JVT: JUST VIDEO TRANSFORMER MAKES PIXEL-SPACE VIDEO DIFFUSION SIMPLE

Anonymous authors
Paper under double-blind review

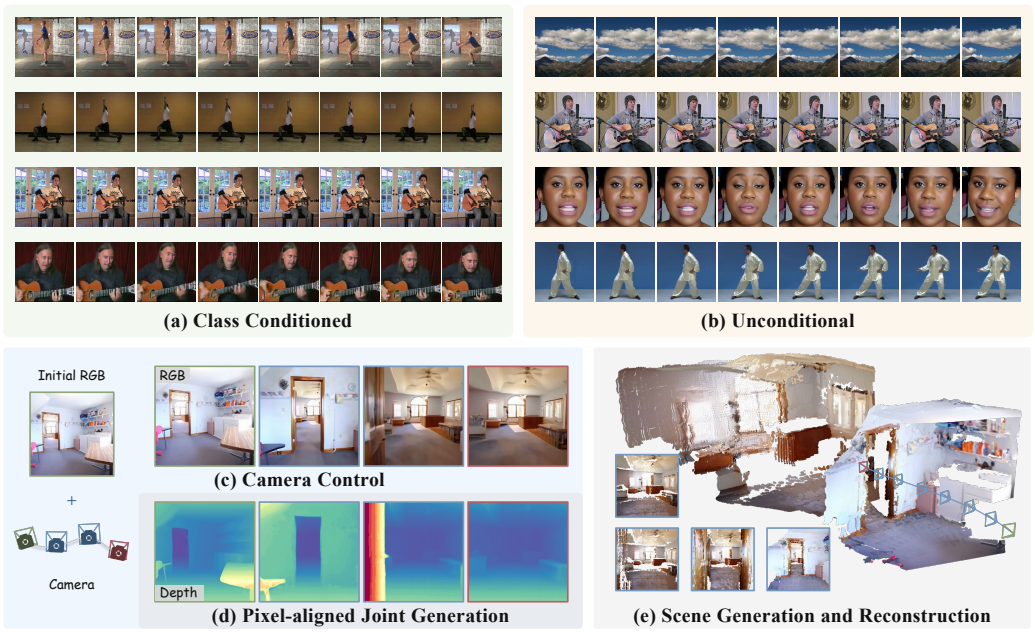


Figure 1: JVT generates videos directly in pixel space with a simple DiT-style transformer, supporting (a) class-conditioned and (b) unconditional generation. Given an initial RGB frame and a target camera trajectory, it jointly generates (c) RGB frames and disparity, visualized as (d) depth maps. These outputs yield (e) colored point clouds through geometric unprojection, without an additional reconstruction model.

ABSTRACT

Direct diffusion modeling of high-dimensional raw video data is challenging, and existing pixel-space methods often rely on specialized architectures. We introduce **JVT** (**J**ust **V**ideo **T**ransformer), a simple large-patch DiT-style Transformer for pixel-space video diffusion, without a pretrained autoencoder, auxiliary detailers, cascaded stages, or U-Net-style skip connections. We systematically examine video patch embeddings, register tokens, and adaptive layer normalization to identify design choices that balance generation quality and computational efficiency. The resulting JVT design achieves generation quality competitive with latent diffusion baselines in both class-conditioned and unconditional settings. The pixel-space formulation of JVT also naturally supports joint generation of pixel-aligned modalities. We demonstrate this capability through camera-conditioned RGB-disparity generation, which improves camera following relative to RGB-only baselines and yields explicit scene geometry without an additional reconstruction model at inference. These results highlight the potential of simple DiT-style Transformers for pixel-space video diffusion, and we hope our findings offer useful insights and practical guidance for future research.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Diffusion and flow-based models have become leading paradigms for visual generative modeling, driven by rapid advances in algorithm designs and model architectures. Early diffusion models (Ho et al., 2020; Dhariwal & Nichol, 2021) operate directly on pixels with classical U-Nets (Ronneberger et al., 2015), but struggle with efficiency and fidelity at high resolutions. To alleviate this difficulty, latent diffusion moves denoising from pixel space to a compact latent space produced by a separately trained autoencoder (Rombach et al., 2022). Combined with the Diffusion Transformer (DiT) architecture (Peebles & Xie, 2023), this formulation has become a standard design for high-resolution visual generation.

Despite its efficiency, latent diffusion is inherently bounded by the autoencoder, including reconstruction error (Yu et al., 2026; Zheng et al., 2026), additional training and inference costs (Leng et al., 2025; Cheng & Yuan, 2025; Wang et al., 2026c), and limited cross-modal capability for joint modeling (Bounoua et al., 2024; Qiu et al., 2025). Thus recent works revisit pixel-space diffusion and demonstrate competitive quality on image generation (Li & He, 2026; Yu et al., 2026; Lu et al., 2026; Wang et al., 2026b). In contrast, video generation is still dominated by latent DiTs (Ma et al., 2025; Kong et al., 2024; Yang et al., 2025; Wan et al., 2025), while pixel-space models mostly retain U-Net-style skip connections (Gu et al., 2024; Bar-Tal et al., 2024; Hoogeboom et al., 2025) or cascaded generation (Menapace et al., 2024; Skorokhodov et al., 2024). This motivates us to investigate how far a simple DiT-style Transformer can go in direct pixel-space video generation.

We introduce JVT, a pixel-space video generative model built on a simple DiT-style Transformer backbone, without detailers, cascaded stages or U-Net-style skip connections. Following JiT (Li & He, 2026), JVT predicts clean videos from noisy pixels under a rectified-flow objective. To make raw videos tractable, we explore variants of video patch embedding for efficient spatiotemporal downsampling. We further examine the design choices of register tokens (Darcet et al., 2024) and adaptive layer normalization (AdaLN) (Peebles & Xie, 2023), which can carry noise level or class condition and amplify it by register guidance (Starodubcev et al., 2026) or classifier-free guidance (Ho & Salimans, 2021). With these design choices, JVT achieves generation quality competitive with latent diffusion baselines in both class-conditioned and unconditional settings.

Pixel-space video modeling also provides a direct interface for jointly generating video and pixel-aligned modalities, which typically requires reusing an RGB-specific autoencoder in latent diffusion (Ke et al., 2024; Zhu et al., 2025; Huang et al., 2026a). We extend JVT to camera-conditioned joint RGB-disparity generation to demonstrate this cross-modal capability. Without an additional reconstruction model at inference, the generated disparity yields a sequence of point clouds in world coordinates, which enables explicit 4D world modeling. Experiments further show that joint RGB-disparity generation improves camera following relative to RGB-only JVT.

The main contributions of this work are as follows:

- We present JVT, the first pixel-space video diffusion model built on a simple DiT-style Transformer backbone, without detailers, cascaded stages, or U-Net-style skip connections.
- We systematically study video patch embeddings and conditioning mechanisms, identifying designs that improve the quality-efficiency trade-off. In class-conditioned and unconditional settings, JVT achieves generation quality competitive with latent diffusion baselines.
- We formulate pixel-aligned joint generation as a native extension of JVT. We further demonstrate this capability through camera-conditioned RGB-disparity video generation, which enables explicit 4D scene representation and improves camera following relative to RGB-only generation.

2 RELATED WORK

High-resolution pixel-space diffusion. Pixel-space diffusion modeling for high-resolution visual data has long relied on specialized architectures, including cascaded diffusion (Ho et al., 2022a; Teng et al., 2024), cross-resolution designs (Gu et al., 2024; Chen et al., 2025b), multiscale or U-shaped backbones (Hoogeboom et al., 2023; Crowson et al., 2024; Hoogeboom et al., 2025; He et al., 2026; Guo et al., 2026), frequency decomposition (Ma et al., 2026b; Lin et al., 2026; Ma et al., 2026a), and

auxiliary detail modules (Wang et al., 2026a; Chen et al., 2026; Yu et al., 2026). Though effective, these approaches often introduce additional architectural or training complexity, which can limit efficiency and scalability. Recently, JiT (Li & He, 2026) achieves high-resolution image generation competitive with latent diffusion using a simple large-patch DiT-style Transformer. This paradigm is also extended to one-step image generation (Lu et al., 2026) and text-to-image generation (Wang et al., 2026b; Jiang et al., 2026), further demonstrating the effectiveness.

