TIV-Diffusion: Towards Object-Centric Movement for Text-driven Image to Video Generation

Xingrui Wang[1], **Xin Li**[1*], **Yaosi Hu**[2], **Hanxin Zhu**[1], **Chen Hou**[1], **Cuiling Lan**, **Zhibo Chen**[1*]
[1]University of Science and Technology of China
[2]The Hong Kong Polytechnic University
{wxrui_18264819595, hanxinzhu, houchen}@mail.ustc.edu.cn, youncyhu@gmail.com,
cuilinglan@outlook.com, {xin.li, chenzhibo}@ustc.edu.cn

Abstract

Text-driven Image to Video Generation (TI2V) aims to generate controllable video given the first frame and corresponding textual description. The primary challenges of this task lie in two parts: (i) how to identify the target objects and ensure the consistency between the movement trajectory and the textual description. (ii) how to improve the subjective quality of generated videos. To tackle the above challenges, we propose a new diffusion-based TI2V framework, termed TIV-Diffusion, via object-centric textual-visual alignment, intending to achieve precise control and high-quality video generation based on textual-described motion for different objects. Concretely, we enable our TIV-Diffuion model to perceive the textual-described objects and their motion trajectory by incorporating the fused textual and visual knowledge through scale-offset modulation. Moreover, to mitigate the problems of object disappearance and misaligned objects and motion, we introduce an object-centric textual-visual alignment module, which reduces the risk of misaligned objects/motion by decoupling the objects in the reference image and aligning textual features with each object individually. Based on the above innovations, our TIV-Diffusion achieves state-of-the-art high-quality video generation compared with existing TI2V methods.

Extended version — https://arxiv.org/abs/2412.10275

1 Introduction

Videos that effectively convey complex visual information play a crucial role in human life, including entertainment, education, and documentation (Aldausari et al. 2022). Recently, as automation advances, employing artificial intelligence algorithms for video generation has drawn numerous research attention. Early unconditional video generation (Vondrick, Pirsiavash, and Torralba 2016; Saito, Matsumoto, and Saito 2017; Tulyakov et al. 2018) relies solely on models to learn from unlabeled video data, yielding uncontrollable outcomes. In the pursuit of controllability, approaches such as Image-to-Video Generation (I2V) (Chang et al. 2021; Lee et al. 2021; Gupta et al. 2022) and Text-to-Video Generation (T2V) (Ho et al. 2022; Singer et al. 2022; Wu et al. 2022b) have come to the forefront, facilitating precise specification of appearance or movement within the generated videos. Controllable video generation enhances the way we express our intention, with diverse applications in creative content creation (Blattmann et al. 2023), data augmentation (Wang et al. 2023a), and various fields.

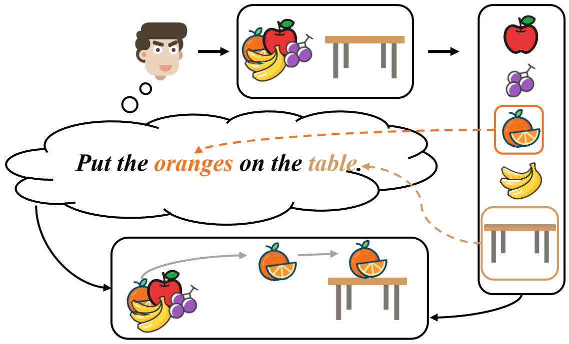


Figure 1: Humans naturally disentangle different objects in their environment. In the image above, the human wants to 'Put the oranges on the table.' First, he will decouple objects with different attributes, then find the target objects according to the intention and complete the action. Our model draws inspiration from this observation.

Compared with images, videos introduce the temporal dimension, which means that the unconditional video generation process should ensure the authenticity of each frame as well as maintain coherence between frames (Vondrick, Pirsiavash, and Torralba 2016; Sun et al. 2023; Tulyakov et al. 2018; Mei and Patel 2023). In contrast, controllable video generation is expected to consider control conditions while meeting the above-mentioned basic requirements. In particular, T2V utilizes text descriptions to specify the visual appearance and motion characteristics of the generated video (Wu et al. 2021, 2022a; Singer et al. 2022). Typically, the text descriptions are encoded using CLIP (Radford et al. 2021) and subsequently control the generation process via the cross-attention mechanism. While for I2V, the image or image sequence primarily restricts the visual aspects of the target video, the model should deduce subsequent object movements based on motion cues (Yang, Srivastava, and Mandt 2022; Tan et al. 2023b).

---

[*] Corresponding authors.

Although T2V and I2V yield impressive results, it is essential to acknowledge that text inherently harbors ambiguity, leading to T2V generating numerous videos that align with a text, and I2V having limited control over motion (Hu, Luo, and Chen 2022). To enhance generation controllability, we direct our attention to Text-Image-to-Video Generation (TI2V), a paradigm where the image shapes the content, and the text guides movement (Song et al. 2022). Nevertheless, the explorations of higher subjective quality in TI2V through diffusion models remain relatively few, and certain challenges hinder the quality of the generated videos. Firstly, given that the model processes text and image simultaneously, there emerges a necessity to align the appearance and motion of objects from different modalities. When the alignment quality is subpar, it will lead to instances where the moving objects do not correspond to the textual intended targets or their motion fails to adhere to the provided instruction (Hu, Luo, and Chen 2022; Xu et al. 2023; Hu, Luo, and Chen 2023). With an increased number of objects in the image, alignment difficulties are further exacerbated. Secondly, we observe that object overlap or occlusion during movement can result in deformations or disappearance of objects in subsequent video frames, as demonstrated in Sec. 4.2.

