FIELD-DIT: DIFFUSION TRANSFORMER ON UNIFIED VIDEO, 3D, AND GAME FIELD GENERATION

Kangfu Mei
Johns Hopkins University
kmei1@jhu.edu

Mo Zhou
Johns Hopkins University
mzhou32@jhu.edu

Vishal M. Patel
Johns Hopkins University
vpatel36@jhu.edu

ABSTRACT

The probabilistic field models the distribution of continuous functions defined over metric spaces. While these models hold great potential for unifying data generation across various modalities, including images, videos, and 3D geometry, they still struggle with long-context generation beyond simple examples. This limitation can be attributed to their MLP architecture, which lacks sufficient inductive bias to capture global structures through uniform sampling. To address this, we propose a new and simple model that incorporates a view-wise sampling algorithm to focus on local structure learning, along with autoregressive generation to preserve global geometry. It adapts cross-modality conditions, such as text prompts for text-to-video generation, camera poses for 3D view generation, and control actions for game generation. Experimental results across various modalities demonstrate the effectiveness of our model, with its 675M parameter size, and highlight its potential as a foundational framework for scalable, architecture-unified visual content generation for different modalities with different weights. Our project page can be found at https://kfmei.com/Field-DiT/.

1 INTRODUCTION

Generative tasks (Rombach et al., 2022; Ramesh et al., 2022) are overwhelmed by diffusion probabilistic models that hold state-of-the-art results on most modalities like audio, images, videos, and 3D geometry. Take image generation as an example, a typical diffusion model (Ho et al., 2020) consists of a forward process for sequentially corrupting an image into standard noise, a backward process for sequentially denoising a noisy image into a clear image, and a score network that learns to denoise the noisy image.

The forward and backward processes are agnostic to different data modalities; however, the architectures of the existing score networks are not. The existing score networks are highly customized towards a single type of modality, which is challenging to adapt to a different modality. For example, a recently proposed multi-frame video generation network (Ho et al., 2022b;a) adapting single-frame image generation networks involves significant designs and efforts in modifying the score networks.

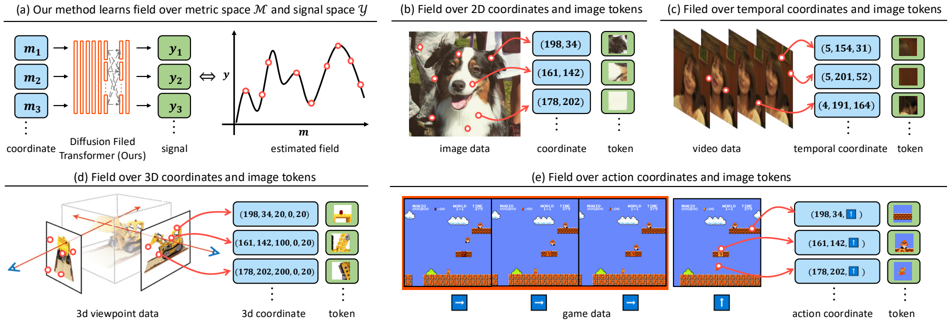


Figure 1: Illustration of the field model's capability to model visual content. The model learns the distribution through attention between coordinate-signal pairs, which is modality-agnostic.

Therefore, it is important to develop a single, versatile architecture that can be applied across different modalities without modification. Such a unified architecture simplifies the development process, reduces the complexity associated with designing modality-specific models, and enables knowledge transfer between modalities.

Field model (Sitzmann et al., 2020; Tancik et al., 2020; Dupont et al., 2022b; Zhuang et al., 2023) is a promising unified score network architecture for different modalities. It learns the distribution over the functional view of data. Specifically, the field f maps the observation from the *metric* space $\mathcal{M}$ (*e.g.*, coordinate or camera pose) into the *signal* space $\mathcal{Y}$ (*e.g.*, RGB pixel) as $f : \mathcal{M} \mapsto \mathcal{Y}$. For instance, an image is represented as $f : \mathbb{R}^2 \mapsto \mathbb{R}^3$ that maps the spatial coordinates (*i.e.*, height and width) into RGB values at the corresponding location, while a video is represented as $f : \mathbb{R}^3 \mapsto \mathbb{R}^3$ that maps the spatial and temporal coordinates (*i.e.*, frame, height, and width) into RGB values. Different modalities usually use different weights to ensure the best performance. Recently, diffusion models are leveraged to characterize the field distributions over the functional view of data (Zhuang et al., 2023) for field generation. Given a set of coordinate-signal pairs $\{(\boldsymbol{m}_i, \boldsymbol{y}_i)\}$, the field f is regarded as the score network for the backward process, which turns a noisy signal into a clear signal $\boldsymbol{y}_i$ in a sequential process with $\boldsymbol{m}_i$ being fixed all the time. The visual content is then composed of the clear signal generated on a grid in the metric space.

Nevertheless, diffusion-based field models for generation still lag behind the modality-specific approaches (Dhariwal & Nichol, 2021; Ho et al., 2022b; He et al., 2022) for learning from dynamic data in high resolution (Bain et al., 2021; Yu et al., 2023a). For example, a 240p video lasting 5 seconds is comprised of up to 10 million coordinate-signal pairs. Due to the memory bottleneck in existing GPU-accelerated computing systems, recent field models (Zhuang et al., 2023) are limited to observe merely a small portion of these pairs (*e.g.*, 1%) that are uniformly sampled during training. This limitation significantly hampers the field models in approximating distributions from such sparse observations (Quinonero-Candela & Rasmussen, 2005). Consequently, diffusion-based field models often struggle to capture the fine-grained local structure of the data, leading to, *e.g.*, unsatisfactory blurry results.

While it is possible to change the pair sampling algorithm to sample densely from local areas instead of uniformly, the global geometry is weakened. To alleviate this issue, it is desirable to introduce some complementary guidance on the global geometry in addition to local sampling.