Video Transformer architectures. Modern video Transformers largely retain the standard blocks of 2D Vision Transformers (ViTs), making patch embedding the key adaptation to video inputs (Tong et al., 2022; Wang et al., 2023; Bardes et al., 2024; Assran et al., 2025). A common choice is is tubelet embedding (Arnab et al., 2021), which encodes non-overlapping spatiotemporal patches into tokens via 3D convolution. This architecture is also widely adopted in latent video DiTs, whereas pixel-space video diffusion methods still typically rely on U-shaped architectures (Gu et al., 2024; Bar-Tal et al., 2024; Hoogeboom et al., 2025) or cascaded generation (Menapace et al., 2024; Skorokhodov et al., 2024). Our work aims to explore whether a simple DiT-style video Transformer can also support direct pixel-space video generation.

3 METHOD

3.1 PRELIMINARIES

Though denoising diffusion models commonly use noise, score or velocity prediction (Dhariwal & Nichol, 2021; Song et al., 2021; Karras et al., 2022; Salimans & Ho, 2022; Lipman et al., 2023; Liu et al., 2023), JiT proposes clean data prediction and velocity loss supervision for pixel-space diffusion (Li & He, 2026). Following this formulation, JVT predicts clean video $\hat{x}_0$ from noisy pixels x_t:

$$x_t = (1-t)x_0 + t\epsilon, \qquad \hat{x}_0 = f_\theta(x_t, t, c), \qquad \hat{v} = (\hat{x}_0 - x_t)/t, \tag{1}$$

where noise $\epsilon \sim \mathcal{N}(0, I)$, noise level $t \in (0, 1]$, and c denotes the available conditioning. Thus the standard rectified-flow velocity loss can be represented as:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}\,\|\hat{v} - v\|_2^2 = \mathbb{E}\|(\hat{x}_0 - x_0)/t\|_2^2, \qquad v = x_0 - \epsilon = (x_0 - x_t)/t. \tag{2}$$

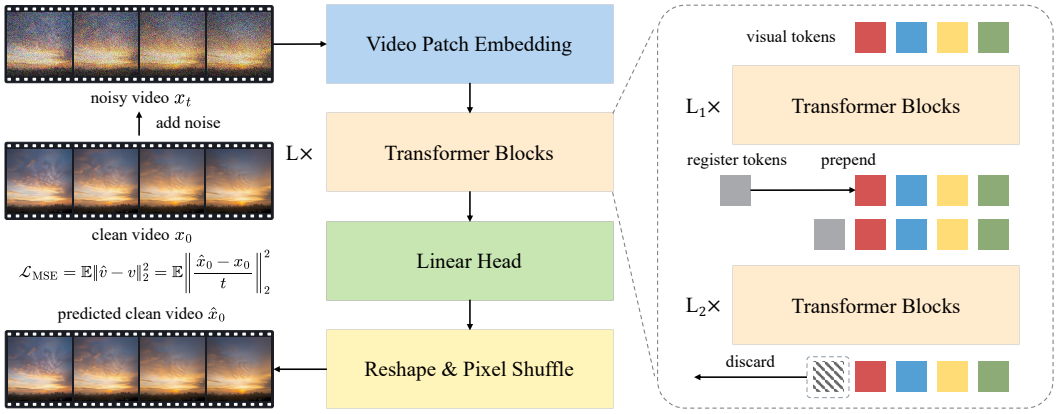


Figure 2: **JVT architecture and register token design**. The network predicts a clean video $\hat{x}_0$ from a noisy input x_t. Register tokens are prepended to visual tokens after the first L_1 blocks, and then discarded after the remaining L_2 blocks, where $L = L_1 + L_2$.

3.2 PATCH EMBEDDING

Patch embedding is central to pixel-space DiTs, as it provides the initial downsampling that latent DiTs obtain from an autoencoder. For videos, it also aggregates spatiotemporal information while controlling the token sequence length.

Following standard video Transformer designs, we use plain tubelet embedding as our baseline. As illustrated in Fig. 3 (a), it uses a single 3D convolution to project each non-overlapping patch of temporal size p_t and spatial size $p_h \times p_w$ into a C-dimensional token.

Bottleneck structure. Following JiT (Li & He, 2026), we introduce a channel bottleneck into tubelet embedding (Fig. 3 (b)). A 3D convolution first projects each patch into a C_b-dimensional representation, which is then expanded to C channels by a pointwise projection. We typically choose $C_b < C$ to encourage compact, low-dimensional representations (Tishby et al., 2000; Rifai et al., 2011; Makhzani & Frey, 2013; Alemi et al., 2017).

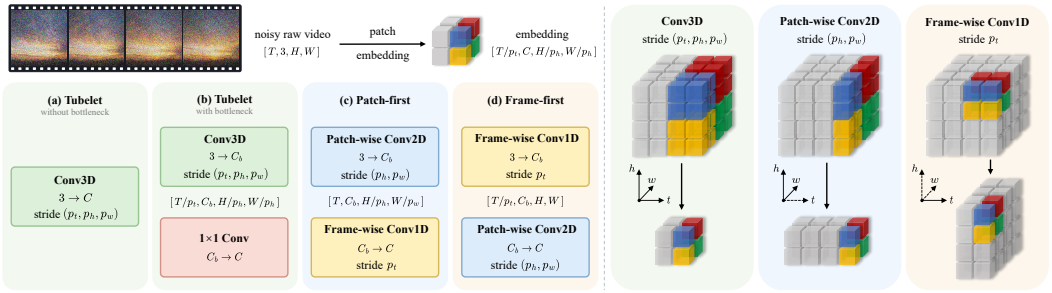


Figure 3: **Video patch embedding variants (left) and their convolutional components (right)**. Assume non-overlapping patches with size $p_t \times p_h \times p_w$. T, H, W are divisible by p_t, p_h, p_w, respectively.

Spatiotemporal factorization. Inspired by the factorization principle of depthwise separable convolution (Chollet, 2017; Howard et al., 2017), we separate the spatial and temporal operations of tubelet embedding. Specifically, a patch-wise 2D convolution extracts spatial features within each frame, while a frame-wise 1D convolution aggregates temporal information at each spatial location. Applying the spatial operation first yields patch-first embedding (Fig. 3 (c)), while reversing the order yields frame-first embedding (Fig. 3 (d)).

3.3 Register Tokens

Register tokens provide additional computational slots in ViTs (Darcet et al., 2024) and also support in-context conditioning and improve generation quality in pixel-space image DiTs (Li & He, 2026; Starodubcev et al., 2026). We extend this mechanism to JVT and examine its roles in conditioning and sampling guidance.

As illustrated in Fig. 2, given $L = L_1 + L_2$ Transformer blocks, register tokens are prepended to the visual sequence after the first L_1 blocks and discarded after the remaining L_2 blocks. Let $X^{(L_1)} \in \mathbb{R}^{N \times C}$ denote the N visual tokens at the insertion point and $R_0 \in \mathbb{R}^{\ell_r \times C}$ the ℓ_r learnable base registers. With $e_t, e_y \in \mathbb{R}^C$ denoting the noise level and class embeddings, respectively, we construct the register tokens and augmented sequence as

$$R = R_0 + \mathrm{Repeat}(e_r, \ell_r), \qquad \widetilde{X}^{(L_1)} = \mathrm{Concat}(R, X^{(L_1)}), \tag{3}$$

where $e_r \in \{0, e_t, e_y, e_t + e_y\}$ and $\mathrm{Repeat}(e_r, \ell_r)$ replicates e_r across the ℓ_r register positions. Setting $e_r = 0$ preserves $R = R_0$ without adding a conditioning embedding, which is different from removing the registers. For $e_r \in \{e_t, e_y, e_t + e_y\}$, the register tokens provide an additional conditioning pathway alongside AdaLN.

Register guidance. We also apply register guidance (RG) (Starodubcev et al., 2026) during sampling. For the same noisy input and condition, let v_r and v_{nr} denote the velocity predictions of the same model with and without registers, respectively. RG amplifies the register contribution through

$$v_{rg} = v_r + w_{rg}(v_r - v_{nr}), \qquad w_{rg} \geq 0. \tag{4}$$

For conditional generation, let v_u denote the unconditional velocity prediction with registers retained. Combining RG with classifier-free guidance (CFG) (Ho & Salimans, 2021) gives

$$v_{cfg+rg} = v_u + w_{cfg}(v_r - v_u) + w_{rg}(v_r - v_{nr}). \tag{5}$$

The scales w_{cfg} and w_{rg} separately weight the CFG and RG terms: setting $w_{rg} = 0$ recovers CFG, while $w_{cfg} = 1$ recovers RG in Eq. (4). Unlike the original RG procedure (Starodubcev et al., 2026), which randomly drops registers during training, we keep them active throughout training and remove them only to compute v_{nr} during sampling. Our experiments show that RG remains effective without register dropout during training.