As illustrated in Fig. 1, motivated by the natural ability of individuals to decouple various objects within their visual field and correlate them with the corresponding text description, we propose a new framework for **TI2V**, dubbed **TIV-Diffusion**, leveraging object disentanglement to improve textual-visual alignment and subjective quality. Specifically, as shown in Fig. 2, TIV-Diffusion encodes the input image and text caption respectively, followed by a fusion of their encoded embeddings, and then modulates them into the autoregressive generation process in a SPADE (Park et al. 2019) manner, which integrates the appearance and motion information. Furthermore, TIV-Diffusion extracts object-centric representations (*i.e.*, slots) from the input image utilizing a Slot Attention (Locatello et al. 2020) encoder. This encoder aims to discover the latent compositional structure from unstructured observations (Jiang et al. 2023), and the resulting slots capture object attributes. To facilitate TIV-Diffusion in comprehending which objects to move and their final destinations, as depicted in Fig. 3, we align slots with the text caption individually. Disentangling objects enables precise alignment between objects and their motion description, fostering a heightened semantic consistency between the generated video and its corresponding text. Moreover, to mitigate object disappearance or deformation, we incorporate slots into the generation process. Hence, with the object attribute information within slots, TIV-Diffusion refines object composition by constant awareness of object features. However, considering that the object stored in each slot is unknown, we use Gumbel-Softmax (Jang, Gu, and Poole 2016) for adaptive slot selection, thereby improving the temporal consistency of objects.

The contributions of this paper are concluded as follows:

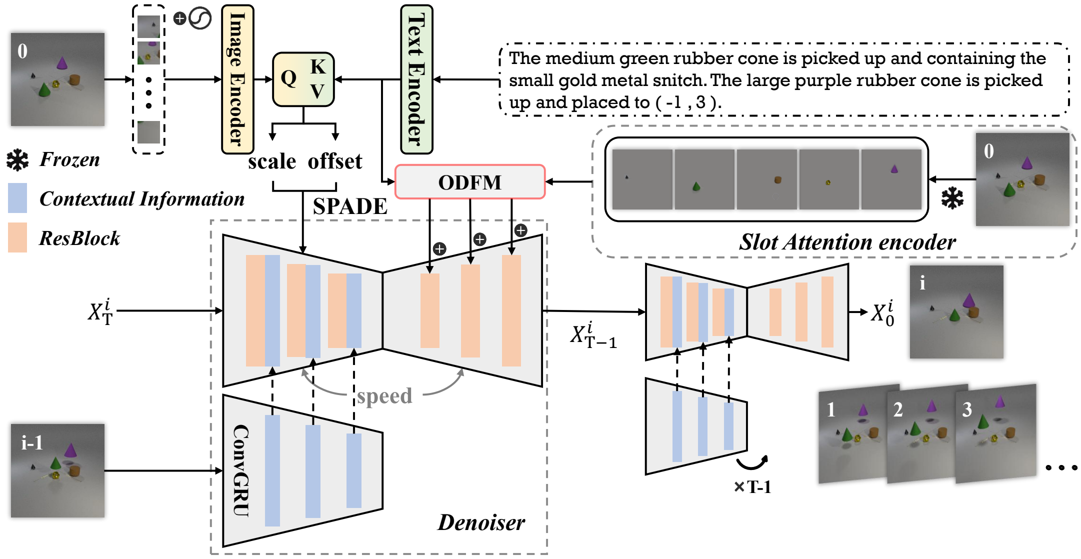
- We introduce a diffusion-based TI2V model that harnesses the potential of object disentanglement, which can generate video frames with high perceptual quality, as well as achieve better controllability.
- Our model extracts object features (*i.e.*, slots) through the Slot Attention encoder to enhance the alignment between text and objects, and incorporates slots during the generation process adaptively to mitigate the likelihood of object deformation and disappearance.
- We perform experiments on two categories of existing datasets, and the model performs well under various control conditions. Extensive results have demonstrated that our proposed method can achieve state-of-the-art performance.

2 Related Work

Text-Image-to-Video Generation. Text-Image-to-Video Generation achieves enhanced controllability through combining image and text. TVP (Song et al. 2022) explores the causality in the text description and then generates step-wise inference embeddings to guide the generation of each frame using GAN (Goodfellow et al. 2014). Concurrent work MAGE (Hu, Luo, and Chen 2022) employs VQ-VAE for frame quantization and leverages a transformer-based approach guided by a motion anchor. Furthermore, the MAGE+ (Hu, Luo, and Chen 2023) variant enhances performance by incorporating a robust compression autoencoder. MMVG (Fu et al. 2023) is similar to the former, except that it combines masks to achieve multi-tasks. TiV-ODE (Xu et al. 2023) considers temporal continuity and implements the Neural ODE approach, but the generated videos have residual shadows. LFDM (Ni et al. 2023) utilizes diffusion models to learn the distribution of optical flow sequences, which are used to synthesize single-object videos. Similar to the motivation of TVP, diffusion-based Seer (Gu et al. 2023) uses fine-grained textual sub-instructions for each frame generation. DynamiCrafter (Xing et al. 2023a), on the other hand, requires expensive computational costs for its implementation.

One of the primary challenges in the TI2V task is the alignment between the text caption and the image. When misaligned, the movement trajectory of the object will be semantically inconsistent with the text description. In contrast to prior works, we enhance this alignment by introducing object-centric features and achieve fine-grained control in a resource-friendly manner.

Other Controllable Video Generation. Text-to-Video Generation refers to the generation of videos guided by textual input. GODIVA (Wu et al. 2021), CogVideo (Hong et al. 2022), and NÜWA (Wu et al. 2022a) leverage codebooks from VQVAE or VQGAN for video content quantization. They employ sparse attention mechanisms to capture textual information and temporal relationships within the video efficiently. Meanwhile, diffusion models excel in image generation and have recently been extended to the realm of video generation (Singer et al. 2022; Zhou et al. 2022; Ho et al. 2022; Wu et al. 2022b; An et al. 2023; Wang et al. 2023b; Wang, Li, and Chen 2024). To adapt U-Net for video data, current works often incorporate temporal convolutions or temporal attention mechanisms while using cross-attention to introduce text guidance into the generation process (Singer et al. 2022; Ho et al. 2022; Zhou et al. 2022;



The medium green rubber cone is picked up and containing the small gold metal snitch. The large purple rubber cone is picked up and placed to (-1 , 3).