In this paper, we propose a new diffusion field transformers, called **Field-DiT**. In contrast to previous methods, Field-DiT preserves both the local structure and the global geometry of the fields during learning by employing a new view-wise sampling algorithm in the coordinate space, and incorporates additional inductive biases from the text descriptions and autoregressive generation. By combining these advancements with our simplified transformer architecture, we demonstrate that modeling capability of our model surpasses previous methods, achieving improved generated results under the same memory constraints. We empirically validate its superiority against previous domain-agnostic methods across three different tasks, including text-to-video generation, 3D novel-view generation, and game generation. Various experiments show that our method achieves compelling performance even when compared to the state-of-the-art domain-specific methods, underlining its potential as a scalable and architecture-unified visual content generation model across various modalities. Our contributions are summarized as follows:

- We propose a new transformer-based diffusion field model for long-context modeling, which comprises of a view-wise sampling algorithm and autoregressive generation for local structure and global geometry model respectively.
- We demonstrate the effectiveness and efficiency of a simple 675M model on different modalities generation including video, 3D, and game in a unified-architecture, which largely closes the performance gap with modality-specific models with different weights.
- We show the potential of action game generation using diffusion models, and we release the benchmarks including both training and testing data for replication and comparisons.

2 Related Work

Generation Models. In recent years, generative models have shown impressive performance in visual content generation. The major families are generative adversarial networks (Goodfellow et al., 2020; Mao et al., 2017; Karras et al., 2019; Brock et al., 2019), variational autoencoders (Kingma & Welling, 2014; Vahdat & Kautz, 2020), auto-aggressive networks (Chen et al., 2020; Esser et al., 2021), diffusion models (Ho et al., 2020; Song et al., 2021), and consistency models (Mei et al., 2024a; Song et al., 2023). Recent diffusion models have obtained significant advancement with stronger network architectures (Dhariwal & Nichol, 2021; Mei et al., 2025b), additional text and image condition (Ramesh et al., 2022; Mei et al., 2024b; 2025a), and pretrained latent space (He et al., 2022). Our method built upon these successes and targets at scaling domain-agnostic models.

Field Models. Field models like SIREN Sitzmann et al. (2020) excel at effectively handling diverse data types, such as images, videos, 3D shapes, and audio, without requiring extensive customization. Compared with the modality-specific models, field models enable scalability by allowing advancements in one domain (e.g., images) to directly enhance others (e.g., 3D modeling and video synthesis), streamlining research and development. In order to model complex field distributions, representative methods like Functa (Dupont et al., 2022b) and GEM (Du et al., 2021) adopt a two-stage modeling paradigm: first parameterizing fields, then learning distributions over the parameterized latent space. However, the learning efficiency of the two-stage methods hinders scaling the models, as their first stage incurs substantial computational costs to compress fields into latent codes. Building on recent exploration Zhuang et al. (2023) into the use of diffusion models, which are more powerful for directly modeling complex data distributions without additional parametrization, we propose to model field distributions using explicit coordinate-signal pairs. Nevertheless, field models struggle with very large or highly diverse datasets, such as high-resolution videos. This is due to the complexity of preserving both local structures and global geometry. In contrast, our method leverages the benefits of a single-stage modeling approach, improving accuracy in preserving both local structures and global geometry.

Long-context Modeling. Our method also differs from the recently proposed domain-specific works for high-resolution, dynamic data, which models specific modalities in a dedicated latent space, including Spatial Functa (Bauer et al., 2023) and PVDM (Yu et al., 2023c). These methods typically compress the high-dimensional data into a low-dimensional latent space. However, the compression is usually specific to a center modality and lacks the flexibility in dealing with different modalities. For instances, PVDM compresses videos into three latent codes that represent spatial and temporal dimensions separately. However, such a compressor cannot be adopted into the other similar modalities like 3D scenes. In contrast, our method owns the unification flexibility and the achieved advancement can be easily transferred into different modalities.

3 METHOD

Definition. Conceptually, the diffusion-based field models sample from field distributions by reversing a gradual noising process. As shown in Fig. 1, in contrast to the data formulation of the conventional diffusion models (Ho et al., 2020) applied to the complete data like a whole image, diffusion-based field models apply the noising process to the sparse observation of the field, which is a kind of parametrized functional representation of data consisting of coordinate-signal pairs, *i.e.*, $f : \mathcal{M} \mapsto \mathcal{Y}$. Specifically, the sampling process begins with a coordinate-signal pair $(\mathbf{m}_i, \mathbf{y}_{(i,T)})$, where the coordinate comes from a field and the signal is a standard noise, and less-noisy signals $\mathbf{y}_{(i,T-1)}, \mathbf{y}_{(i,T-2)}, \ldots$, are progressively generated until reaching the final clear signal $\mathbf{y}_{(i,0)}$, with $\mathbf{m}_i$ being constant. Diffusion Probabilistic Field (DPF) (Zhuang et al., 2023) is one of the recent representative diffusion-based field models. It parameterizes the denoising process with a transformer-based network $\epsilon_\theta(\cdot)$, which takes noisy coordinate-signal pairs as input and predicts the noise component ϵ of $\mathbf{y}_{(i,t)}$. The less-noisy signal $\mathbf{y}_{(i,t-1)}$ is then sampled from the noise component ϵ using a denoising process (Ho et al., 2020).