3.4 Pixel-aligned Joint Generation

Pixel-aligned signals about geometry, motion, and semantic structure provide complementary cues for video generation. By channel-wise concatenating RGB frames with these signals (Ke et al., 2024), diffusion models can jointly generate RGB content and spatially aligned auxiliary modalities (Zhu et al., 2025; Chen et al., 2025a; Huang et al., 2026a).

Pixel-space modeling allows these modalities to be represented on a shared spatiotemporal grid without an RGB-specific autoencoder. After modality-specific normalization, we concatenate an RGB video $x_{\text{rgb}} \in \mathbb{R}^{T\times 3\times H\times W}$ with M pixel-aligned modalities along the channel dimension:

$$x_0 = [x_{\text{rgb}}; x^{(1)}; \ldots; x^{(M)}] \in \mathbb{R}^{T\times C_{\text{joint}}\times H\times W}, \tag{6}$$

where $x^{(m)} \in \mathbb{R}^{T\times C_m\times H\times W}$ denotes the m-th aligned modality and $C_{\text{joint}} = 3 + \sum_{m=1}^{M} C_m$ is the total number of channels. These modalities may include depth, surface normals, ray maps, segmentation masks, and optical flow. All modalities share the same noise level and are predicted jointly during denoising.

We instantiate this formulation with RGB and disparity, a reparameterization of depth, for camera-conditioned video generation (Sec. 4.3). At inference, we convert the predicted disparity back to depth and unproject it using the camera parameters. Combined with the generated RGB frames, this yields a sequence of colored point clouds without an additional reconstruction model.

4 Experiments

4.1 Main Results

Datasets. For class-conditioned generation and unconditional generation, we conduct experiments on four datasets: UCF-101 (Soomro et al., 2012), FaceForensics (Rössler et al., 2018), SkyTimelapse (Xiong et al., 2018), and Taichi-HD (Siarohin et al., 2019). Following settings of previous works (Skorokhodov et al., 2022; Ma et al., 2025), we use both training and test splits of UCF-101, and only the training split of the other datasets. We randomly extract 16-frame video clips with a frame interval of 3, and resize all video frames to 256×256 resolution for training.

Evaluation metrics. We evaluate generation quality with Fréchet Video Distance (FVD) (Unterthiner et al., 2019), using 2,048 generated samples. For efficiency metrics, we report parameter count, floating point operations (FLOPs), peak allocated GPU memory and inference latency. Additional details about these metrics are given in Appendix (A.3, A.6).

Baselines. Our main baseline is Latte (Ma et al., 2025), a representative DiT-based latent diffusion model. Other competing methods include GAN-based methods (Skorokhodov et al., 2022; Yu et al., 2022), autoregressive models (Yan et al., 2021; Wang et al., 2024a), U-Net-based latent diffusion (He et al., 2022; Yu et al., 2023b) and DiT-based latent diffusion (Liu et al., 2024).

Table 1 reports class-conditioned results on UCF-101 and unconditional results on the four datasets. JVT outperforms all competing methods in the table, demonstrating the effectiveness of direct pixel-space video modeling. Additional qualitative results are provided in Fig. 8 of the Appendix.

4.2 Ablation Study

Video patch embedding. We compare the tubelet, patch-first, and frame-first embeddings introduced in Sec. 3.2 while varying the bottleneck width C_b. Under identical training settings, all bottlenecked variants in Tab. 2 achieve lower final training loss than the plain tubelet baseline. With an appropriate choice of C_b, tubelet and patch-first embeddings also improve FVD while reducing computational cost. Taking both generation quality and efficiency into consideration, we adopt

Table 1: **FVD@2048 across video generation benchmarks.** Cond. and Uncond. denote class-conditioned and unconditional generation, respectively. Res. denotes spatial resolution, and all methods use 16-frame clips. Gray scores marked with †/‡ are taken from StyleGAN-V/DIGAN, and other values are sourced from the original papers. Taichi-HD scores marked with $*$ are trained on both training and test splits, excluded from ranking. JVT with and without "+RG" denote JVT with register guidance and JVT without registers, respectively. – denotes unreported results.

Method	**Res.**	**Cond.**	**Uncond.**			
		UCF-101	UCF-101	FaceForensics	SkyTimelapse	Taichi-HD
MoCoGAN (Tulyakov et al., 2018)	256^2	–	2886.9†	124.7†	206.6†	–
MoCoGAN-HD (Tian et al., 2021)	256^2	–	1729.6†	111.8†	164.1†	144.7‡
VideoGPT (Yan et al., 2021)	256^2	–	2880.6†	185.9†	222.7†	–
StyleGAN-V (Skorokhodov et al., 2022)	256^2	–	1431.0	47.41	79.52	–
MoStGAN-V (Shen et al., 2023)	256^2	–	1380.3	39.7	65.3	–
CogVideo (Hong et al., 2023)	160^2	626	–	–	–	–
DIGAN (Yu et al., 2022)	128^2	465	577	–	114.6	128.1*
DIGAN (Yu et al., 2022)	256^2	–	1630.2†	62.5†	83.11†	–
LVDM (He et al., 2022)	256^2	–	372	–	95.2	99.0*
PVDM (Yu et al., 2023b)	256^2	–	343.6	–	55.41	–
VIDM (Mei & Patel, 2023)	128^2	–	–	–	–	121.9
VIDM (Mei & Patel, 2023)	256^2	–	294.7	–	57.4	–
GLOBER (Sun et al., 2023)	128^2	151.5	239.5	–	–	124.2
GLOBER (Sun et al., 2023)	256^2	168.9	252.7	–	78.1	–
OmniTokenizer (Wang et al., 2024a)	256^2	191	–	–	–	–
VDT (Lu et al., 2024)	64^2	–	225.7	–	–	–
FVDM (Liu et al., 2024)	256^2	468.23	–	55.01	106.09	194.61
Latte (Ma et al., 2025)	256^2	477.97	–	34.00	59.82	159.60
JVT (Ours)	256^2	136.5	413.5	32.86	**46.58**	**119.1**
JVT+RG (Ours)	256^2	**129.2**	**215.9**	**31.80**	47.80	124.3

Table 2: **Performance of different video patch embedding variants.** Efficiency metrics are measured on one NVIDIA 4090 GPU. Denoiser rows measure the complete model, while Emb. rows measure the isolated video patch embedding module on a $1 \times 16 \times 3 \times 256 \times 256$ input. Embedding peak allocated memory includes the input and embedding parameters.

Embedding	**Tubelet**				**Patch-first**			**Frame-first**		
C_b	-	32	64	128	16	32	64	8	16	32
FVD (↓)	145.0	147.4	135.6	142.3	143.6	**135.0**	143.5	151.2	149.7	146.6
Final loss (↓)	0.0151	0.0150	0.0147	0.0143	0.0146	0.0143	**0.0141**	0.0146	0.0146	0.0142
Denoiser Params (M)	682.2	679.2	679.7	680.8	678.7	678.8	679.0	681.0	683.3	688.1
Denoiser FLOPs (TFLOPs)	1.089	1.083	1.084	1.086	1.082	1.082	1.082	1.086	1.091	1.101
Denoiser Latency (ms, ↓)	16.07	15.96	15.99	15.58	15.53	15.59	15.59	15.42	15.64	15.89
Denoiser Peak (GiB)	2.674	2.650	2.652	2.659	2.648	2.648	2.648	2.658	2.673	2.721
Emb. Params (M)	3.540	0.542	1.082	2.164	0.087	0.173	0.345	2.361	4.720	9.439
Emb. FLOPs (GFLOPs)	7.248	1.107	2.215	4.429	0.252	0.503	1.007	4.882	9.764	19.529
Emb. Latency (ms, ↓)	1.095	1.047	1.059	1.072	0.382	0.383	0.245	0.416	0.478	0.676
Emb. Peak (MiB)	89.257	51.817	56.630	68.254	34.669	35.333	37.536	47.508	78.008	136.510

patch-first embedding with $C_b = 32$ as the default configuration for JVT. Further patch embedding ablation details are provided in Appendix A.5.

