96.00	98.21	**53.98**	50.75	65.77	72.25
OmniVDiff(ours)	**97.78**	**96.26**	**99.21**	49.69	**51.47**	**67.13**	**72.78**

Table 1: VBench metrics for text-conditioned video generation. We compare our method with the prior baseline CogVideoX. For each metric group, the best performance is shown in **bold**.

Model	subj. cons.	b.g. cons.	motion smoothness	dynamic degree	aesthetic quality	imaging quality	weighted average
text+depth							
Control-A-Video(Chen et al. 2023)	89.99	91.63	91.90	40.62	48.67	68.69	68.53
ControlVideo(Zhang et al. 2023)	95.50	94.17	97.80	18.35	**57.56**	70.09	70.71
Make-your-video(Xing et al. 2024)	90.04	92.48	97.64	51.95	44.67	**70.26**	70.17
VideoX-Fun(aigc-apps 2024)	96.25	95.73	98.90	50.43	55.81	55.38	72.85
OmniVDiff(ours)	**97.96**	**96.66**	**99.18**	**53.32**	52.95	67.26	**73.45**
text+canny							
CogVideoX+CTRL(TheDenk 2024)	96.26	94.53	98.42	53.44	49.34	55.56	70.13
Control-A-Video(Chen et al. 2023)	89.81	91.27	97.86	41.79	47.23	**68.77**	69.31
ControlVideo(Zhang et al. 2023)	95.23	94.00	97.12	17.58	**55.81**	55.38	67.72
VideoX-Fun(aigc-apps 2024)	96.69	95.41	99.15	50.78	52.99	66.76	72.73
OmniVDiff(ours)	**97.84**	**95.55**	**99.23**	**53.53**	52.34	67.14	**73.14**
text+segment							
OmniVDiff(ours)	**97.97**	**95.81**	**99.31**	**53.18**	**53.37**	**67.51**	**73.42**

Table 2: VBench metrics for depth-, canny-, and segmentation-conditioned video generation. For each condition type, the best performance is shown in **bold**, and the second-best is marked with an underline.

serves as a conditioning input, its denoising loss is skipped, ensuring it only guides generation without being explicitly optimized. The final objective is:

$$\mathcal{L} = \sum_{m, m \notin Cond} \mathbb{E}_{\mathbf{x}_m, t, \epsilon, m} \left[\| \epsilon - \epsilon_\theta(\mathbf{x}_m^{t,'}, t, e_m) \|^2 \right] \quad (7)$$

This approach provides adaptive supervision, enabling flexible modality roles and allowing the model to transition seamlessly between generation and conditioning.

Experiments

Implementation Details

We fine-tune our model based on CogVideoX (Yang et al. 2024b), a large-scale text-to-video diffusion model. Specifically, we adopt CogVideoX1.5-5B as the base model for our fine-tuning. The fine-tuning process follows a two-stage training strategy, progressively adapting the model from multi-modality video generation to multi-modal controllable video synthesis with the support of X-conditioned video generation and video visual understanding. We train the model using a learning rate of 2e-5 on 8 H100 GPUs for 40K steps. The model is optimized using a batch size of 8, with each training stage consisting of 20K steps. To evaluate the performance of video generation, we follow (Team et al. 2025) and report evaluation metrics follow VBench (Huang et al. 2024), a standard benchmark for video generation.

Omni Controllable Video Generation

We evaluate our approach against state-of-the-art methods on three tasks: text-conditioned video generation, X-conditioned video generation, and video understanding.

Text-conditioned video generation Given a text prompt, *OmniVDiff* generates multi-modal video sequences simultaneously within a single diffusion process. To provide a comprehensive evaluation of our generation performance, we compare our method with the baseline video diffusion model CogVideoX (Yang et al. 2024b) on rgb video generation and assess the generation quality on VBench(Huang et al. 2024) metrics. Note that for this comparison, we focus on the rgb modality to ensure consistency with CogVideoX, which does not support multi-modal outputs. Table 1 presents a quantitative comparison, where our model achieves a comparable VBench metric with CogVideoX, demonstrating superior generation quality. Although our focus is on multi-modal training, the joint optimization may provide stronger regularization than using rgb alone, potentially resulting in more coherent and consistent predictions.