Figure 2: Illustration of the TIV-Diffusion framework. Given an image and its corresponding text caption, TIV-Diffusion can autoregressively generate subsequent video frames. To improve the alignment between objects' appearance and movement from different modalities, we introduce object-centric representations. We employ a slot attention encoder to extract slots and further fuse them with text information through the Object Distanglement Fusion Module (ODFM), which will be introduced thoroughly in Sec. 3.4. Additionally, the video generation process adaptively incorporates object attribute information from slots to address object deformation and disappearance.

Chen et al. 2023). Besides, some works choose to generate videos in the latent space, utilizing pre-trained Stable Diffusion (Rombach et al. 2022) model weights for initialization, and exclusively fine-tuning temporal layers for efficient training (An et al. 2023; Blattmann et al. 2023; Xing et al. 2023b; Qiu et al. 2023; Wang, Li, and Chen 2024; Wu et al. 2025).

Video-to-Video Generation focuses on predicting future video frames using motion cues from past frames. Recurrent-based models, including ConvLSTM (Shi et al. 2015) and ConvGRU (Ballas et al. 2015) structures, are extensively studied in this context. LMC-memory (Lee et al. 2021) employs memory to store long-term motion context and leverages this information to assist ConvLSTM in predicting future frames, while MAU (Chang et al. 2021) introduces a motion-aware unit to expand the temporal receptive field. Compared to recurrent-based models, recurrent-free models (Tan et al. 2023a,b), which process multiple frames concurrently, offer a trade-off between efficiency and performance. Furthermore, the Transformer (Vaswani et al. 2017) architecture is a promising option (Weissenborn, Täckström, and Uszkoreit 2019; He et al. 2022; Gupta et al. 2022; Sun et al. 2023). Diffusion models, exemplified by MCVD (Voleti, Jolicoeur-Martineau, and Pal 2022) and RaMViD (Höppe et al. 2022), excel in multi-tasks such as frame prediction and infilling. Unlike conventional methods that directly forecast video frames, RVD (Yang, Srivastava, and Mandt 2022) chooses to predict frame residuals.

3 Method

3.1 Preliminary

Diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Li et al. 2023; Ren et al. 2025) primarily acquire knowledge of the unknown data distribution through the dual processes of *diffusion* and *denoising*. *Diffusion* gradually disrupts data distribution with Gaussian noise, avoiding extra parameter training. This process can be represented as a Markov chain denoted by $q\left(X_t \mid X_{t-1}\right) = \mathcal{N}\left(X_t \mid \sqrt{1-\beta_t} X_{t-1}, \beta_t \mathrm{I}\right)$, where $\beta_t \in (0,1)$ and X_t represent the tth step of noise addition, with a total of T steps. *Denoising* is used for data structure reconstruction, and in DDPM (Ho, Jain, and Abbeel 2020), ϵ-prediction is introduced. This method involves learning the denoising function f_θ through mean square error loss minimization, represented as $L(\theta) = \mathbb{E}_{X_0,t,\epsilon}\left\|\epsilon - f_\theta\left(X_t, t\right)\right\|^2$. Following model training, desired data samples can be acquired by sampling Gaussian noise and continuously applying denoising processes.

3.2 Overall Framework

Given an image X^0 along with a text caption $S = \{s_1, \cdots, s_L\}$ of length L, the model aims to generate a new sequence of images $X^{1:N} = \left\{X^1, \cdots, X^N\right\}$. These generated images should maintain visual consistency with X^0 while adhering to the motion described in S, represented by the conditional distribution $p\left(X^{1:N} \middle| X^0, S\right)$.

Fig. 2 illustrates the overall architecture of the proposed TIV-Diffusion. Our method is founded on the diffusion model, which allows for controllable video generation through extensions discussed in Sec. 3.3. Previous works lack disentanglement of objects within the image, leading to suboptimal alignment between the text caption and the image. In Sec. 3.4, we incorporate object-centric representations to enhance this alignment.

3.3 Autoregressive TI2V Generation

To reduce computational demand, we employ an autoregressive generation approach here (Voleti, Jolicoeur-Martineau, and Pal 2022; Yang, Srivastava, and Mandt 2022).

Text-Image Fusion. The text encoder, comprised of trainable Transformer (Vaswani et al. 2017) encoder layers, establishes correlations within the input data while encoding the text S as $e_S \in \mathbb{R}^{L\times d}$, where d denotes the dimension of the text embeddings. The image, denoted as $X^0 \in \mathbb{R}^{H\times W\times C}$, is partitioned into non-overlapping patches of uniform size (Dosovitskiy et al. 2020), where H and W represent the height and width, and C is the number of channels. The image token is also encoded using Transformer encoder layers and position embeddings are learned during training. Since each patch has a size of $f_p \times f_p$, the encoding of X^0 is represented as $e_{X^0} \in \mathbb{R}^{\frac{H}{f_p}\times\frac{W}{f_p}\times d}$. Following this, information from both modalities is fused through cross-attention, with e_{X^0} as the Query and e_S as the Key and Value:

$$c = \mathrm{CrossAttention}\left(Q\left(e_{X^0}\right), K\left(e_S\right), V\left(e_S\right)\right). \tag{1}$$

Here, Q, K, and V are learnable linear projections. The fused information c aligns the textual description of objects with their respective counterparts in the image, specifying their motion. c will be incorporated into the downsampling section of Unet (Ronneberger, Fischer, and Brox 2015) in the SPADE (Park et al. 2019) format, where it is utilized for calculating multi-scale scaling and offset values to modulate the feature maps $\{\mathbf{F}^m_{\text{Unet}}\}^M_{m=1}$ within the residual blocks of Diffusion. M represents the number of down or up-sampling layers.

$$\begin{gathered}\hat{\mathbf{F}}^m_{\text{Unet}} = (1+\gamma^m)\odot \mathbf{F}^m_{\text{Unet}} + \beta^m,\\ \gamma^m, \beta^m = \mathcal{M}^m_\theta\left(c\right).\end{gathered} \tag{2}$$

In Eq. 2, the scale γ^m and offset β^m are the modulation parameters, $\mathcal{M}^m_\theta$ encompasses multiple convolutional layers and activation functions, and m denotes the current layer.