Register tokens and guidance. In Tab. 3, we ablate both the information carried by register tokens and the effect of register guidance (RG) in class-conditioned generation. Under the default 50-step Euler sampling setup, adding registers without RG yields no FVD improvement over the register-free baseline. With RG, however, all tested register variants outperform this baseline, regardless of the conditioning information they carry. The 50-step Heun and 100-step Euler results further show

Table 3: **Register and sampling ablations for class-conditioned video generation on UCF-101.** NFE denotes the number of function evaluations, and No reg. denotes the register-free baseline.

Guidance		Sampling			FVD@2048 (↓)				
CFG	RG	Solver	Steps	NFE	No reg.	$e_r = 0$	$e_r = e_y$	$e_r = e_t$	$e_r = e_y + e_t$
✓	✗	Euler	50	100	**136.5**	137.9	138.6	<u>137.2</u>	<u>140.4</u>
✓	✗	Euler	100	200	<u>145.2</u>	<u>135.1</u>	<u>136.8</u>	138.0	142.5
✓	✗	Heun	50	200	167.4	158.7	149.1	157.2	160.8
✓	✓	Euler	50	150	–	**129.2**	**130.4**	**133.3**	**132.0**

that RG retains its advantage over higher-NFE baselines using more sampling steps or a higher-order solver, consistent with findings of prior works (Reynaud et al., 2024; Lu et al., 2025; Salajegheh & Mudur, 2026).

Although prior studies demonstrate the effectiveness of in-context class conditioning for pixel-space image generation (Starodubcev et al., 2026; Li & He, 2026; Lu et al., 2026), we observe no clear benefit from any of the three variants of register-based conditioning in Tab. 3. Instead, registers with $e_r = 0$ achieve the lowest FVD of 129.2. These results suggest that, under this sampling setup, the benefits of register tokens depend on RG, rather than register-based conditioning.

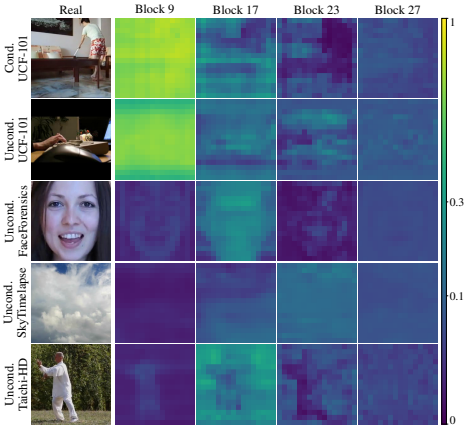


Figure 4: **Patch-to-register attention interactions across Transformer blocks.** Register tokens are prepended to visual tokens after the first $L_1 = 9$ blocks.

As shown in Tab. 1, RG works well on UCF-101, but brings no obvious gains on the other 3 datasets. Without register guidance, JVT variants with and without registers exhibit broadly similar FVD scores and intermediate feature-map smoothness. However, as illustrated in Fig. 4, visual tokens attend strongly to register tokens immediately after their insertion ($L_1 = 9$) on UCF-101, while this interaction is noticeably weaker on FaceForensics, SkyTimelapse, and Taichi-HD.

This dataset-dependent difference may be attributed to the global semantics encoded by register tokens (Starodubcev et al., 2026). UCF-101 spans 101 action categories and exhibits greater semantic diversity than the other three datasets. We hypothesize that this diversity, together with stronger patch-to-register interactions, makes the register-enabled prediction v_r a more informative guide for its register-free counterpart v_{nr}. In contrast, the more homogeneous semantics of the other datasets may limit the benefits of RG.

Adaptive layer normalization. Standard DiTs inject noise level and class information through adaptive layer normalization (AdaLN) (Peebles & Xie, 2023). However, recent works start to explore AdaLN-free DiTs for image generation, using in-context tokens, cross-attention, or multimodal joint attention for conditioning (Geng et al., 2026; Lu et al., 2026; Zeng et al., 2026; Wang et al., 2026b). In Tab. 4, we examine whether register-based conditioning offers a reliable alternative to AdaLN in JVT. Relative to the baseline, neither adding class-conditioned registers nor removing noise level conditioning improves FVD. With class-only conditioning, replacing AdaLN with register tokens further worsens FVD to 156.3. We also observe

Table 4: **Ablation study of AdaLN on UCF-101.** We report class-conditioned FVD@2048. – denotes removal of AdaLN or register tokens.

Variant	AdaLN condition	e_r	FVD (↓)
baseline	$e_t + e_y$	–	**136.5**
w/ $e_r = e_y$	$e_t + e_y$	e_y	<u>138.6</u>
w/o noise level	e_y	–	140.6
w/o AdaLN	–	e_y	156.3



Figure 5: **Joint RGB-disparity generation and point-cloud reconstruction on RealEstate10K**. We show generated RGB frames alongside aligned disparity maps. Colored point clouds are obtained by unprojecting depth recovered from disparity using the camera parameters, without an additional reconstruction model at inference.

no clear efficiency gain from removing AdaLN. Unconditional ablations show no consistent FVD improvement from jointly removing noise level conditioning and AdaLN (Appendix A.5, Tab. 10). These results support retaining standard AdaLN as JVT's default conditioning mechanism.

4.3 Camera-conditioned Video Generation

We instantiate the pixel-aligned joint generation framework in Sec. 3.4 for camera-conditioned video generation. Given an initial RGB frame and a target camera trajectory, we jointly generate RGB frames and aligned disparity maps. After recovering depth from disparity, the camera parameters allow the predictions to be unprojected into a sequence of colored point clouds, providing an explicit 4D representation of the generated scene, as shown in Fig. 5.

Experimental setting. We conduct 8-frame camera-conditioned video generation experiments on RealEstate10K (Zhou et al., 2018) at a resolution of 256×256. As dense depth annotations are unavailable, a frozen lightweight 3D reconstruction model, Pi3X (Wang et al., 2026d) supplies depth targets online during training. Following prior works (Song et al., 2026; Zhu et al., 2025), we reversibly transform depth D to square-root disparity $d = (D + 1)^{-1/2}$ and concatenate it with RGB as a four-channel target.

Evaluation. We evaluate RGB video quality using FVD, and assess camera following by comparing Pi3X-estimated relative poses with the target camera trajectory. Following prior works (Wang et al., 2024b; 2025a; 2026d), we report relative rotation accuracy (RRA), relative translation accuracy (RTA), and AUC as camera-following metrics. Specifically, RRA and RTA measure the percentages of evaluated frame pairs whose rotation and translation-direction errors fall below a given threshold, respectively, while AUC summarizes their joint error distribution. We use thresholds of 5° and 30°.

We compare JVT with pixel-space U-ViT models (Song et al., 2025; Wu et al., 2026) and a latent-space DiT (Huang et al., 2026b). As shown in Tab. 5, both JVT variants achieve lower FVD than these baselines. Joint RGB-disparity generation achieves the highest RTA and AUC among all compared methods, with only a slight increase in FVD relative to RGB-only JVT. The qualitative results in Fig. 6 further illustrate the differences in camera following.

Across methods, RRA is consistently high, whereas RTA remains substantially lower. This gap highlights a key limitation of these camera-conditioned video generators: they can follow prescribed camera rotations relatively reliably, but are less accurate at reproducing camera translations. Joint RGB-disparity generation narrows this gap, suggesting that explicit depth cues help the model relate

Table 5: **Camera-conditioned video generation on RealEstate10K.** FVD evaluates RGB video quality. Relative rotation accuracy (RRA), relative translation accuracy (RTA), and AUC evaluate agreement between the generated video and the target camera trajectory.

Method	Objective	FVD (↓)	RRA (↑)		RTA (↑)		AUC (↑)	
			@5	@30	@5	@30	@5	@30
DFoT (Song et al., 2025)	RGB	35.3	**99.9**	100	77.7	97.4	58.3	87.2
Geometry Forcing (Wu et al., 2026)	RGB	46.8	99.8	100	73.6	96.4	52.1	84.7
LIVE (Huang et al., 2026b)	RGB	42.5	99.1	100	49.8	92.6	28.4	73.3
JVT (Ours)	RGB	**27.0**	**99.9**	100	72.4	96.0	52.8	84.1
JVT (Ours)	RGB-disparity	27.5	**99.9**	100	**79.5**	**98.4**	**58.6**	**88.6**

camera motion to scene geometry. Within a unified pixel-space generative framework, JVT natively supports this joint modeling of RGB and geometry and improves camera following.