In practice, when handling low-resolution data consisting of N coordinate-signal pairs with DPF, the scoring network $\epsilon_\theta(\cdot)$ takes all pairs $\{(\mathbf{m}_i, \mathbf{y}_{(i,T)})\}$ as input at once. For high-resolution data with a large number of coordinate-signal pairs that greatly exceed the modern GPU capacity, (Zhuang et al., 2023) uniformly sample a subset of pairs from the data as input. They subsequently condition the diffusion model on the other non-overlapping subset, referred to as *context pairs*. Specifically, the sampled pairs interact with the query pairs through cross-attention blocks. (Zhuang et al., 2023) show that the ratio between the context pairs and the sampling pairs is strongly related to the quality

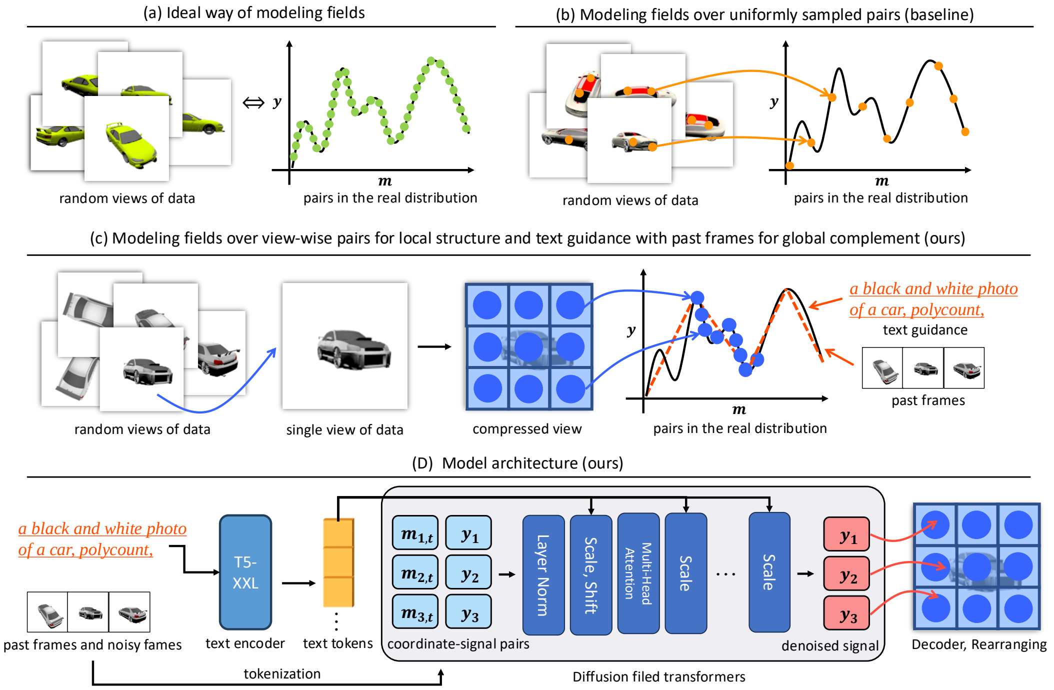

(a) Ideal way of modeling fields
⇔ y
m
random views of data
pairs in the real distribution
(b) Modeling fields over uniformly sampled pairs (baseline)
y
m
random views of data
pairs in the real distribution
(c) Modeling fields over view-wise pairs for local structure and text guidance with past frames for global complement (ours)
random views of data
single view of data
compressed view
y
m
pairs in the real distribution
a black and white photo of a car, polycount,
text guidance
past frames
(D) Model architecture (ours)
a black and white photo of a car, polycount,
past frames and noisy fames
T5-XXL
text encoder
text tokens
tokenization
$m_{1,t}$ y_1
$m_{2,t}$ y_2
$m_{3,t}$ y_3
coordinate-signal pairs
Layer Norm
Scale, Shift
Multi-Head Attention
Scale
...
Scale
y_1
y_2
y_3
denoised signal
Decoder, Rearranging
Diffusion filed transformers

Figure 2: (a) Ideally, all pairs within a field (green points) should be used for training, but this is impractical due to memory limitations. (b) Previous methods uniformly sample a sparse set of pairs (orange points) to represent the field to mitigate memory limitations. (c) Compared to uniform sampling, our local sampling extracts high-fidelity pairs (blue points), better covering the local structure. The text prompt and past frames serve as an approximation to complement the global geometry. (d) Visualization of our sampling pipeline. Note that the input coordinates include the diffusion timesteps of each input frames.

of the generated fields, and the quality decreases as the context pair ratio decreases. Due to the practical memory bottleneck, DPF can only support a maximum 64×64 resolution, let alone being extended to long context such as multi-frame video generation.

3.1 Diffusion Field Transformer

In order to scale diffusion-based field models for high-resolution, dynamic data generation, we build upon the recent DPF model (Zhuang et al., 2023) and address its limitations in preserving the local structure of fields, as it can hardly be captured when the uniformly sampled coordinate-signal pairs are too sparse. Specially, our method not only can preserve the local structure, but also introduce additional inductive biases for capturing the global geometry, such as text descriptions and past frames in autoregressive generation.

In order to preserve the local structure of fields, we propose a new view-wise sampling algorithm that samples local coordinate-signal pairs for better representing the local structure of fields. For instance, the algorithm samples the coordinate-signal pairs belonging to a single or several ($n \geqslant 1$; n denotes the number of views) views for video data, where a view corresponds to a single frame. It samples pairs belonging to a single or several rendered images for 3D viewpoints, where a view corresponds to an image rendered at a specific camera pose. A view of an image is the image itself.

This approach restricts the number of interactions among pairs to be modeled and reduces the learning difficulty on high-resolution, dynamic data. Nevertheless, even a single high-resolution view , *e.g.*, in merely 128×128 resolution) can still consist of 10K pairs, which in practice will very easily reach the memory bottleneck if we leverage a large portion of them at one time, and hence hinder scaling the model for generating high-resolution dynamic data.