Text-Image-to-Video Generation. The video generator, an extension of DDPM (Ho, Jain, and Abbeel 2020), integrates a timing relationship capture module. Specifically, it conditions the generation of video frame X^n on $X^{<n}$ (all frames before n), and information extraction from $X^{<n}$ is achieved iteratively, using ConvGRU (Ballas et al. 2015). Initially, X^0 is processed through ConvGRU to extract contextual information, and then concatenated with $\left\{\hat{\mathbf{F}}^m_{\text{Unet}}\right\}^M_{m=1}$ at multiple scales to produce the video frame X^1. Similarly, X^1 is used as a condition to generate X^2 by passing it through ConvGRU for status updates. Continuing this process, each X^n generation involves sending X^{n-1} to the timing module to gather dynamic information from frame 0 to $n-1$. With the concurrent guidance from c and $X^{<n}$, we employ frame reconstruction loss (Zhou et al. 2022) directly to train the diffusion model:

$$\begin{gathered}\mathcal{L}\left(\theta\right) = \mathbb{E}_t \sum_{n=1}^{N} \left\| X^n_0 - f_\theta\left(X^n_t | t, c, X^{<n}, v\right)\right\|^2_2,\\ v = \phi\left(\eta\right).\end{gathered} \tag{3}$$

To enhance video diversity, we control objects' speed using the parameter η, encoding it with a linear layer ϕ, and introduce it via cross-attention.

3.4 Utilize Object Disentanglement for Improved Alignment

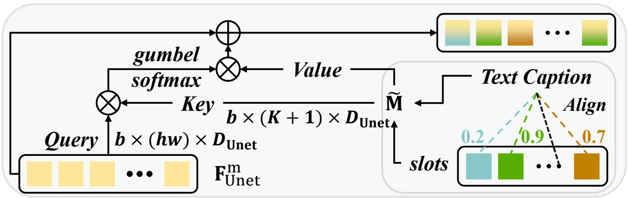


Figure 3: Object Distanglement Fusion Module (ODFM). We initially identify the relevant target moving objects within the text caption based on the provided slots and then align slots with the text caption individually. Subsequently, we utilize the text-enhanced slots to facilitate the video generation process.

Assuming there are K objects in the initial image X^0, along with the background, a total of $K+1$ corresponding learnable variables are required. We utilize the Slot Attention (Locatello et al. 2020) encoder to extract object features from X^0, as shown in Fig. 2. Specifically, the $K+1$ slots, each possessing dimension D_{slots}, are initialized using a Gaussian distribution featuring the learnable μ and σ, denoted as slots $\in \mathbb{R}^{(K+1)\times D_{\text{slots}}}$. Given an image X^0 and processed by CNNs (LeCun et al. 1989), we get inputs $\in \mathbb{R}^{N_{\text{inputs}}\times D_{\text{inputs}}}$. Subsequently, we employ Scaled Dot-Product Attention (Vaswani et al. 2017) $\mathrm{Softmax}\left(\frac{1}{\sqrt{D}}k\left(\text{inputs}\right)\cdot q\left(\text{slots}\right)^T, \text{axis} = \text{“slots”}\right)$ to foster competition among slots and iteratively update the content within them, thereby yielding object-centric representations. Before that, we should apply learnable linear transformations to map D_{inputs} and D_{slots} to a shared dimension D. The Slot Attention encoder is trained via image reconstruction:

$$\hat{X}^0 = Dec\left(Enc\left(X^0\right)\right). \tag{4}$$

Enc (*i.e.*, Slot Attention encoder) in this context encompasses both the CNNs responsible for feature extraction from X^0 and the slots update module, while *Dec* uses slots to reconstruct X^0 (Locatello et al. 2020). *Enc* and *Dec* are pre-trained before training the TI2V diffusion model. *Dec* is a CNN-structured auxiliary module designed to aid in the training of *Enc*, and it is omitted in Fig. 2. Following this, object-centric representations are obtained using Eq. 5:

$$\text{slots} = Enc\left(X^0\right). \tag{5}$$

Fig. 3 shows how object slots can be used to improve the effects of the generated video. The input text caption specifies the objects to be controlled. As a result, the currently obtained slots make it easier to extract relevant information from the text, improving the identification of the target objects. Additionally, slots also store object attribute information. Combining the above two aspects, we get Eq. 6:

$$M = \text{slots} + \text{CrossAttention}\left(Q\left(\text{slots}\right), K\left(e_S\right), V\left(e_S\right)\right). \tag{6}$$

We align slots with the text caption individually, and the slots augmented with textual information are denoted as $M \in \mathbb{R}^{(K+1)\times D}$. The experiments in Sec. 4 demonstrate that M effectively enhances text-image alignment and mitigates object deformation. Relying on similarity within the embedding space, we enable the diffusion model to autonomously select suitable slots, thereby enhancing the generation of video frames. In order to match the inner dimension D_{Unet} of Unet, we first linearly project M to $\tilde{M} = \text{Linear}\left(M\right) \in \mathbb{R}^{(K+1)\times D_{\text{Unet}}}$. Gumbel-Softmax (Jang, Gu, and Poole 2016) is used for interaction between $\mathbf{F}^m_{\text{Unet}}$ and $\tilde{M}$ to introduce randomness and allow the model to explore during the training process. Due to the unpredictable assignment of objects to slots, this exploration process increases model diversity to handle uncertainty effectively.