Figure 6: **Qualitative comparison of camera-conditioned video generation on RealEstate10K**, with four selected frames per scene. GT denotes ground truth and GF denotes Geometry Forcing. JVT and JVT w/ disp. denote the RGB-only and joint RGB-disparity variants, respectively.

5 Conclusion

We demonstrate pixel-space video generation with a simple DiT-style Transformer, achieving generation quality competitive with latent diffusion baselines in both class-conditioned and unconditional video generation benchmarks. Our ablations study suggests patch-first embedding for spatiotemporal tokenization and standard AdaLN for conditioning. For the register mechanism, register-free JVT provides an acceptable baseline, while register guidance can offer additional gains, particularly on the more diverse UCF-101 dataset.

Pixel-space modeling makes pixel-aligned joint generation straightforward by representing RGB and aligned modalities on a common spatiotemporal grid, without encoding them into a shared pretrained latent space. When extending JVT to camera-conditioned joint RGB-disparity generation, we find camera following is improved relative to RGB-only baselines, with a small FVD trade-off.

These findings highlight the potential of simple DiT-style Transformers for pixel-space video generation and pixel-aligned joint modeling. We hope our empirical findings help guide future efforts to scale pixel-space video generation to larger model size, more datasets, longer videos and higher resolutions.

AI USE STATEMENT

In this work, we used generative AI tools for [assist with translation]. We have not used generative AI tools for [help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research, methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, interpret results], and [Generate synthetic data sets, assist in the writing of proofs] are not applicable to this work. Additionally, we used generative AI tools for [sourcing/searching for information, grammar correction, polishing writing]. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyxQzBceg.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6836–6846, October 2021.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, 2024.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-JEPA: Latent video prediction for visual representation learning, 2024. URL https://openreview.net/forum?id=WFYbBOEOtv.

Mustapha Bounoua, Giulio Franzese, and Pietro Michiardi. Multi-modal latent diffusion. *Entropy*, 26(4):320, 2024.

Junyi Chen, Haoyi Zhu, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, and Tong He. Deepverse: 4d autoregressive video generation as a world model. *arXiv preprint arXiv:2506.01103*, 2025a.

Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. Pixelflow: Pixel-space generative models with flow. *arXiv preprint arXiv:2504.07963*, 2025b.

Zhennan Chen, Junwei Zhu, Xu Chen, Jiangning Zhang, Xiaobin Hu, Hanzhen Zhao, Chengjie Wang, Jian Yang, and Ying Tai. Dip: Taming diffusion models in pixel space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36136–36146, June 2026.

Yu Cheng and Fajie Yuan. Leanvae: An ultra-efficient reconstruction vae for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15692–15702, October 2025.

Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Katherine Crowson, Stefan Andreas Baumann, Alex Birch, Tanishq Mathew Abraham, Daniel Z Kaplan, and Enrico Shippole. Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9550–9575. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/crowson24a.html.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, Jun Xiao, and Long Chen. Ca2-VDM: Efficient autoregressive video diffusion model with causal generation and cache sharing. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=YbtH1aoE1V.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30467–30476, June 2026.

Sahil Goyal, Debapriya Tula, Gagan Jain, Pradeep Shenoy, Prateek Jain, and Sujoy Paul. Masked generative nested transformers with decode time scaling. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=h4lZMgGyAz.

Sahil Goyal, Swayam Agrawal, Gautham Govind Anil, Prateek Jain, Sujoy Paul, and Aditya Kusupati. Elt: Elastic looped transformers for visual generation. *arXiv preprint arXiv: 2604.09168*, 2026.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Joshua M. Susskind, and Navdeep Jaitly. Matryoshka diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tOzCcDdH9O.

Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv: 2503.19325*, 2025.

Zipeng Guo, Lichen Ma, Yu He, Xiaolong Fu, Jingling Fu, Junshi Huang, and Yan Li. Pixelu: A u-shaped transformer for efficient end-to-end pixel diffusion. *arXiv preprint arXiv: 2606.27760*, 2026.

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

Yu He, Lichen Ma, Zipeng Guo, Xinyuan Shan, Jingling Fu, Dong Chen, Junshi Huang, and Yan Li. Hyperdit: Hyper-connected transformers for high-fidelity pixel-space diffusion. *arXiv preprint arXiv:2605.15741*, 2026.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David Fleet. Video diffusion models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646. Curran Associates, Inc., 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/39235c56aef13fb05a6adc95eb9d8d66-Paper-Conference.pdf.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=rB6TpjAuSRy.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13213–13232. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/hoogeboom23a.html.

Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion: 1.5 fid on imagenet512 with pixel-space diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18062–18071, June 2025.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv: 1704.04861*, 2017.

Jinyi Hu, Shengding Hu, Yuxuan Song, Yufei Huang, Mingxuan Wang, Hao Zhou, Zhiyuan Liu, Wei-Ying Ma, and Maosong Sun. ACDit: Interpolating autoregressive conditional modeling and diffusion transformer. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=OuFNXESoCO.

Jiehui Huang, Yuechen Zhang, Xu He, Yuan Gao, Zhi Cen, Bin Xia, Yan Zhou, Xin Tao, Pengfei Wan, and Jiaya Jia. Unityvideo: Unified multi-modal multi-task learning for enhancing world-aware video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4471–4481, June 2026a.

Junchao Huang, Ziyang Ye, Xinting Hu, Tianyu He, Guiyu Zhang, Shaoshuai Shi, Jiang Bian, and Li Jiang. LIVE: Long-horizon interactive video world modeling. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=iKKZmgxkbG.

Dengyang Jiang, Ruoyi Du, Zhennan Chen, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Xiangpeng Yang, Huanqia Cai, Aiming Hao, Yuming Jiang, Peng Gao, Harry Yang, and Steven Hoi. An empirical study of training pixel-space text-to-image diffusion models. *arXiv preprint arXiv: 2608.16887*, 2026.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577. Curran Associates, Inc., 2022. doi: 10.52202/068431-1926. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a98846e9d9cc01cfb87eb694d946ce6b-Paper-Conference.pdf.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502, June 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv: 2412.03603*, 2024.

Minh Khoa Le, Kien Do, Duc Thanh Nguyen, and Truyen Tran. Framedit: Diffusion transformer with matrix attention for efficient video generation. *arXiv preprint arXiv: 2603.09721*, 2026.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18262–18272, October 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36115–36125, June 2026.

Mingfeng Lin, Jiakun Chen, Liang Han, and Liqiang Nie. Frepix: Frequency-heterogeneous flow matching for pixel-space image generation. *arXiv preprint arXiv:2605.06421*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Yaofang Liu, Yumeng Ren, Xiaodong Cun, Aitor Artola, Yang Liu, Tieyong Zeng, Raymond H. Chan, and Jean michel Morel. Redefining temporal modeling in video diffusion: The vectorized timestep approach. *arXiv preprint arXiv: 2410.03160*, 2024.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025.

Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. VDT: General-purpose video diffusion transformers via mask modeling. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Un0rgm9f04.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=WUK8JIeetF.

Lichen Ma, Zipeng Guo, Yu He, Xiaolong Fu, Luohang Liu, Jingling Fu, Junshi Huang, and Yan Li. Frequencybooster: Full-frequency modeling for high-fidelity pixel diffusion. *arXiv preprint arXiv:2605.17759*, 2026a.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research*, 2025.

Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. Deco: Frequency-decoupled pixel diffusion for end-to-end image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43600–43610, June 2026b.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Kangfu Mei and Vishal M. Patel. VIDM: video implicit diffusion models. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 9117–9125. AAAI Press, 2023. doi: 10.1609/AAAI.V37I8.26094. URL https://doi.org/10.1609/aaai.v37i8.26094.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Peijie Qiu, Wenhui Zhu, Sayantan Kumar, Xiwen Chen, Jin Yang, Xiaotong Sun, Abolfazl Razi, Yalin Wang, and Aristeidis Sotiras. Multimodal variational autoencoder: A barycentric view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20060–20068, 2025.

Hadrien Reynaud, Matthew Baugh, Mischa Dombrowski, Sarah Cechnicka, Qingjie Meng, and Bernhard Kainz. Jvid: Joint video-image diffusion for visual-quality and temporal-consistency in video generation. *arXiv preprint arXiv:2409.14149*, 2024.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv: 1803.09179*, 2018.