To address this issue, our method begins with increasing the signal resolution of coordinate-signal pairs and hence reducing memory usage in the score network. Specifically, we replace the signal

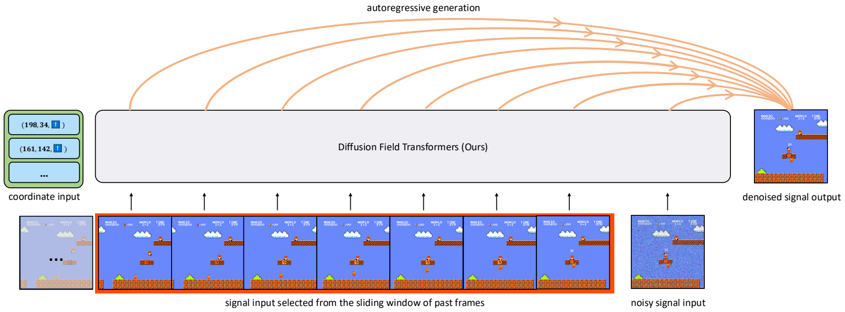

autoregressive generation
(198, 34, ⬆)
(161, 142, ⬆)
...
coordinate input
Diffusion Field Transformers (Ours)
denoised signal output
signal input selected from the sliding window of past frames
noisy signal input

Figure 3: Autoregressive next-frame prediction. Our model takes past frames selected from a sliding window and next action coordinates, such as actions like jump or move, as input. It then generates the next frame, reflecting both the action and the long context of the past frames.

space with a compressed latent space, and employ a more efficient network architecture that only contains decoders. This improvement in efficiency allows the modeling of interactions among pairs representing higher-resolution data while keeping the memory usage constrained. Based on this, one can then model the interactions of pairs within a single or several views of high-resolution data. The overall diagram of the proposed sampling method can be found in Fig. 2.

View-wise Sampling. Based on the high-resolution signal and decoder-only network architecture, our method represents field distributions by using view-consistent coordinate-signal pairs, *i.e.*, collections of pairs that belong to a single or several ($n \geqslant 1$) views of the data, such as one or several frames in a video, and one or several viewpoints of a 3D geometry. In particular, take the spatial and temporal coordinates of a video in $H \times W$ resolution lasting for T frames as an example, for all coordinates $\{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_i, \ldots, \mathbf{m}_{H\times W\times T}\}$, we randomly sample a consecutive sequence of length $H \times W$ that correspond to a single frame, *i.e.*, $\{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_i, \ldots, \mathbf{m}_{H\times W}\}$. For data consisting of a large amount of views (*e.g.* $T >> 16$), we randomly sample n views (sequences of length $H \times W$), resulting in an $H \times W \times n$ sequence set. Accordingly, different from the transformers in previous works (Zhuang et al., 2023) that model interaction among all pairs across all views, ours only models the interaction among pairs that belongs to the same view, which reduces the complexity of field model by limiting the number of interactions to be learned.

3.2 Long-context Conditioning

To complement our effort in preserving local structures that may weaken global geometry learning, since the network only models the interaction of coordinate-signal pairs in the same view, we propose to supplement the learning with a long-context conditioning of the field, avoiding issues in cross-view consistency like worse spatial-temporal consistency between frames in video generation.

In particular, we propose to condition diffusion models on long-context such as text-prompt and past frames related to the fields. Text-prompt can represent data in compact but highly expressive features (Devlin et al., 2019; Brown et al., 2020; Raffel et al., 2020), and serve as a low-rank approximation of data (Radford et al., 2021). Past frames are especially useful in autoregressive generation, such as in game data. By conditioning diffusion models on long-context, we demonstrate that our method can capture the global geometry for generating long videos and game sequences.

Text-prompt for Cross-view Condition Consistency. In order to model the dependency variation between views belonging to the same field, *i.e.*, the global geometry of the field, we condition the diffusion model on the text embeddings of the field description or equivalent embeddings (*i.e.*, the language embedding of a single view in the CLIP latent space (Radford et al., 2021)). Our approach leverages the adaptive layer normalization layers in GANs (Brock et al., 2019; Karras et al., 2019), and adapts them by modeling the statistics from the text embeddings of shape $Z \times D$. For pairs that make up a single view, we condition on their represented tokens $Z \times D$, (Z tokens of size D), by modulating them with the scale and shift parameters regressed from the text embeddings. For pairs $(T \times Z) \times D$ that make up multiple views, we condition on the view-level pairs by modulating feature in $Z \times D$ for each of the T views with the same scale and shift parameters. Specifically, each transformer blocks of our score network learns to predict statistic features β_c and γ_c from the text embeddings per channel. These statistic features then modulate the transformer features F_c as: $\mathrm{adLNorm}(F_c|\beta_c, \gamma_c) = \mathrm{Norm}(F_c) \cdot \beta_c + \beta_c$.

Past frames for Autoregressive Generation. Recent work (Mei & Patel, 2023) shown that generating long videos and games can be formulated as autoregressive generation, where each frame depends only on past frames and current actions. In Figure. 3, we illustrate the input and output of our model, where it conditions on the past T frames and generates the next $T+1$ frame. Additional inputs include coordinates consisting of spatiotemporal coordinates and one-hot encoded actions, such as *jump* and *move*, from the last frame. Due to memory constraints, the input past frames are limited to a fixed number of n past frames, acting as a sliding window for long-context modeling. The generation of the next n frames on the diffusion field can be simplified as

$$p\left(\mathbf{y}_1, \mathbf{y}_2 \ldots, \mathbf{y}_{n-1}, \mathbf{y}_{(n,t-1)}\right) = \prod_{i=1}^{n} p\left(\mathbf{y}_{(n,t-1)} \mid \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n-1}\right), \tag{1}$$

where $p(\cdot)$ represents the modeled signal probability conditioned on the past frames. Additional conditions also include coordinate inputs $(\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_n)$ and the diffusion timestep t. Empirically, we use the last 16 frames as the context length for game generation, and the last 8 frames as the context length for text-to-video generation.