$$A^m_{i,j} = \frac{\exp\left(W_q \mathbf{F}^m_{\text{Unet},i} \cdot W_k \tilde{M}^m_j + \varepsilon_j\right)}{\sum_{k'=1}^{K+1} \exp\left(W_q \mathbf{F}^m_{\text{Unet},i} \cdot W_k \tilde{M}^m_{k'} + \varepsilon_{k'}\right)} \tag{7}$$

In practice, we introduce $\tilde{M}$ using Eq. 7 after each resolution's ResBlock in the upsampling section of Unet. W_q and W_k are the weights of learned linear projections, and $\{\varepsilon_j\}$ are i.i.d. random samples from the $\text{Gumbel}(0, 1)$ distribution. We empower the model to adaptively choose the required object features by conducting the argmax operation across all acquired slots, and further employ a straight-through trick to address the issue of the argmax operation being non-differentiable (Xu et al. 2022; Van Den Oord, Vinyals et al. 2017):

$$\hat{A}^m = \text{one} - \text{hot}\left(A^m_{\text{argmax}}\right) + A^m - \text{sg}\left(A^m\right). \tag{8}$$

Furthermore, we incorporate them into the feature maps of the original Unet using a residual approach, as illustrated in Fig. 2:

$$\mathbf{F}^{m+1}_{\text{Unet}} = \mathbf{F}^m_{\text{Unet}} + \frac{\sum_{j=1}^{K+1} \hat{A}^m_{i,j} W_v \tilde{M}^m_j}{\sum_{j=1}^{K+1} \hat{A}^m_{i,j}}. \tag{9}$$

Finally, Eq. 3 is modified to:

$$\mathcal{L}\left(\theta\right) = \mathbb{E}_t \sum_{n=1}^{N} \left\| X^n_0 - f_\theta\left(X^n_t | t, c, X^{<n}, v, \tilde{M}\right)\right\|^2_2. \tag{10}$$

During training, the parameters of Enc remain fixed.

4 Experiments

4.1 Datasets and Evaluation Metrics

MAGE (Hu, Luo, and Chen 2022) introduces five datasets for evaluating this task, comprising three MNIST datasets and two CATER datasets.

MNIST datasets. Single Moving MNIST contains a single digit, whereas Double Moving MNIST (Mittal, Marwah, and Balasubramanian 2017) features pairs of digits moving in various directions: top to bottom, bottom to top, left to right, and right to left. Modified Double Moving MNIST differs in movement, with digits stopping or bouncing once at boundaries and a random static digit inserted in the background. It's worth noting that we slightly modify these three datasets because an examination of the source code reveals that the digits exhibit uneven speeds when reaching the boundaries and rebounding, which affects the evaluation under varying speed conditions. Consequently, these datasets are adjusted to ensure uniform motion. The video resolution is 64×64 pixels.

CATER datasets. MAGE (Hu, Luo, and Chen 2022) provides two datasets, namely CATER-GEN-v1 and CATER-GEN-v2, which remain unaltered in this paper. CATER-GEN-v1 comprises two objects (a cone and a snitch), while CATER-GEN-v2 encompasses 3 to 8 objects, each defined by four randomly selected attributes like shape and material. The atomic actions of objects in these two datasets are the same, namely, "rotate", "contain", "pick-place", and "slide". The video resolution is 256×256, but it will be resized to 128×128 during the experiments.

In accordance with prior work (Xu et al. 2023; Hu, Luo, and Chen 2022), we select evaluation metrics including Peak Signal to Noise Ratio (PSNR) (Huynh-Thu and Ghanbari 2008), Structural Similarity Index Measure (SSIM) (Wang et al. 2004), Learned Perceptual Image Patch Similarity (LPIPS) (Dosovitskiy and Brox 2016), Fréchet Inception Distance (FID) (Heusel et al. 2017) and Fréchet-Video-Distance (FVD) (Unterthiner et al. 2018). In the computation of LPIPS, the VGG (Simonyan and Zisserman 2014) network was employed.

4.2 Qualitative Results

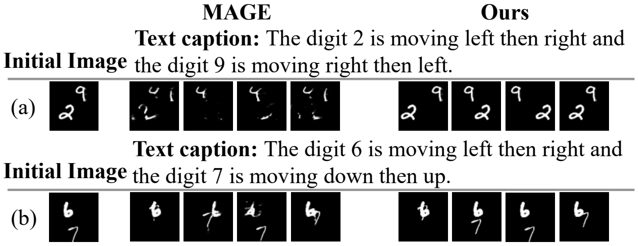


Figure 4: Comparison Results on Double Moving MNIST. Under the constraints of object-centric representations, the shapes of the digits in the videos generated by our model remain consistent even when they overlap. To facilitate visualization, we have extracted specific frames.

This section presents the qualitative results on the datasets mentioned in Sec. 4.1.

Comparison Results. In Figs. 4, 5 and 6, we show video generation results comparing our model with MAGE (Hu, Luo, and Chen 2022). In cases of text-image misalignment,

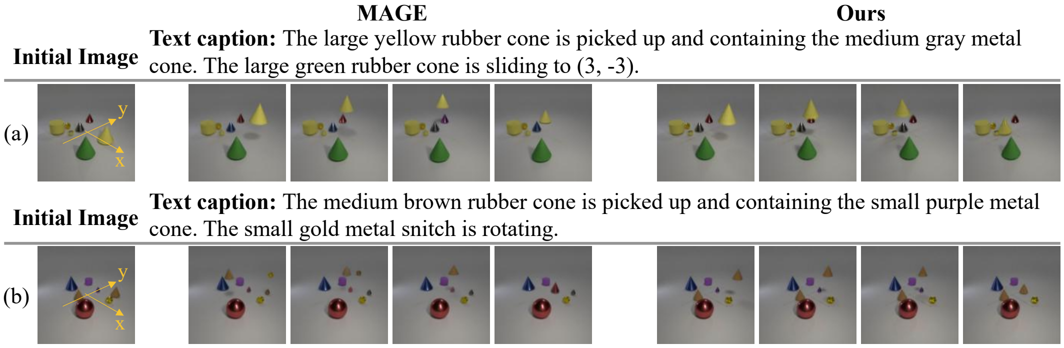


Figure 5: Comparison Results on CATER-GEN-v2. We propose object disentanglement for enhanced alignment between text and image in video generation. This yields improved semantic consistency, enabling precise identification and motion in accordance with textual descriptions.