Farzad Salajegheh and Sudhir Mudur. When higher order hurts: Pre-asymptotic order collapse in generative ode sampling: A theory of discretization–learning interaction. In *European Conference on Computer Vision*, pp. 151–167. Springer, 2026.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Mostgan-v: Video generation with temporal motion styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5652–5661, June 2023.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/31c0b36aef265d9221af80872ceb62f9-Paper.pdf.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3626–3636, June 2022.

Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, and Sergey Tulyakov. Hierarchical patch diffusion models for high-resolution video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7569–7579, June 2024.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=j8Vr3E3vhy.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Ziyang Song, Zerong Wang, Bo Li, Hao Zhang, Ruijie Zhu, Li Liu, Peng-Tao Jiang, and Tianzhu Zhang. Depthmaster: Taming diffusion models for monocular depth estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2026. doi: 10.1109/TCSVT.2026.3681436.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv: 1212.0402*, 2012.

Nikita Starodubcev, Ilia Sudakov, Ilya Drobyshevskiy, Artem Babenko, and Dmitry Baranchuk. Registers matter for pixel-space diffusion transformers. *arXiv preprint arXiv: 2605.16147*, 2026.

Mingzhen Sun, Weining Wang, Zihan Qin, Jiahui Sun, Sihan Chen, and Jing Liu. Glober: Coherent non-autoregressive video generation via global guided video decoder. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 76120–76136. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/efe36e55d80a94d1726f660b8d237a0f-Paper-Conference.pdf.

Mingzhen Sun, Weining Wang, Gen Li, Jiawei Liu, Jiahui Sun, Wanquan Feng, Shanshan Lao, Siyu Zhou, Qian He, and Jing Liu. Ar-diffusion: Asynchronous video generation with auto-regressive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7364–7373, June 2025.

Jiayan Teng, Wendi Zheng, Ming Ding, Wenyi Hong, Jianqiao Wangni, Zhuoyi Yang, and Jie Tang. Relay diffusion: Unifying diffusion process across resolutions for image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=qTlcbLSm4p.

Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6puCSjH3hwA.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/416f9cb3276121c42eebb86352a4354a-Paper-Conference.pdf.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

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

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 (CVPR)*, pp. 5294–5306, June 2025a.

Junke Wang, Yi Jiang, Zehuan Yuan, Binyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 28281–28295. Curran Associates, Inc., 2024a. doi: 10.52202/079017-0887. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/31994923f58ae5b2d661b300bd439107-Paper-Conference.pdf.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14560, June 2023.

Shuai Wang, Ziteng Gao, Chenhui Zhu, Weilin Huang, and Limin Wang. Pixnerd: Pixel neural field diffusion. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=BDnOrExHmt.

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 (CVPR)*, pp. 20697–20709, June 2024b.

Xianbang Wang, Hanhong Zhao, Yiyang Lu, Kangyang Zhou, Linrui Ma, and Kaiming He. Minit2i: A minimalist baseline for text-to-image generation, 2026b. URL https://peppaking8.github.io/#/post/minit2i.

Yan Wang, Shijie Zhao, Junlin Li, and Li zhang. Eliminating VAE for fast and high-resolution generative detail restoration. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=1uu4Hg2Nlk.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *The Fourteenth International Conference on Learning Representations*, 2026d. URL https://openreview.net/forum?id=DTQIjngDta.

Yuqing Wang, Shuhuai Ren, Zhijie Lin, Yujin Han, Haoyuan Guo, Zhenheng Yang, Difan Zou, Jiashi Feng, and Xihui Liu. Parallelized autoregressive visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12955–12965, June 2025b.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ULXYZCms41.

Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv: 2104.10157*, 2021.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10459–10469, June 2023a.

Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David Ross, and Lu Jiang. Language model beats diffusion - tokenizer is key to visual generation. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 765–783, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/036912a83bdbb1fd792baf6532f102d8-Paper-Conference.pdf.

Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Czsdv-S4-w9.

Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18456–18466, June 2023b.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. Pixeldit: Pixel diffusion transformers for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14273–14282, June 2026.

Boya Zeng, Tianze Luo, Shu Pu, Jucheng Shen, Taiming Lu, Gabriel Sarch, and Zhuang Liu. i1: A simple and fully open recipe for strong text-to-image models. *arXiv preprint arXiv:2606.11289*, 2026.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0u1LigJaab.

Deyu Zhou, Quan Sun, Yuang Peng, Kun Yan, Runpei Dong, Duomin Wang, Zheng Ge, Nan Duan, and Xiangyu Zhang. Taming teacher forcing for masked autoregressive video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7374–7384, June 2025.

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

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8535–8546, October 2025.

A Appendix

A.1 Implementation Details

Our implementation follows the pixel-space clean-data prediction design of JiT (Li & He, 2026), with video generation and camera conditioning implemented in the DFoT framework (Song et al., 2025). Table 6 specifies reference model and optimization configurations.

Table 6: **Reference model and optimization configurations.** The standard-generation column specifies the selected reference recipe, and the camera columns summarize the archived RGB and RGB–disparity configurations. Registers vary across ablations. Camera batch sizes are the configured per-rank batch multiplied by the recorded world size, without gradient accumulation.

	Standard RGB	**Camera RGB**	**Camera RGB–disp.**
Backbone	JVT-XL	JVT-L	JVT-L, joint I/O
Depth / width / heads	28 / 1,152 / 16	24 / 1,024 / 16	24 / 1,024 / 16
Video size	$16 \times 256 \times 256$	$8 \times 256 \times 256$	$8 \times 256 \times 256$
Input/output channels	3	3	4
Spatial patch size	16×16	16×16	16×16
Temporal factor	4	1	1
Visual tokens	1,024	2,048	2,048
Embedding	Patch-first, $b = 32$	-	-
Registers, when used	32, after block 9	None	None
Optimizer	AdamW	AdamW	AdamW
Adam β	$(0.9, 0.99)$	$(0.9, 0.99)$	$(0.9, 0.99)$
Learning rate	10^{-4}	8×10^{-5}	8×10^{-5}
Warm-up updates	4,000	4,000	1,000
LR after warm-up	Constant	Constant	Constant
Weight decay	0.01	0.01	0.01
EMA decay	0.9999	0.9999	0.9999
Configured global batch	256	256	256
Noise shift s	4	4	4
Training noise levels	Shared across a clip	Frame-wise independent	Frame-wise independent
Prediction / loss	Clean data / velocity MSE	Clean data / velocity MSE	Clean data / velocity MSE
Division floor t_ϵ	0.005	0.005	0.005

A.2 Dataset Settings

Standard video-generation datasets. UCF-101, FaceForensics, SkyTimelapse, and Taichi-HD use 16-frame clips with frame interval three, spanning 46 source frames, at 256×256 input resolution. UCF-101 uses the combined training and test videos, while the other datasets use their training splits. No additional image-training dataset is used for these standard-generation recipes.

RealEstate10K clips and augmentation. The camera configurations use eight-frame 256×256 clips, one context frame, and a nominal training frame interval of 20 with temporal augmentation. The RGB and RGB–disparity configurations both specify horizontal flipping and temporal reversal with probability 0.5, back-and-forth augmentation with probability 0.1.

Disparity targets. A frozen Pi3X processes the RGB clip, resized by area interpolation to 252×252, together with intrinsics and camera-to-world pose priors. Its camera-space depth, taken from the third coordinate of local points, is resized bilinearly to 256×256 as the disparity labels.

A.3 Evaluation Settings

Sampling and Guidance The reference sampling configuration uses Euler solver with 50 steps (Euler-50) and noise shift four. The register ablation also evaluates Euler-100 and Heun-50. These choices are not uniform across all main results.

We count denoiser-forward equivalents, where a batch containing two conditioning branches counts as two forwards. Euler normally uses one, two, two, and three equivalents per step for no guidance, CFG, nondegenerate RG, and CFG+RG, respectively. Heun doubles these counts in the present implementation.

Camera-conditioned Generation and Geometry We use vanilla history guidance with a scale of 4, while retaining the target camera trajectory in both history branches. The context frame is restored after sampling. Generated disparity is analytically converted to 3D geometry by unnormalizing and clipping it to $10^{-5}, 1$, computing depth as $D = d^{-2} - 1$, and unprojecting it to world coordinates using the prescribed cameras. No learned reconstruction model is used in this post-processing step.

Efficiency Evaluation Details The efficiency experiments use JVT-XL with 28 Transformer blocks, width 1,152, 16 attention heads, and spatial patch size 16. Each case processes one 16-frame 256×256 video in pixel space. With a temporal downsampling factor of four, the denoiser receives 1,024 visual tokens. The register-enabled configuration inserts 32 class-initialized in-context tokens after the first 9 blocks, resulting in 1,056 tokens for the remaining blocks.