The proposed autoregressive generation not only preserves global geometry of the data but also significantly improves efficiency in long-context generation. Typical autoregressive transformer models like GPT (Radford et al., 2019) depend on the number of generated tokens, as each new token is conditioned on all previously generated tokens. In contrast to GPT, our method achieves linear complexity with respect to the number of generated frames, similar to the parallel generation efficiency. Each new frame depends only on a fixed number of the most recently generated frames, where conditioning frames are updated in the sliding window. Our game generation maximizes this efficiency, enabling the stable generation of games with an infinite number of frames.

4 Experimental Results

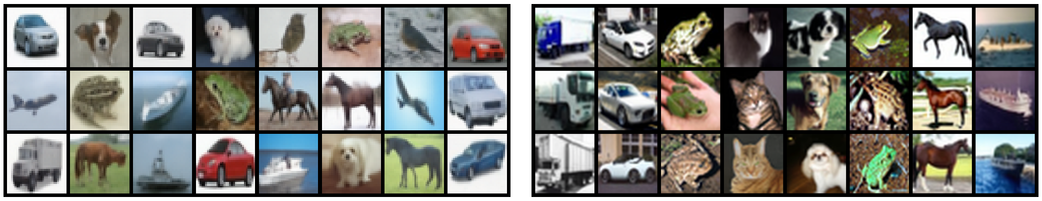
We demonstrate the effectiveness of our method on multiple modalities, including 2D image data on a spatial metric space $\mathbb{R}^2$, 3D video data on a spatial-temporal metric space $\mathbb{R}^3$, and 3D viewpoint data on a camera pose and intrinsic parameter metric space $\mathbb{R}^6$, game data on a action and spatial-temporal metric space $\mathbb{R}^4$, while the score network implementation remains identical across different modalities, except for the embedding size.

Experimental Details. In the interest of maintaining simplicity, we adhere to the methodology outlined by Dhariwal et al. (Dhariwal & Nichol, 2021) and utilize a 256-dimensional frequency embedding to encapsulate input denoising timesteps. This embedding is then refined through a two-layer Multilayer Perceptron (MLP) with Swish (SiLU) activation functions. Our model aligns with the size configuration of DiT-XL (Peebles & Xie, 2023), which includes retaining the number of transformer blocks (*i.e.* 28), the hidden dimension size of each transformer block (*i.e.*, 1152), and the number of attention heads (*i.e.*, 16). Our model derives text embeddings employing T5-XXL (Raffel et al., 2020), culminating in a fixed length token sequence (*i.e.*, 256) which matches the length of the noisy tokens. To further process each text embedding token, our model compresses them via a single layer MLP, which has a hidden dimension size identical to that of the transformer block. Our model uses classifier-free guidance in the backward process with a fixed scale of 8.5. To keep consistency with DiT-XL (Peebles & Xie, 2023), we only applied guidance to the first three channels of each denoised token.

Generative Metrics. In video generation, we use FVD (Unterthiner et al., 2018) to evaluate the video spatial-temporal coherency, FID (Heusel et al., 2017) to evaluate the frame quality, and CLIPSIM (Radford et al., 2021) to evaluate relevance between the generated video and input text. As all metrics are sensitive to data scale during testing, we randomly select 2,048 videos from the test data and generate results as the "real" and "fake" part in our metric experiments. For FID, we uniformly sample 4 frames from each video and use a total of 8,192 images. For CLIPSIM, we calculate the average score across all frames. We use the "openai/clip-vit-large-patch14" model for extracting features in CLIPSIM calculation.

Model	CIFAR10 64×64		CelebV-Text 256×256×128			ShapeNet-Cars 128×128×251			
	FID (↓)	IS (↑)	FVD (↓)	FID (↓)	CLIPSIM (↑)	FID (↓)	LPIPS (↓)	PSNR (↑)	SSIM (↑)
Functa (Dupont et al., 2022a)	31.56	8.12	✗	✗	✗	80.30	0.183	N/A	N/A
GEM (Du et al., 2021)	23.83	8.36	✗	✗	✗	✗	✗	✗	✗
DPF (Zhuang et al., 2023)	15.10	8.43	✗	✗	✗	43.83	0.158	18.6	0.81
DiT (Peebles & Xie, 2023)	7.53	8.97	✗	✗	✗		✗	✗	✗
TFGAN (Balaji et al., 2019)	✗	✗	571.34	784.93	0.154	✗	✗	✗	✗
MMVID (Han et al., 2022b)	✗	✗	109.25	82.55	0.174	✗	✗	✗	✗
MMVID-interp (Han et al., 2022b)	✗	✗	80.81	70.88	0.176	✗	✗	✗	✗
VDM (Ho et al., 2022b)	✗	✗	81.44	90.28	0.162	✗	✗	✗	✗
CogVideo (Hong et al., 2023)	✗	✗	99.28	54.05	0.186	✗	✗	✗	✗
Latte (Ma et al., 2024)	✗	✗	67.97	39.69	0.201	✗	✗	✗	✗
EG3D-PTI (Chan et al., 2022)	✗	✗	✗	✗	✗	20.82	0.146	19.0	0.85
ViewFormer (Kulhánek et al., 2022)	✗	✗	✗	✗	✗	27.23	0.150	19.0	0.83
pixelNeRF (Yu et al., 2021)	✗	✗	✗	✗	✗	65.83	0.146	23.2	0.90
Zero-1-to-3 (Liu et al., 2023)	✗	✗	✗	✗	✗	**17.901**	**0.093**	23.1	0.80
Field-DiT (Ours)	**7.29**	**9.31**	**42.03**	**24.33**	**0.220**	24.36	0.118	**23.9**	**0.90**

Table 1: Sample quality comparison with state-of-the-art field models and representative modality-specific models for each task. "✗" denotes that the method cannot be applied to the modality due to its design or impractical computational costs.