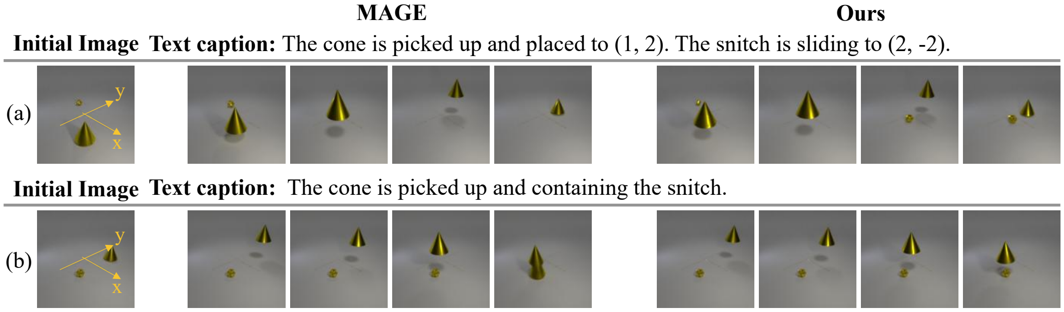


Figure 6: Comparison Results on CATER-GEN-v1. We can observe the object's disappearance or deformation as the video continues. Instead, our model alleviates it by utilizing the object attribute information in slots to improve the generation process.

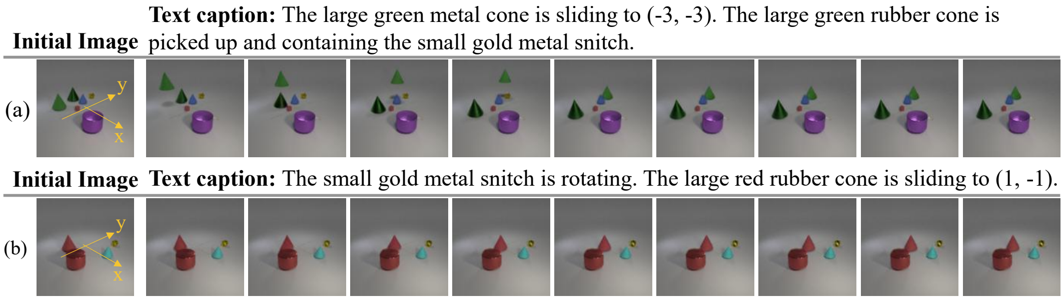


Figure 7: Samples generated from CATER-GEN-v2. The initial image on the far left is provided by the user, and above the images is a text caption describing the movement. The generated videos are coherent and of high perceptual quality, maintaining semantic consistency with textual descriptions.

two scenarios emerge: firstly, the model fails to align object motion with the textual description, and secondly, it mistakenly identifies the objects to be moved. Fig. 5(a) illustrates the first scenario, wherein MAGE successfully identifies the large yellow rubber cone but struggles to execute the action containing the medium gray metal cone. Fig. 5(b) corresponds to the second scenario, where MAGE misidentifies the medium brown rubber cone requiring movement. Our model employs object disentanglement to separate co-located objects, storing object attribute information in discrete slots. This strategy facilitates the alignment of textual descriptions, resulting in improved generation results.

As the number of frames increases, objects in MAGE-generated video frames deform. Instead, our approach utilizes object-centric representations to enhance the generation process, thus ensuring constant awareness of object attributes. This is evident in the comparison results in Fig. 4 and Fig. 6(b). When objects overlap, as observed in Fig. 6(a) of the MAGE generation results, one of the objects may disappear. However, with the assistance of object features, our method improves the temporal consistency of objects.

Generated Results. The CATER-GEN-v2 dataset poses significant challenges due to varying object sizes and multiple objects with identical shapes, demanding a high level of discrimination capability from the model. In Fig. 7(a), there are cones exhibiting distinct attributes, including two cones of green color. Distinguishing between these two cones relies on the model's analysis of their unique material properties. It can be observed that TIV-Diffusion successfully recognizes the difference between metal and rubber, executing the correct motion instructions.

4.3 Quantitive Results

We compare TIV-Diffusion with several state-of-the-art models using SSIM, PSNR, LPIPS, FID, and FVD metrics. Throughout both the training and testing phases, objects in the video move at random speeds. Leveraging a given text caption and initial image, we generate nine subsequent frames, excluding the initial image from quantitative metric calculations. To maintain consistency, we reproduce MAGE's (Hu, Luo, and Chen 2022) results on slightly modified MNIST datasets. Additionally, for the CATER datasets, we replicate the results of MAGE and MAGE+ (Hu, Luo, and Chen 2023) using official weights to ensure uniform test conditions. We follow the fine-tuning procedures of Seer (Gu et al. 2023) to train and test on the CATER datasets. Quantitative results for TiV-ODE and TVP are sourced from (Xu et al. 2023) and (Song et al. 2022), respectively, and are summarized in Tab. 1.