Sampling uses Euler integration with 50 steps. Register-enabled cases include no guidance, CFG, RG, and CFG+RG, while register-disabled cases include no guidance and CFG. These four guidance settings use one, two, two, and three denoiser-forward equivalents per step, respectively.

All measurements use bfloat16 with TF32 enabled and `torch.compile` disabled on one NVIDIA 4090 GPU. Each case runs in a fresh Python process. Denoiser latency uses 20 warm-up iterations followed by 100 timed iterations, and sampling latency uses 5 warm-up samples followed by 20 timed samples. The complete matrix is repeated in three independent sweeps. We record model parameters, FLOPs, baseline and peak allocated GPU memory, incremental peak memory, latency statistics, and throughput. FLOPs for one denoiser forward pass are counted with `torch.utils.flop_counter.FlopCounterMode`. Sampling FLOPs are computed from the matching denoiser FLOPs, 50 sampling steps, and the number of denoiser-forward equivalents per step.

A.4 Efficiency Results

The following tables use the protocol in Sec. A.3. Table 7 covers the ten embedding configurations in Tab. 2, with additional register-free measurements for five of them. Table 8 reports the corresponding 30 sampling measurements.

Table 7: **Complete JVT-XL denoiser efficiency results on one NVIDIA 4090 GPU.** Embedding FLOPs count the convolution operations in the patch embedder. The embedding share is their percentage of denoiser FLOPs. Peak denotes allocated GPU memory. C_b is the bottleneck width.

Embedding	C_b	In-context	Tokens	Params (M)	FLOPs			Peak (GiB)
					Total (TFLOPs)	Embed. (GFLOPs)	Embed. share (%)	
Tubelet	–	yes	1056	682.2	1.089	7.248	0.666	2.674
Tubelet	32	yes	1056	679.2	1.083	1.107	0.102	2.650
Tubelet	64	yes	1056	679.7	1.084	2.215	0.204	2.652
Tubelet	128	yes	1056	680.8	1.086	4.429	0.408	2.659
Patch-first	16	yes	1056	678.7	1.082	0.252	0.023	2.648
Patch-first	32	yes	1056	678.8	1.082	0.503	0.047	2.648
Patch-first	64	yes	1056	679.0	1.082	1.007	0.093	2.648
Frame-first	8	yes	1056	681.0	1.086	4.882	0.449	2.658
Frame-first	16	yes	1056	683.3	1.091	9.764	0.895	2.673
Frame-first	32	yes	1056	688.1	1.101	19.529	1.774	2.721
Tubelet	–	no	1024	682.1	1.063	7.248	0.682	2.674
Tubelet	32	no	1024	679.1	1.057	1.107	0.105	2.649
Tubelet	64	no	1024	679.7	1.058	2.215	0.209	2.651
Patch-first	32	no	1024	678.8	1.057	0.503	0.048	2.648
Frame-first	32	no	1024	688.0	1.076	19.529	1.815	2.721

Table 8: **Complete 50-step sampling efficiency results on one NVIDIA 4090 GPU.** "Fwd." is the number of denoiser-forward equivalents per step. Latency is milliseconds per generated $16 \times 256 \times 256$ video and is reported as mean $\pm$ standard deviation across three sweeps. FLOPs are nominal estimates based on 50 steps and the listed forward equivalents. Register-free branch costs are approximated by the listed configuration's denoiser cost.

Embedding	In-context	Guidance	Fwd.	Latency (ms)	Peak (GiB)	Est. FLOPs (TFLOPs)
Tubelet (plain)	yes	None	1	802.3 ± 1.3	2.689	54.4
Tubelet (plain)	yes	CFG	2	1280.8 ± 1.9	2.751	108.9
Tubelet (plain)	yes	RG	2	1529.9 ± 0.3	2.707	108.9
Tubelet (plain)	yes	CFG+RG	3	2029.3 ± 2.5	2.768	163.3
Tubelet (plain)	no	None	1	739.3 ± 0.8	2.689	53.2
Tubelet (plain)	no	CFG	2	1190.5 ± 2.4	2.751	106.3
Tubelet $C_b = 32$	yes	None	1	788.2 ± 0.9	2.662	54.1
Tubelet $C_b = 32$	yes	CFG	2	1229.7 ± 1.9	2.724	108.3
Tubelet $C_b = 32$	yes	RG	2	1544.9 ± 1.9	2.683	108.3
Tubelet $C_b = 32$	yes	CFG+RG	3	1992.3 ± 1.6	2.747	162.4
Tubelet $C_b = 32$	no	None	1	743.9 ± 2.1	2.662	52.9
Tubelet $C_b = 32$	no	CFG	2	1168.5 ± 1.1	2.723	105.7
Tubelet $C_b = 64$	yes	None	1	790.3 ± 1.9	2.663	54.2
Tubelet $C_b = 64$	yes	CFG	2	1238.7 ± 0.7	2.726	108.4
Tubelet $C_b = 64$	yes	RG	2	1532.5 ± 1.7	2.685	108.4
Tubelet $C_b = 64$	yes	CFG+RG	3	1978.1 ± 3.3	2.749	162.5
Tubelet $C_b = 64$	no	None	1	729.9 ± 1.6	2.663	52.9
Tubelet $C_b = 64$	no	CFG	2	1158.2 ± 1.1	2.726	105.8
Patch-first $C_b = 32$	yes	None	1	767.8 ± 1.5	2.661	54.1
Patch-first $C_b = 32$	yes	CFG	2	1199.6 ± 1.6	2.721	108.2
Patch-first $C_b = 32$	yes	RG	2	1498.1 ± 1.3	2.682	108.2
Patch-first $C_b = 32$	yes	CFG+RG	3	1937.0 ± 1.2	2.746	162.3
Patch-first $C_b = 32$	no	None	1	717.8 ± 0.4	2.660	52.8
Patch-first $C_b = 32$	no	CFG	2	1144.0 ± 1.1	2.721	105.7
Frame-first $C_b = 32$	yes	None	1	796.8 ± 1.5	2.737	55.0
Frame-first $C_b = 32$	yes	CFG	2	1245.8 ± 2.1	2.781	110.1
Frame-first $C_b = 32$	yes	RG	2	1507.7 ± 2.1	2.755	110.1
Frame-first $C_b = 32$	yes	CFG+RG	3	1964.4 ± 3.3	2.799	165.1
Frame-first $C_b = 32$	no	None	1	721.6 ± 1.2	2.737	53.8
Frame-first $C_b = 32$	no	CFG	2	1152.5 ± 0.3	2.780	107.6

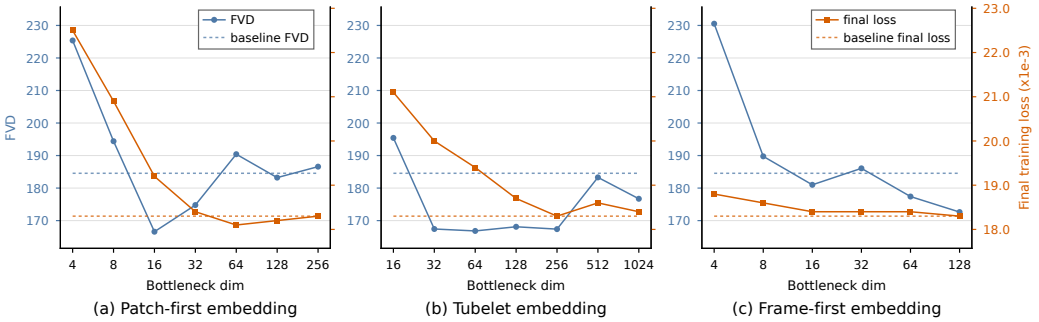


Figure 7: Effect of configured bottleneck width on video patch embedding with batch size 40. FVD uses the left axis and final training loss the right axis. Dashed lines show the corresponding plain-tubelet baseline values. The two vertical scales are independent.

A.5 Additional Ablations

Figure 7 gives the bottleneck width sweep curve. In this experiment, we use batch size of 40, whereas Table 2 reports full-scale runs with batch size 256. For tubelet and patch-first embedding, FVD varies non-monotonically with C_b: an overly narrow bottleneck may discard task-relevant spatiotemporal information, whereas an excessively wide bottleneck offers limited additional benefit. As for frame-first embedding, FVD improves progressively as C_b increases, but the efficiency metrics declines sharply.