(a) DPF (Zhuang et al., 2023)

(b) **Field-DiT** (ours)

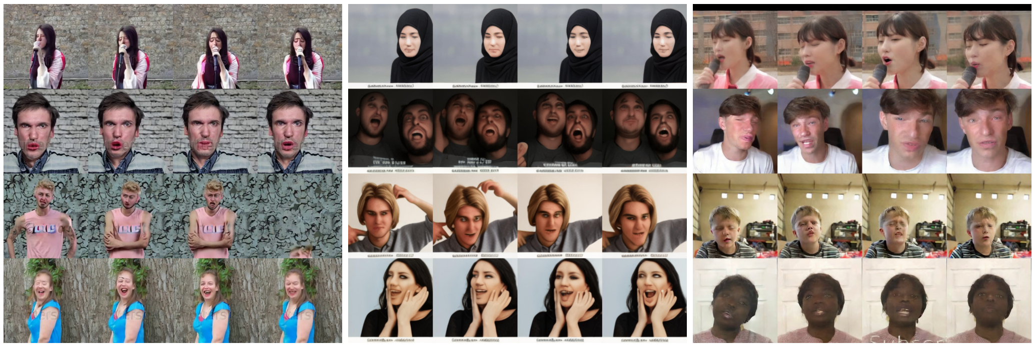
Figure 4: Qualitative comparisons of domain-agnostic methods and ours on CIFAR-10. Our results show better visual quality with more details than the others, while being domain-agnostic as well.

Images. For image generation, we use the standard benchmark dataset, *i.e.*, CIFAR10 64×64 (Krizhevsky et al., 2009) as a sanity test, in order to compare with other domain-agnostic and domain-specific methods. For the low-resolution CIFAR10 dataset, we compare our method with the previous domain-agnostic methods including DPF (Zhuang et al., 2023) and GEM (Du et al., 2021). We report Fréchet Inception Distance (FID) Heusel et al. (2017) and Inception Score (IS) (Salimans et al., 2016) or quantitative comparisons.

The experimental results can be found in Tab. 1. Specifically, Field-DiT outperforms all domain-agnostic models in the FID and IS metrics. The qualitative comparisons in Fig. 4 further demonstrate our method's superiority in images. Note that our method does not use text descriptions for ensuring a fair comparison. It simply learns to predict all coordinate-signal pairs of a single image during training without using additional text descriptions or embeddings.

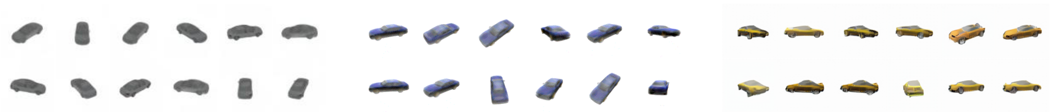
Videos. To show our model's capacity for more complex data, *i.e.*, high-resolution, dynamic video, we conduct experiments on the recent text-to-video benchmark: CelebV-Text 256×256×128 (Yu et al., 2023b) (128 frames). As additional spatial and temporal coherence is enforced compared to images, video generation is relatively underexplored by domain-agnostic methods. We compare our method with the representative domain-specific methods including TFGAN (Balaji et al., 2019), MMVID (Han et al., 2022a), CogVideo (Hong et al., 2023), VDM (Ho et al., 2022b), and Latte (Ma et al., 2024). We report Fréchet Video Distance (FVD) (Unterthiner et al., 2018), FID, and CLIPSIM (Wu et al., 2021), *i.e.*, the cosine similarity between the CLIP embeddings (Radford et al., 2021) of the generated images and the corresponding texts.

Our method achieves the comparable performance in both the video quality (FVD) and signal frame quality (FID) in Tab. 1, compared with the recent domain-specific text-to-video models. Moreover, our model learns more semantics as suggested by the CLIPSIM scores. The results show that our model, as a domain-*agnostic* method, can achieve a performance on par with domain-*specific* methods in modeling long-context. The qualitative comparisons in Fig. 5 further support our model in text-to-video generation compared with the recent state-of-the-art methods.



(a) VDM (Ho et al., 2022b) (b) CogVideo (Hong et al., 2023) (c) **Field-DiT** (Ours)

Figure 5: Qualitative comparisons between domain-specific text-to-video models and ours. Compared to VDM Ho et al. (2022b), our results are more continuous. Compared to CogVideo Hong et al. (2023), our results feature more realistic textures. Please see `https://kfmei.com/Field-DiT/` for the input prompt and video results.



(a) pixelNeRF (Yu et al., 2021) (b) Functa (Dupont et al., 2022b) (c) **Field-DiT** (Ours)

Figure 6: Qualitative comparisons between representative 3D novel view generation methods and ours. Our results demonstrate competitive quality without explicitly using 3D modeling.

PSNR (dB)	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
DPF (Zhuang et al., 2023)	24.00	21.97	20.87	20.66	✗	✗	✗	✗	✗	✗
Field-DiT (Ours)	44.30	43.96	43.87	44.16	42.92	42.20	42.42	42.51	42.07	42.22

Table 2: We demonstrate the long-context modeling capability of our model by showing its next-frame generation accuracy on game data, where a total of 100 frames are evaluated. ✗ denotes out-of-memory results when the model cannot handle such a long context.