Datasets	Method	SSIM↑	PSNR↑	LPIPS↓	FID↓	FVD↓
Single	TVP	0.46	17.24	0.30	-	-
	MAGE†	0.98	32.98	0.02	5.94	1.20
	Ours	**0.99**	**36.68**	**0.01**	**0.49**	**0.11**
Double	MAGE†	0.92	26.13	0.06	7.83	2.55
	Ours	**0.96**	**32.99**	**0.03**	**1.04**	**0.47**
Modified	MAGE†	**0.92**	24.24	0.06	10.03	10.01
	Ours	0.91	**24.80**	**0.06**	**5.44**	**4.03**
CATERv1	MAGE†	0.97	35.03	0.20	40.09	20.75
	TiV-ODE	0.96	-	0.12	11.98	-
	MAGE+†	0.96	36.16	0.09	6.41	21.15
	Seer	0.90	20.49	0.40	95.07	647.39
	Ours	**0.97**	**37.15**	**0.06**	**6.06**	**13.82**
CATERv2	MAGE†	0.95	32.54	0.22	28.16	51.81
	TiV-ODE	0.93	-	0.18	38.12	-
	MAGE+†	0.95	32.58	0.11	23.56	34.93
	Seer	0.86	18.51	0.42	79.31	589.74
	Ours	**0.95**	**32.66**	**0.09**	**6.67**	**16.48**

Table 1: Quantitive comparison of our model and other models. † means reproducing the results.

Our method demonstrates superiority in FID, FVD, and LPIPS metrics while remaining competitive in SSIM and PSNR compared to other models. The powerful generation capabilities of diffusion models have significantly heightened the perceptual quality of the resulting videos. Additionally, we enhance the alignment between text captions and images through object disentanglement, ensuring semantic consistency. The mitigation of issues related to object deformation or disappearance further improves visual quality of the generated video. These results collectively underscore the competitiveness of TIV-Diffusion for TI2V tasks.

Datasets	Method	SSIM↑	PSNR↑	LPIPS↓	FID↓	FVD↓
Single	Ours w/o	0.99	34.50	0.01	1.66	0.52
	Ours	**0.99**	**36.68**	**0.01**	**0.49**	**0.11**
Double	Ours w/o	0.96	31.06	0.03	1.55	0.51
	Ours	**0.96**	**32.99**	**0.03**	**1.04**	**0.47**
Modified	Ours w/o	0.90	24.34	0.06	6.37	6.47
	Ours	**0.91**	**24.80**	**0.06**	**5.44**	**4.03**
CATER-v1	Ours w/o	0.97	36.59	0.07	6.99	30.87
	Ours	**0.97**	**37.15**	**0.06**	**6.06**	**13.82**
CATER-v2	Ours w/o	0.95	32.06	0.09	10.66	32.72
	Ours	**0.95**	**32.66**	**0.09**	**6.67**	**16.48**

Table 2: Quantitive results of the ablation study. "Ours w/o" signifies the absence of object-centric representations, while "Ours" indicates their presence.

4.4 Ablation Study

Object Disentanglement. We train the encoder Enc for object disentanglement on MNIST datasets and CATER datasets respectively. During its initial image processing, Enc extracts slots, with each slot corresponding to an object. To demonstrate the advantages of object disentanglement, we compare the results with and without using object-centric representations, as shown in Tab. 2, where object-centric textual-visual alignment enhances generation results.

5 Conclusion

In this paper, we propose a model for Text-Image-to-Video Generation, named TIV-Diffusion. Our model extends the diffusion model to enable autoregressive video frame generation, reducing the computational resources required for training. In addition to incorporating the fused textual and visual knowledge with the scale-offset modulation, we introduce object-centric representations to improve cross-modal alignment. After the slot attention encoder processes the input image, we obtain the corresponding slot for each object, which contains object attribute information. Subsequently, we identify the target objects to be manipulated based on the accompanying text. This process parallels the way humans comprehend natural phenomena. To mitigate object disappearance or deformation, and given the stochastic nature of objects stored in slots, our model employs Gumbel-Softmax to fuse object attributes. Experimental results confirm that our model attains state-of-the-art performance on existing datasets.

Acknowledgments

This work was supported in part by NSFC under Grant 623B2098, 62021001, and 62371434. This work was mainly completed before March 2024.

References

Aldausari, N.; Sowmya, A.; Marcus, N.; and Mohammadi, G. 2022. Video generative adversarial networks: a review. *ACM CSUR*, 55(2): 1–25.

An, J.; Zhang, S.; Yang, H.; Gupta, S.; Huang, J.-B.; Luo, J.; and Yin, X. 2023. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*.

Ballas, N.; Yao, L.; Pal, C.; and Courville, A. 2015. Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 22563–22575.

Chang, Z.; Zhang, X.; Wang, S.; Ma, S.; Ye, Y.; Xinguang, X.; and Gao, W. 2021. Mau: A motion-aware unit for video prediction and beyond. *NeurIPS*, 34: 26950–26962.

Chen, H.; Xia, M.; He, Y.; Zhang, Y.; Cun, X.; Yang, S.; Xing, J.; Liu, Y.; Chen, Q.; Wang, X.; et al. 2023. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dosovitskiy, A.; and Brox, T. 2016. Generating images with perceptual similarity metrics based on deep networks. *NeurIPS*, 29.

Fu, T.-J.; Yu, L.; Zhang, N.; Fu, C.-Y.; Su, J.-C.; Wang, W. Y.; and Bell, S. 2023. Tell me what happened: Unifying text-guided video completion via multimodal masked video generation. In *CVPR*, 10681–10692.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *NeurIPS*, 27.

Gu, X.; Wen, C.; Song, J.; and Gao, Y. 2023. Seer: Language Instructed Video Prediction with Latent Diffusion Models. *arXiv preprint arXiv:2303.14897*.

Gupta, A.; Tian, S.; Zhang, Y.; Wu, J.; Martín-Martín, R.; and Fei-Fei, L. 2022. Maskvit: Masked visual pre-training for video prediction. *arXiv preprint arXiv:2206.11894*.

He, Y.; Yang, T.; Zhang, Y.; Shan, Y.; and Chen, Q. 2022. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851.

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*.

Höppe, T.; Mehrjou, A.; Bauer, S.; Nielsen, D.; and Dittadi, A. 2022. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*.