X-conditioned video generation We evaluate our unified framework on X-conditioned video synthesis, comparing it with specialized baselines that leverage visual cues such as depth, canny, or segmentation. As shown in Table 2 and Figure 3, our model outperforms depth-specific baselines in depth-conditioned video generation, exhibiting superior structural fidelity and stronger alignment with the depth guidance signal. Furthermore, Table 2 also demonstrates that our approach surpasses existing modality-specific methods in segmentation- and canny-guided synthesis. Benefiting from a unified diffusion architecture, our model enables controllable video synthesis across multiple modalities within a cohesive framework. See the supplementary for details.

Rgb-conditioned video understanding To assess video understanding, we compare our model with baselines de-

	subject consistency	b.g. consistency	motion smoothness	dynamic degree	aesthetic quality	imaging quality	weighted average
w/o modality embedding	97.11	95.59	98.97	41.80	50.25	66.43	71.54
w/o AMCS	97.31	96.19	99.01	33.28	50.82	**67.31**	71.21
w/o MSPH	96.76	95.44	99.12	41.41	50.26	65.81	71.35
OmniVDiff(Ours)	**97.78**	**96.26**	**99.21**	**49.69**	**51.47**	67.13	**72.78**

Table 3: VBench metrics for the ablation study under different training settings. For each group of metrics, the best performance is highlighted in **bold**, and the second-best is indicated with an underline.

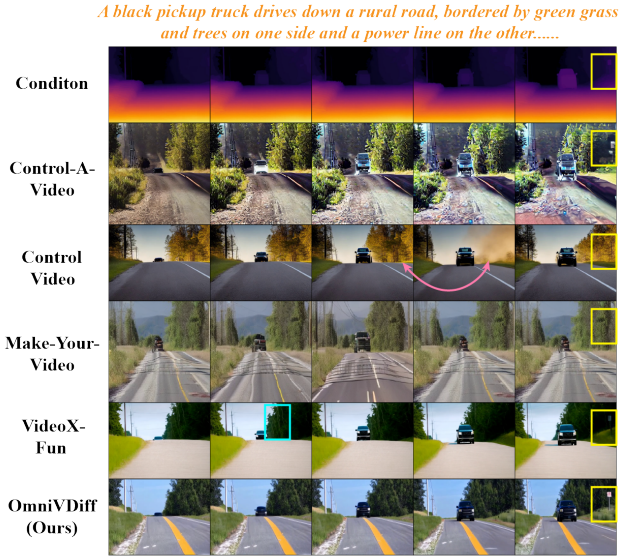

Figure 3: Visual comparison for depth-guided video generation. Yellow boxes highlight regions where our method better aligns with the provided depth compared to the baseline. Red arrows indicate temporal flickering, while cyan boxes denote artifacts in the rgb outputs.

signed for depth and segmentation estimation.

For depth estimation, we follow the Video Depth Anything protocol (Chen et al. 2025) and evaluate the zero-shot performance on the ScanNet dataset (Dai et al. 2017). As shown in Table 4, *OmniVDiff* achieves state-of-the-art performance among all baselines, delivering results comparable to the expert model VDA-S. Notably, VDA-S serves as our teacher model and is trained with high-quality ground-truth depth supervision, while *OmniVDiff* is trained solely with pseudo labels generated by VDA-S.

Although designed for controllable video diffusion, our model may benefit from high-quality ground-truth data for understanding tasks. We ablate this by introducing a small set of 10k synthetic samples into the training data. With this setting, *OmniVDiff*-Syn surpasses VDA-S in accuracy and produces sharper, more precise geometric details (Figure 4). This demonstrates the model's ability to leverage small amounts of high-quality data for significant gains.