Table 9: **Register-token and sampling ablation for unconditional video generation.** We report FVD@2048, lower is better. NFE denotes the number of function evaluations, and No reg. denotes the register-free baseline.

RG	NFE	Solver	Steps	UCF-101			FaceForensics			SkyTimelapse			Taichi-HD		
				No reg.	$e_r = 0$	$e_r = e_t$	No reg.	$e_r = 0$	$e_r = e_t$	No reg.	$e_r = 0$	$e_r = e_t$	No reg.	$e_r = 0$	$e_r = e_t$
✗	100	Heun	50	413.5	378.6	474.2	**32.9**	37.7	33.1	46.6	54.9	52.6	119.1	144.6	127.1
✗	200	Euler	200	421.2	392.4	484.5	35.2	40.2	34.9	49.0	58.3	55.3	114.5	143.6	124.3
✗	200	Heun	100	**387.7**	354.9	**451.5**	34.0	39.0	34.0	**45.5**	54.0	51.2	**114.4**	137.5	**122.0**
✓	200	Heun	50	–	**215.9**	462.4	–	**34.9**	**31.8**	–	**47.8**	**51.1**	–	**132.3**	124.3

Table 10: **Effect of timestep conditioning and adaptive layer normalization on unconditional video generation.** We report FVD@2048, where lower is better. For the AdaLN condition and e_r, – denotes the absence of the corresponding conditioning pathway.

Variant	AdaLN condition	e_r	FaceForensics	SkyTimelapse	Taichi-HD
baseline	e_t	–	32.9	46.6	119.1
w/ $e_r = e_t$	e_t	e_t	33.1	53.6	127.1
w/o AdaLN	–	–	40.6	60.8	118.2

A.6 Additional FVD Results

Training strategies and algorithm designs greatly affect model performance. Some methods adopt the image-video joint training strategy proposed in VDM (Ho et al., 2022b), which markedly improves video generation quality and can also be incorporated into our method. Tab. 11 compares our method without the joint training strategy against these methods. We further compare our method with recent hybrid approaches of autoregressive modeling and diffusion, and show the results in Tab. 12.

Different evaluation protocols can also directly influence the reported FVD scores. Following StyleGAN-V (Skorokhodov et al., 2022) and Latte (Ma et al., 2025), we compute FVD using

2048 generated video samples (FVD@2048), whereas several other methods use 10000 samples (FVD@10000). Since the original FVD paper (Unterthiner et al., 2019) shows that the number of samples used for evaluation can greatly affect the FVD score, we additionally report our FVD@10000 result in Tab. 13.

Table 11: **FVD@2048 on different datasets.** We report class-conditioned generation on UCF-101 and unconditional generation on other datasets. "+IMG" means video-image joint training.

Method	UCF101	FaceForensics	SkyTimelapse	Taichi-HD
Latte (Ma et al., 2025)	477.97	34.00	59.82	159.6
Latte+IMG (Ma et al., 2025)	202.23	27.08	42.67	97.1
FrameDiT+IMG (Le et al., 2026)	170.1	16.6	39.5	95.5
JVT (Ours)	136.5	32.86	46.58	119.1

Table 12: **FVD@2048 on UCF-101.** We compare class-conditioned and unconditional generation with hybrid AR+diffusion methods. All methods use $16 \times 256 \times 256$ resolution.

Method	FVD w/ class	FVD w/o class
AR-Diffusion (Sun et al., 2025)	186.6	-
Ca2-VDM (Gao et al., 2025)	184.5	-
ACDiT-XL (Hu et al., 2026)	111	-
MAGI (Zhou et al., 2025)	-	297.8
FAR-XL (Gu et al., 2025)	108	279
JVT (Ours)	136.5	215.9

Table 13: **FVD@10000 on UCF-101.** All methods use class condition.

Method	Resolution	FVD@10000
MAGVIT (Yu et al., 2023a)	$16 \times 128 \times 128$	76
MAGVIT-v2 (Yu et al., 2024)	$16 \times 128 \times 128$	58
MaGNeTS (Goyal et al., 2025)	$16 \times 128 \times 128$	96.4
PAR (Wang et al., 2025b)	$16 \times 128 \times 128$	94.1
ELT (Goyal et al., 2026)	$16 \times 128 \times 128$	60.8
JVT (Ours)	$16 \times 256 \times 256$	91.2

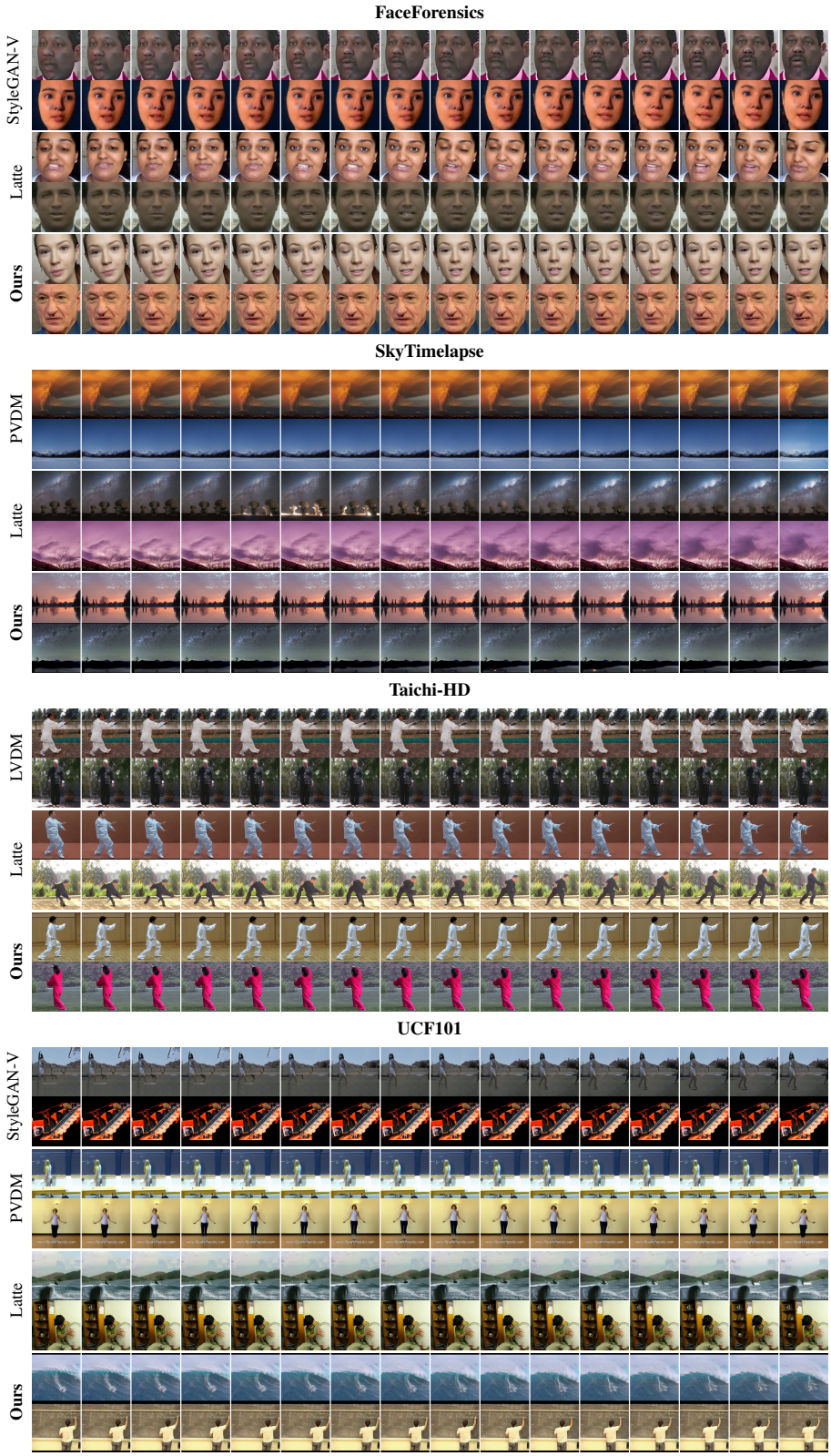


Figure 8: **Video generation results across four datasets.** Two 16-frame samples are shown for each method on FaceForensics, SkyTimelapse, Taichi-HD, and UCF-101.