3D novel views. We also evaluate our method on 3D novel view generation with the ShapeNet dataset (Chang et al., 2015). Specifically, we use the "car" class of ShapeNet which involves 3514 different cars. Each car object has 50 random viewpoints, where each viewpoint is in 128 × 128 resolution. Unlike previous domain-agnostic methods (Du et al., 2021; Zhuang et al., 2023) that model 3D geometry over voxel grids at 64^3 resolution, we model over rendered camera views based on their corresponding camera poses and intrinsic parameters, similar to recent domain-specific methods (Sitzmann et al., 2019; Yu et al., 2021). This approach allows us to extract more view-wise coordinate-signal pairs while voxel grids only have 6 views. We report our results in comparison with the state-of-the-art view-synthesis algorithms including pixelNeRF (Yu et al., 2021), viewFormer (Kulhánek et al., 2022), EG3D-PTI (Chan et al., 2022), and Zero-1-to-3 Liu et al. (2023). Among the compared methods, only Zero-1-to-3 is a zero-shot generation approach, while the others are trained on the ShapeNet dataset, either by their authors or by us. Zero-1-to-3 is pretrained on large-scale data, making it robust to out-of-domain data. Note that our model performs one-shot novel view synthesis by conditioning on the text embedding of a random view. Compared to recent methods specifically designed for 3D modalities, our approach achieves higher fidelity metrics, such as PSNR and SSIM, while producing comparable scores in LPIPS. Although methods like EG3D-PTI and Zero-1-to-3, which directly fine-tune pretrained 2D image generation models like StyleGAN and Stable-Diffusion, achieve better FID scores, this metric prioritizes 2D visual quality. However, it does not strictly reflect 3D consistency, which limits its relevance for 3D evaluation.

Concurrent 3D novel view generation methods, such as Eschernet (Kong et al., 2024), Wonder3D (Long et al., 2024), and SV3D (Voleti et al., 2025), have demonstrated their effectiveness in modeling sparse views of 3D objects. While our method may underperform in perceptual quality due to its modality-unified design, its unique ability to generate continuous, long-context 3D views offers a fresh perspective to the 3D modeling community.

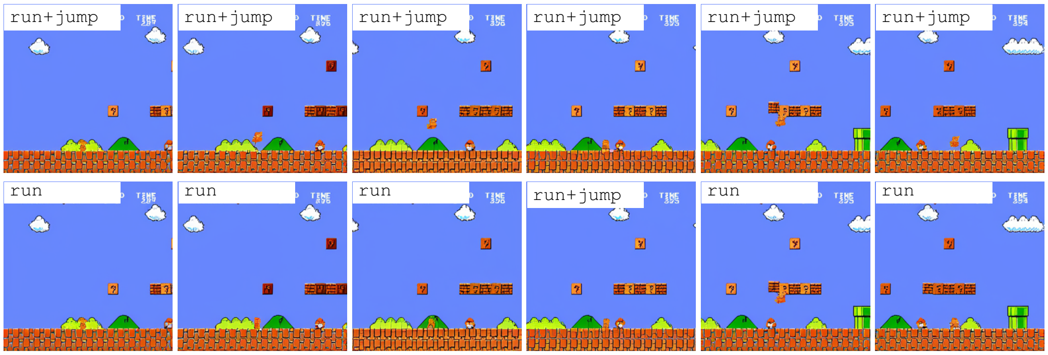


Figure 7: Visualization of our generated game (1/8 sampling rate at 50 frames), showcasing how our method generalizes to different actions within the same context. Each frame's action is labeled in the top-left corner. Please see `https://kfmei.com/Field-DiT/` for videos.

Text	Cross-view consistent noise	Resolution	Training Views n	FVD (↓)	FID (↓)	CLIPSIM (↑)	MACs	Mems
✗	N/A	16.0	8	608.27	34.10	-	113.31G	15.34Gb
✓	✗	16.0	8	401.64	75.81	0.198	117.06G	15.34Gb
✓	✓	1.0	8	115.20	40.34	0.187	7.314T	22.99Gb
✓	✓	16.0	1	320.02	21.27	0.194	117.06G	15.34Gb
✓	✓	16.0	4	89.83	23.69	0.194	117.06G	15.34Gb
✓	✓	16.0	8	42.03	24.33	0.220	117.06G	15.34Gb

Table 3: Ablation analysis of the text-to-video results of our proposed method under different settings. All computation costs (MACs) and GPU memory usage (Mems) are estimated for generating a single view, regardless of the resolution, to ensure a fair comparison. The mark in the text column indicates whether a text prompt is used. The number in the resolution column denotes the usage of a latent encoder, where a resolution equal to 1 means the model is directly trained in pixel space.

Games. Game generation is an under-explored area and lacks data and benchmarks. We demonstrate the game generation capability of our method by showing the accuracy of predicted frames compared with the frame of the real game when using the same action. Specially, we model the World 1-1 of Super Mario Bros (NES version) with a sliding window size of 16, and we test it with new actions for next-frame generation. Fig. 7 shows the visual results generated from two different actions starting from the same scene. Tab. 2 demonstrates our long-context modeling capability compared with the DPF, where ours performance loss is minor compared with DPF.

4.1 Ablations and Discussions

In this section, we demonstrate the effectiveness of each of our proposed components and analyze their contributions to the quality of the final result, as well as the computation cost. The quantitative text-to-video generation results under various settings are shown in Table 3.

Effect of text condition. To verify the effectiveness of the text condition for capturing the global geometry of the data, we use two additional settings. **(1)** The performance of our model when the text condition is removed is shown in the first row of Tab. 3. The worse FVD means that the text condition play a crucial role in preserving the global geometry, specifically the spatial-temporal coherence in videos. **(2)** When the text condition is added, but not the cross-view consistent noise, the results can be found in the second row of Tab. 3. The FVD is slightly improved compared to the previous setting, but the FID is weakened due to underfitting against cross-view inconsistent noises. In contrast to our default setting, these results demonstrate the effectiveness of the view-consistent noise. Furthermore, we note that more detailed text descriptions can significantly improve the generated video quality.

Effect of number of views. We investigate the model performance change with varying number of views (n) for representing fields, as shown in the 2nd and 3rd rows of Tab. 3. Compared to the default setting of $n = 8$, reducing n to 1 leads to non-continuous frames and abrupt identity changes, as indicated by the low FVD. When n is increased to 4, the continuity between frames is improved, but still worse than $n = 8$ for the dynamics between frames. Thus, we can conclude that a larger number of views leads to a higher performance, along with a higher computation cost.