Similarly, Table 5 presents quantitative comparisons on segmentation estimation, where our method achieves superior performance over baseline methods. Additional results are provided in the supplementary material.

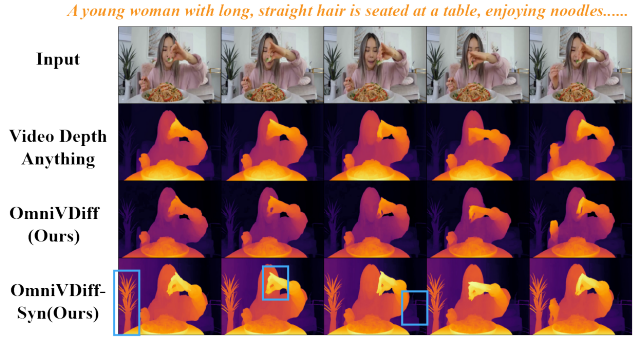

Figure 4: Qualitative comparison of video depth estimation. Yellow boxes highlight areas where both *OmniVDiff*-Syn succeed in capturing sharper details and achieving superior geometric fidelity.

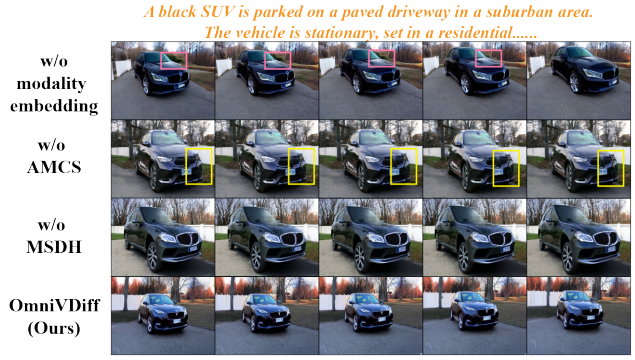

Figure 5: Qualitative comparison of ablation variants under different training configurations. Red boxes highlight missing rearview mirrors in the generated vehicles, while yellow boxes indicate visual artifacts.

Ablation study We conduct an ablation study to assess the contributions of key design components, focusing specifically on the *modality embedding*, *adaptive modality control strategy (AMCS)*, and the *modality-specific projection heads (MSPH)*. As shown in Table 3 and Figure 5, the full model consistently outperforms all ablated variants across all modalities. Introducing modality embeddings improves the model's understanding of each modality's role, whether as conditioning or generation input. The use of adaptive modality control facilitates flexible multi-modal control and understanding. Moreover, modality-specific projections allow the model to better capture the unique characteristics of each modality. Together, the results confirm that these designs play a crucial role in enabling precise control and faithful synthesis in our unified diffusion framework.

Method	AbsRel ↓	δ_1 ↑
DAv2-L(Yang et al. 2024a)	0.150	0.768
NVDS(Wang et al. 2023)	0.207	0.628
NVDS + DAv2-L	0.194	0.658
ChoronDepth (Shao et al. 2024)	0.199	0.665
DepthCrafter(Hu et al. 2024)	0.169	0.730
VDA-S (e)(Chen et al. 2025)	<u>0.110</u>	<u>0.876</u>
OmniVDiff(Ours)	0.125	0.852
OmniVDiff-Syn(Ours)	**0.100**	**0.894**

Table 4: Zero-shot video depth estimation results. We compare our method with representative single-image and video depth estimation models. "VDA-S(e)" denotes the expert model with a ViT-Small backbone. The **best** and <u>second-best</u> results are highlighted.

Method	COCO Val 2017(Lin et al. 2015)	
	Point (Max) 1-IoU ↑	Point (Oracle) 1-IoU ↑
SAM (B)	52.1	68.2
SAM (L)	55.7	70.5
Semantic-SAM (T)	54.5	73.8
Semantic-SAM (L)(e)	**57.0**	**74.2**
OmniVDiff(ours)	<u>56.0</u>	<u>73.9</u>

Table 5: Comparison with prior methods on point-based interactions, evaluated on COCO Val2017. "Max" selects the prediction with the highest confidence score, while "Oracle" uses the one with highest IoU against the target mask.