Hu, Y.; Luo, C.; and Chen, Z. 2022. Make it move: controllable image-to-video generation with text descriptions. In *CVPR*, 18219–18228.

Hu, Y.; Luo, C.; and Chen, Z. 2023. A Benchmark for Controllable Text-Image-to-Video Generation. *TMM*.

Huynh-Thu, Q.; and Ghanbari, M. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics letters*, 44(13): 800–801.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Jiang, J.; Deng, F.; Singh, G.; and Ahn, S. 2023. Object-centric slot diffusion. *arXiv preprint arXiv:2303.10834*.

LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; and Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4): 541–551.

Lee, S.; Kim, H. G.; Choi, D. H.; Kim, H.-I.; and Ro, Y. M. 2021. Video prediction recalling long-term motion context via memory alignment learning. In *CVPR*, 3054–3063.

Li, X.; Ren, Y.; Jin, X.; Lan, C.; Wang, X.; Zeng, W.; Wang, X.; and Chen, Z. 2023. Diffusion Models for Image Restoration and Enhancement–A Comprehensive Survey. *arXiv preprint arXiv:2308.09388*.

Locatello, F.; Weissenborn, D.; Unterthiner, T.; Mahendran, A.; Heigold, G.; Uszkoreit, J.; Dosovitskiy, A.; and Kipf, T. 2020. Object-centric learning with slot attention. *NeurIPS*, 33: 11525–11538.

Mei, K.; and Patel, V. 2023. Vidm: Video implicit diffusion models. In *AAAI*, volume 37, 9117–9125.

Mittal, G.; Marwah, T.; and Balasubramanian, V. N. 2017. Sync-draw: Automatic video generation using deep recurrent attentive architectures. In *ACM MM*, 1096–1104.

Ni, H.; Shi, C.; Li, K.; Huang, S. X.; and Min, M. R. 2023. Conditional Image-to-Video Generation with Latent Flow Diffusion Models. In *CVPR*, 18444–18455.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2337–2346.

Qiu, H.; Xia, M.; Zhang, Y.; He, Y.; Wang, X.; Shan, Y.; and Liu, Z. 2023. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Ren, Y.; Li, X.; Li, B.; Wang, X.; Guo, M.; Zhao, S.; Zhang, L.; and Chen, Z. 2025. Moe-diffir: Task-customized diffusion priors for universal compressed image restoration. In *ECCV*, 116–134. Springer.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer.

Saito, M.; Matsumoto, E.; and Saito, S. 2017. Temporal generative adversarial nets with singular value clipping. In *ICCV*, 2830–2839.

Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *NeurIPS*, 28.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, X.; Chen, J.; Zhu, B.; and Jiang, Y.-G. 2022. Text-driven Video Prediction. *arXiv preprint arXiv:2210.02872*.

Sun, M.; Wang, W.; Zhu, X.; and Liu, J. 2023. MOSO: Decomposing MOtion, Scene and Object for Video Prediction. In *CVPR*, 18727–18737.

Tan, C.; Gao, Z.; Wu, L.; Xu, Y.; Xia, J.; Li, S.; and Li, S. Z. 2023a. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *CVPR*, 18770–18782.

Tan, C.; Li, S.; Gao, Z.; Guan, W.; Wang, Z.; Liu, Z.; Wu, L.; and Li, S. Z. 2023b. OpenSTL: A Comprehensive Benchmark of Spatio-Temporal Predictive Learning. *arXiv preprint arXiv:2306.11249*.

Tulyakov, S.; Liu, M.-Y.; Yang, X.; and Kautz, J. 2018. Mocogan: Decomposing motion and content for video generation. In *CVPR*, 1526–1535.

Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *NeurIPS*, 30.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

Voleti, V.; Jolicoeur-Martineau, A.; and Pal, C. 2022. MCVD-masked conditional video diffusion for prediction, generation, and interpolation. *NeurIPS*, 35: 23371–23385.

Vondrick, C.; Pirsiavash, H.; and Torralba, A. 2016. Generating videos with scene dynamics. *NeurIPS*, 29.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023a. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Wang, X.; Li, X.; and Chen, Z. 2024. CoNo: Consistency Noise Injection for Tuning-free Long Video Diffusion. *arXiv preprint arXiv:2406.05082*.

Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. 2023b. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4): 600–612.

Weissenborn, D.; Täckström, O.; and Uszkoreit, J. 2019. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*.

Wu, C.; Huang, L.; Zhang, Q.; Li, B.; Ji, L.; Yang, F.; Sapiro, G.; and Duan, N. 2021. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*.

Wu, C.; Liang, J.; Ji, L.; Yang, F.; Fang, Y.; Jiang, D.; and Duan, N. 2022a. Nüwa: Visual synthesis pre-training for neural visual world creation. In *ECCV*, 720–736. Springer.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, W.; Gu, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2022b. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*.

Wu, T.; Si, C.; Jiang, Y.; Huang, Z.; and Liu, Z. 2025. Freeinit: Bridging initialization gap in video diffusion models. In *ECCV*, 378–394. Springer.

Xing, J.; Xia, M.; Zhang, Y.; Chen, H.; Wang, X.; Wong, T.-T.; and Shan, Y. 2023a. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*.

Xing, Z.; Dai, Q.; Hu, H.; Wu, Z.; and Jiang, Y.-G. 2023b. SimDA: Simple Diffusion Adapter for Efficient Video Generation. *arXiv preprint arXiv:2308.09710*.

Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; and Wang, X. 2022. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, 18134–18144.

Xu, Y.; Li, N.; Goel, A.; Guo, Z.; Yao, Z.; Kasaei, H.; Kasaei, M.; and Li, Z. 2023. Controllable Video Generation by Learning the Underlying Dynamical System with Neural ODE. *arXiv preprint arXiv:2303.05323*.

Yang, R.; Srivastava, P.; and Mandt, S. 2022. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*.

Zhou, D.; Wang, W.; Yan, H.; Lv, W.; Zhu, Y.; and Feng, J. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*.