Comparison with Context Query Pairs. Even though context query pairs introduced by DPF (Zhuang et al., 2023) has been justified to achieve better performance than using latent space (which needs reconstruction training) in small models and low-resolution modalities, it is shown (Zhuang et al., 2023) to be impossible to largely reduce the memory footprint (by sampling less context pairs) while preserving its original modeling capability and performance. To scale up our model, we replace the context query pairs with latent space in our method. It can significantly reduce memory usage (*e.g.* using less than 2% pairs while maintaining a competitive performance) so that handling a larger model size becomes possible with high-resolution, long views. Based on these, the benefit of scaling using the latent space outweighs the potential performance loss led by the latent space, as backed by Tab. 1.

Comparison with Modality Unified Models. Our method shares the motivation of modality-unified models like SIREN and Functa for handling diverse data modalities but differs in complexity and scope. SIREN uses sinusoidal activations in MLPs to represent continuous signals, excelling in modeling structured data and solving mathematical problems like PDEs with high fidelity but is limited to simpler datasets due to its MLP architecture. In contrast, our diffusion transformer framework handles more diverse and complex data, integrating view-wise sampling for local structure and autoregressive generation for global consistency. Additionally, text and past frame conditioning enable Field-DiT to scale effectively to complex multi-modal tasks, making it more versatile for dynamic and high-dimensional datasets compared to SIREN's structured focus.

5 Limitations.

(1) Our method only applies to visual modalities interpretable by views. For modalities such as temperature manifold Hersbach et al. (2019) where there is no "views" of such field, our method does not apply. As long as the data in the new domain (e.g., 3D dynamic scene and MRI) can be interpreted by views, our method can reuse the same latent autoencoder Rombach et al. (2022) without switching to domain-specific autoencoders. (2) Our method aligns with the standard practice outlined in DPF, using comparison methods with weights trained separately for each modality. While it performs exceptionally well on individual modalities, achieving strong performance across multiple modalities simultaneously is hindered by the inherent challenges of the multi-task problem. We believe our approach provides a solid foundation for the future versatile generative models.

6 Conclusion

We have introduced a new transformer-based diffusion field model that addresses the limitations of current probabilistic field models in capturing global structures and long-context dependencies. By utilizing a view-wise sampling algorithm for local structure learning and incorporating autoregressive generation to preserve global geometry, our approach overcomes the shortcomings of MLP-based architectures. The proposed model can generate high-fidelity data across multiple modalities, including text-to-video, 3D view generation, and game control while maintaining scalability and unifying diverse modalities.

Acknowledgments. Vishal M. Patel was supported by NSF CAREER award 2045489.

References

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 2021.

Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional GAN with discriminative filter generation for text-to-video synthesis. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 2019.

Matthias Bauer, Emilien Dupont, Andy Brock, Dan Rosenbaum, Jonathan Schwarz, and Hyunjik Kim. Spatial functa: Scaling functa to imagenet classification and generation. *ArXiv preprint*, 2023.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2022.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *ArXiv preprint*, 2015.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021.

Yilun Du, Katie Collins, Josh Tenenbaum, and Vincent Sitzmann. Learning signal-agnostic manifolds of neural fields. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021.

Emilien Dupont, Hyunjik Kim, S. M. Ali Eslami, Danilo Jimenez Rezende, and Dan Rosenbaum. From data to functa: Your data point is a function and you can treat it like one. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research, 2022a.

Emilien Dupont, Hyunjik Kim, SM Eslami, Danilo Rezende, and Dan Rosenbaum. From data to functa: Your data point is a function and you should treat it like one. *ArXiv preprint*, 2022b.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, (11), 2020.

Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris N. Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2022a.

Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris N. Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2022b.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent Video Diffusion Models for High-Fidelity Video Generation with Arbitrary Lengths. *ArXiv preprint*, 2022.

H Hersbach, B Bell, P Berrisford, G Biavati, A Horányi, J Muñoz Sabater, J Nicolas, C Peubey, R Radu, I Rozum, et al. Era5 monthly averaged data on single levels from 1979 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*, 2019.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, and David J. Fleet. Imagen video: High definition video generation with diffusion models. *ArXiv preprint*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, 2022b.

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.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. Eschernet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9503–9513, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Jonáš Kulhánek, Erik Derner, Torsten Sattler, and Robert Babuška. Viewformer: Nerf-free neural rendering from few images using transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*. Springer, 2022.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9970–9980, 2024.

Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017.

Kangfu Mei and Vishal Patel. Vidm: Video implicit diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 9117–9125, 2023.

Kangfu Mei, Mauricio Delbracio, Hossein Talebi, Zhengzhong Tu, Vishal M Patel, and Peyman Milanfar. Codi: Conditional diffusion distillation for higher-fidelity and faster image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9048–9058, 2024a.

Kangfu Mei, Luis Figueroa, Zhe Lin, Zhihong Ding, Scott Cohen, and Vishal M Patel. Latent feature-guided diffusion models for shadow removal. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4313–4322, 2024b.

Kangfu Mei, Nithin Gopalakrishnan Nair, and Vishal M Patel. Improving conditional diffusion models through re-noising from unconditional diffusion priors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025a.

Kangfu Mei, Zhengzhong Tu, Mauricio Delbracio, Hossein Talebi, Vishal M. Patel, and Peyman Milanfar. Bigger is not always better: Scaling properties of latent diffusion models. *Transactions on Machine Learning Research*, 2025b. ISSN 2835-8856.

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

Joaquin Quinonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 2005.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 2020.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv preprint*, 2022.

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*, 2022.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016.

Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *ArXiv preprint*, 2018.

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pp. 439–457. Springer, 2025.

Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *ArXiv preprint*, 2021.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2021.

Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023a.

Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023b.

Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023c.

Peiye Zhuang, Samira Abnar, Jiatao Gu, Alex Schwing, Joshua M. Susskind, and Miguel Ángel Bautista. Diffusion Probabilistic Fields. In *International Conference on Learning Representations*, 2023.