Inference efficiency Our unified model offers significant efficiency advantages by supporting multi-modal video outputs within a single framework. Compared to CogVideoX, which generates only rgb videos, our model additionally produces segmentation and depth outputs with comparable inference speed and memory usage (Table 6). Moreover, unlike pipelines that rely on separate expert models for each modality—incurring substantial overhead (e.g., segmentation requires 30 seconds via separate inference)—our unified design reduces total inference time and eliminates the need to deploy multiple networks.

Applications

Our unified model provides significant advantages in controllability and flexibility. In this section, we showcase its versatility through two representative applications:

Video-to-video style control *OmniVDiff* can be directly applied to video-to-video style control, enabling structure-preserving video generation guided by text prompts. Given a reference video (Figure 6 (a)), *OmniVDiff* first estimates depth modality as an intermediate representation, which is then used to generate diverse scene styles (Figure 6 (b)) (e.g., winter), while preserving the original spatial layout. Thanks to joint training, *OmniVDiff* achieves this without relying on external depth experts, ensuring structural consistency. We further provide a quantitative comparison of video-to-video style control using *OmniVDiff*'s estimated depth versus expert-provided depth, demonstrating comparable consistency and visual quality (see supplementary for details).

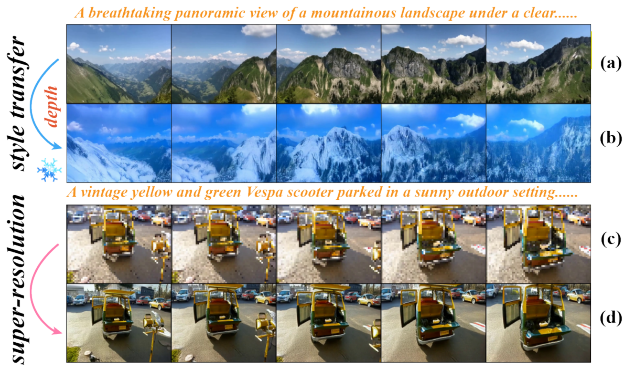

Figure 6: Applications: (a, b): Video-to-video style control. (c, d): Adapt to new tasks: video super-resolution.

Methods	Paras	Time	Memory
Video Depth Anything	28.4M	4s	13.62GB
Semantic-Sam & SAM2	222.8 & 38.9M	30s	6.75GB
CogVideoX	5B	41s	26.48GB
OmniVDiff(Ours)	**5B+11.8M**	**44s**	**26.71GB**

Table 6: Comparison of Model Inference Time, Memory Usage, and Parameter Size. *OmniVDiff* demonstrates its inference efficiency among compared models.

Adaptability to new modalities/tasks To evaluate our model's adaptability to new modalities and applications, we conduct experiments on a representative task: video super-resolution. Specifically, we fine-tune *OmniVDiff* for 2k steps, repurposing an existing modality slot (canny) to handle low-resolution rgb videos during training. At inference, these inputs serve as conditioning signals (Figure 6 (c)), enabling the model to generate high-resolution outputs (Figure 6 (d)), demonstrating its flexibility in handling unseen modalities with minimal adjustments.

Conclusion

In this paper, we present *OmniVDiff*, a unified framework for multi-modal video generation and understanding that extends diffusion models to support text-to-video, modality-conditioned generation, and visual understanding within a single architecture. By simultaneously generating multiple modalities (i.e., rgb, depth, segmentation, and canny) and incorporating an adaptive modality control strategy, our approach flexibly handles diverse generation and conditioning scenarios. Furthermore, our unified design eliminates the need for separate expert models and sequential processing pipelines, offering a scalable and efficient solution that easily adapts to new modalities while maintaining high performance across video tasks. Future research can explore expanding modality support, adopting more powerful pre-trained models (like WAN (Wan et al. 2025)), and enhancing real-time efficiency, further advancing the capabilities of unified video diffusion models.

References

aigc-apps. 2024. VideoX-Fun: A Video Generation Pipeline for AI Images and Videos. https://github.com/aigc-apps/VideoX-Fun. GitHub repository, accessed 2025-07-21.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Byung-Ki, K.; Dai, Q.; Hyoseok, L.; Luo, C.; and Oh, T.-H. 2025. JointDiT: Enhancing RGB-Depth Joint Modeling with Diffusion Transformers. *arXiv preprint arXiv:2505.00482*.

Canny, J. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 679–698.

Chefer, H.; Singer, U.; Zohar, A.; Kirstain, Y.; Polyak, A.; Taigman, Y.; Wolf, L.; and Sheynin, S. 2025. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*.

Chen, H.; Zhang, Y.; Cun, X.; Xia, M.; Wang, X.; Weng, C.; and Shan, Y. 2024a. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7310–7320.

Chen, S.; Guo, H.; Zhu, S.; Zhang, F.; Huang, Z.; Feng, J.; and Kang, B. 2025. Video Depth Anything: Consistent Depth Estimation for Super-Long Videos. *arXiv:2501.12375*.

Chen, W.; Ji, Y.; Wu, J.; Wu, H.; Xie, P.; Li, J.; Xia, X.; Xiao, X.; and Lin, L. 2023. Control-A-Video: Controllable Text-to-Video Diffusion Models with Motion Prior and Reward Feedback Learning. *arXiv preprint arXiv:2305.13840*.

Chen, X.; Zhang, Z.; Zhang, H.; Zhou, Y.; Kim, S. Y.; Liu, Q.; Li, Y.; Zhang, J.; Zhao, N.; Wang, Y.; Ding, H.; Lin, Z.; and Hengshuang. 2024b. UniReal: Universal Image Generation and Editing via Learning Real-world Dynamics. *arXiv preprint arXiv:2412.07774*.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. arXiv:1702.04405.

Feng, R.; Weng, W.; Wang, Y.; Yuan, Y.; Bao, J.; Luo, C.; Chen, Z.; and Guo, B. 2024. Ccedit: Creative and controllable video editing via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6712–6722.

Gan, Q.; Ren, Y.; Zhang, C.; Ye, Z.; Xie, P.; Yin, X.; Yuan, Z.; Peng, B.; and Zhu, J. 2025. HumanDiT: Pose-Guided Diffusion Transformer for Long-form Human Motion Video Generation. *arXiv preprint arXiv:2502.04847*.

Guo, Y.; Yang, C.; Rao, A.; Agrawala, M.; Lin, D.; and Dai, B. 2024. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, 330–348. Springer.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. *Advances in Neural Information Processing Systems*, 35: 8633–8646.

Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*.

Hu, L.; Wang, G.; Shen, Z.; Gao, X.; Meng, D.; Zhuo, L.; Zhang, P.; Zhang, B.; and Bo, L. 2025. Animate Anyone 2: High-Fidelity Character Image Animation with Environment Affordance. *arXiv preprint arXiv:2502.06145*.

Hu, W.; Gao, X.; Li, X.; Zhao, S.; Cun, X.; Zhang, Y.; Quan, L.; and Shan, Y. 2024. DepthCrafter: Generating Consistent Long Depth Sequences for Open-world Videos. arXiv:2409.02095.

Huang, T.; Zheng, W.; Wang, T.; Liu, Y.; Wang, Z.; Wu, J.; Jiang, J.; Li, H.; Lau, R. W. H.; Zuo, W.; and Guo, C. 2025. Voyager: Long-Range and World-Consistent Video Diffusion for Explorable 3D Scene Generation. arXiv:2506.04225.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; Wang, Y.; Chen, X.; Wang, L.; Lin, D.; Qiao, Y.; and Liu, Z. 2024. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jiang, Z.; Han, Z.; Mao, C.; Zhang, J.; Pan, Y.; and Liu, Y. 2025. VACE: All-in-One Video Creation and Editing. *arXiv preprint arXiv:2503.07598*.

Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15954–15964.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.

Le, D. H.; Pham, T.; Lee, S.; Clark, C.; Kembhavi, A.; Mandt, S.; Krishna, R.; and Lu, J. 2024. One Diffusion to Generate Them All. arXiv:2411.16318.

Li, F.; Zhang, H.; Sun, P.; Zou, X.; Liu, S.; Yang, J.; Li, C.; Zhang, L.; and Gao, J. 2023. Semantic-SAM: Segment and Recognize Anything at Any Granularity. *arXiv preprint arXiv:2307.04767*.

Liang, R.; Gojcic, Z.; Ling, H.; Munkberg, J.; Hasselgren, J.; Lin, Z.-H.; Gao, J.; Keller, A.; Vijaykumar, N.; Fidler, S.; et al. 2025. DiffusionRenderer: Neural Inverse and Forward Rendering with Video Diffusion Models. *arXiv preprint arXiv:2501.18590*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312.

Liu, C.; Li, R.; Zhang, K.; Lan, Y.; and Liu, D. 2024. StableV2V: Stablizing Shape Consistency in Video-to-Video Editing. *arXiv preprint arXiv:2411.11045*.

Lv, J.; Huang, Y.; Yan, M.; Huang, J.; Liu, J.; Liu, Y.; Wen, Y.; Chen, X.; and Chen, S. 2024. Gpt4motion: Scripting

physical motions in text-to-video generation via blender-oriented gpt planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1430–1440.

Polyak, A.; Zohar, A.; Brown, A.; Tjandra, A.; Sinha, A.; Lee, A.; Vyas, A.; Shi, B.; Ma, C.-Y.; Chuang, C.-Y.; Yan, D.; Choudhary, D.; Wang, D.; Sethi, G.; Pang, G.; Ma, H.; Misra, I.; Hou, J.; Wang, J.; Jagadeesh, K.; Li, K.; Zhang, L.; Singh, M.; Williamson, M.; Le, M.; Yu, M.; Singh, M. K.; Zhang, P.; Vajda, P.; Duval, Q.; Girdhar, R.; Sumbaly, R.; Rambhatla, S. S.; Tsai, S.; Azadi, S.; Datta, S.; Chen, S.; Bell, S.; Ramaswamy, S.; Sheynin, S.; Bhattacharya, S.; Motwani, S.; Xu, T.; Li, T.; Hou, T.; Hsu, W.-N.; Yin, X.; Dai, X.; Taigman, Y.; Luo, Y.; Liu, Y.-C.; Wu, Y.-C.; Zhao, Y.; Kirstain, Y.; He, Z.; He, Z.; Pumarola, A.; Thabet, A.; Sanakoyeu, A.; Mallya, A.; Guo, B.; Araya, B.; Kerr, B.; Wood, C.; Liu, C.; Peng, C.; Vengertsev, D.; Schonfeld, E.; Blanchard, E.; Juefei-Xu, F.; Nord, F.; Liang, J.; Hoffman, J.; Kohler, J.; Fire, K.; Sivakumar, K.; Chen, L.; Yu, L.; Gao, L.; Georgopoulos, M.; Moritz, R.; Sampson, S. K.; Li, S.; Parmeggiani, S.; Fine, S.; Fowler, T.; Petrovic, V.; and Du, Y. 2025. Movie Gen: A Cast of Media Foundation Models. arXiv:2410.13720.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shao, J.; Yang, Y.; Zhou, H.; Zhang, Y.; Shen, Y.; Guizilini, V.; Wang, Y.; Poggi, M.; and Liao, Y. 2024. Learning Temporally Consistent Video Depth from Video Diffusion Priors. arXiv:2406.01493.

Team, A.; Zhu, H.; Wang, Y.; Zhou, J.; Chang, W.; Zhou, Y.; Li, Z.; Chen, J.; Shen, C.; Pang, J.; and He, T. 2025. Aether: Geometric-Aware Unified World Modeling. arXiv:2503.18945.

TheDenk. 2024. cogvideox-controlnet: ControlNet Extensions for CogVideoX. https://github.com/TheDenk/cogvideox-controlnet. GitHub repository, commit `<YOUR-COMMIT-HASH>`, accessed 2025-07-21.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; Zeng, J.; Wang, J.; Zhang, J.; Zhou, J.; Wang, J.; Chen, J.; Zhu, K.; Zhao, K.; Yan, K.; Huang, L.; Feng, M.; Zhang, N.; Li, P.; Wu, P.; Chu, R.; Feng, R.; Zhang, S.; Sun, S.; Fang, T.; Wang, T.; Gui, T.; Weng, T.; Shen, T.; Lin, W.; Wang, W.; Wang, W.; Zhou, W.; Wang, W.; Shen, W.; Yu, W.; Shi, X.; Huang, X.; Xu, X.; Kou, Y.; Lv, Y.; Li, Y.; Liu, Y.; Wang, Y.; Zhang, Y.; Huang, Y.; Li, Y.; Wu, Y.; Liu, Y.; Pan, Y.; Zheng, Y.; Hong, Y.; Shi, Y.; Feng, Y.; Jiang, Z.; Han, Z.; Wu, Z.-F.; and Liu, Z. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*.

Wang, J.; Wang, Z.; Pan, H.; Liu, Y.; Yu, D.; Wang, C.; and Wang, W. 2025. Mmgen: Unified multi-modal image generation and understanding in one go. *arXiv preprint arXiv:2503.20644*.

Wang, Q.; Shi, Y.; Ou, J.; Chen, R.; Lin, K.; Wang, J.; Jiang, B.; Yang, H.; Zheng, M.; Tao, X.; et al. 2024a. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. *arXiv preprint arXiv:2410.08260*.

Wang, Y.; Shi, M.; Li, J.; Huang, Z.; Cao, Z.; Zhang, J.; Xian, K.; and Lin, G. 2023. Neural video depth stabilizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9466–9476.

Wang, Z.; Xia, X.; Chen, R.; Yu, D.; Wang, C.; Gong, M.; and Liu, T. 2024b. LaVin-DiT: Large Vision Diffusion Transformer. *arXiv preprint arXiv:2411.11505*.

Xing, J.; Xia, M.; Liu, Y.; Zhang, Y.; Zhang, Y.; He, Y.; Liu, H.; Chen, H.; Cun, X.; Wang, X.; et al. 2024. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics*.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024a. Depth Anything V2. arXiv:2406.09414.

Yang, L.; Qi, L.; Li, X.; Li, S.; Jampani, V.; and Yang, M.-H. 2025. Unified Dense Prediction of Video Diffusion. arXiv:2503.09344.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024b. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Zhai, Y.; Lin, K.; Li, L.; Lin, C.-C.; Wang, J.; Yang, Z.; Doermann, D.; Yuan, J.; Liu, Z.; and Wang, L. 2024. Idol: Unified dual-modal latent diffusion for human-centric joint video-depth generation. In *European Conference on Computer Vision*, 134–152. Springer.

Zhang, Y.; Wei, Y.; Jiang, D.; Zhang, X.; Zuo, W.; and Tian, Q. 2023. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*.

Zhao, C.; Liu, M.; Zheng, H.; Zhu, M.; Zhao, Z.; Chen, H.; He, T.; and Shen, C. 2025. DICEPTION: A Generalist Diffusion Model for Visual Perceptual Tasks. *arXiv preprint arXiv:2502.17157*.

Zhao, Y.; Xie, E.; Hong, L.; Li, Z.; and Lee, G. H. 2023. Make-a-protagonist: Generic video editing with an ensemble of experts. *arXiv preprint arXiv:2305.08850*.