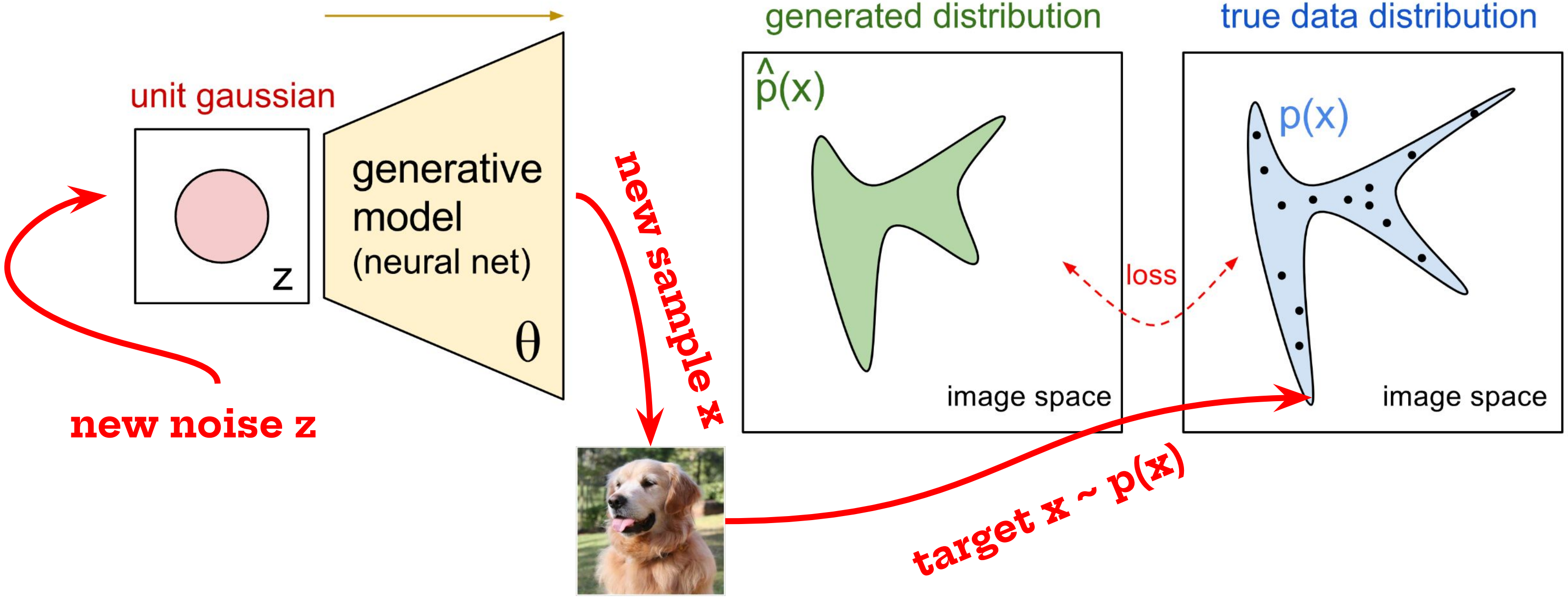
Field-DiT: Diffusion Transformer on Unified Video, 3D, and Game Field Generation

Kangfu Mei*   Mo Zhou   Vishal M. Patel

Johns Hopkins University, ECE Department
* Now at Google Research

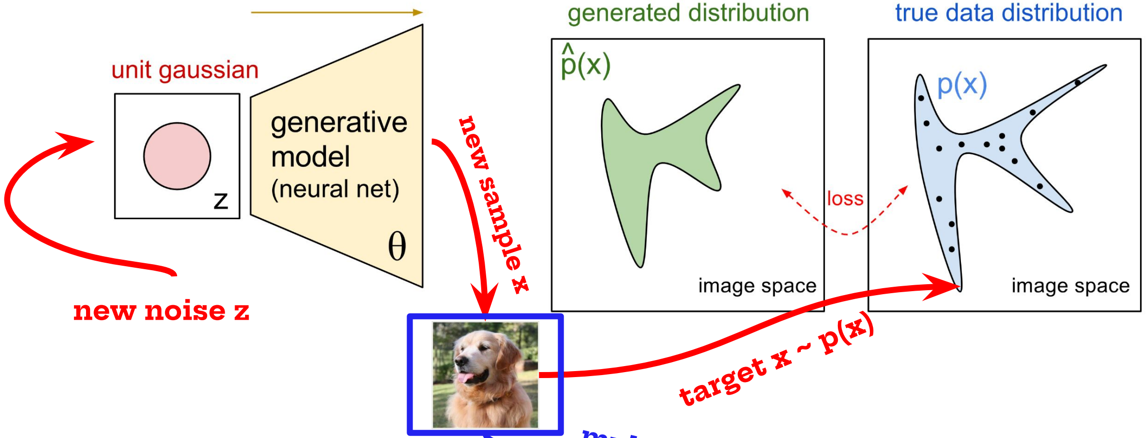
Background: Quadratic complexity in multimodal modeling

- Generative models learn to replicate the distribution of training data for creating new samples.

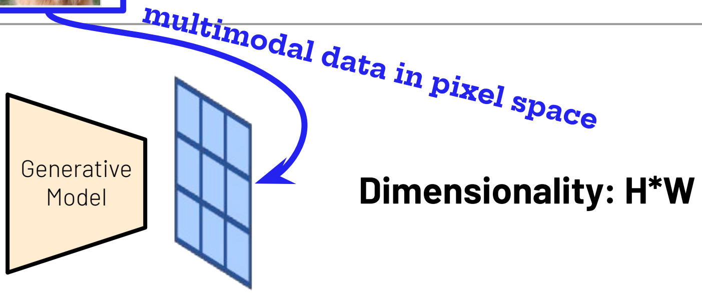


Background: Quadratic complexity in multimodal modeling

- Generative models learn to replicate the distribution of training data for creating new samples.

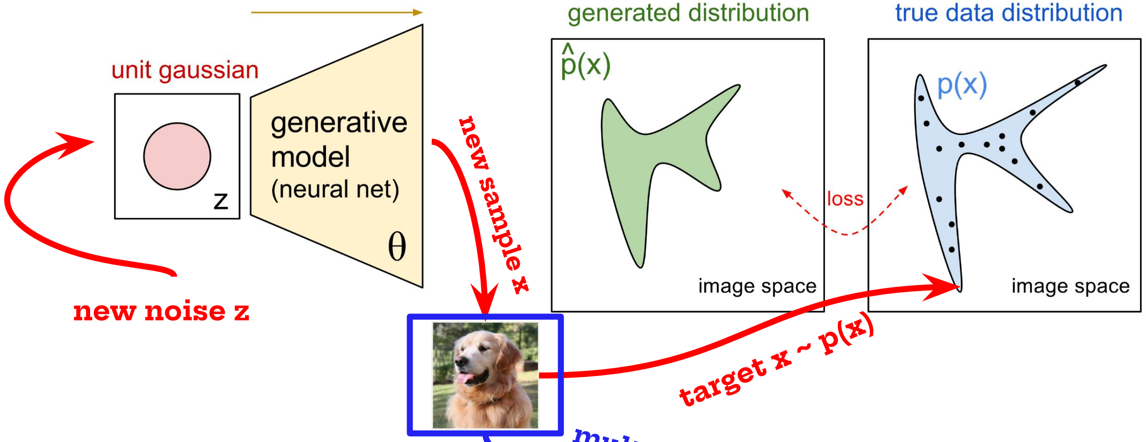


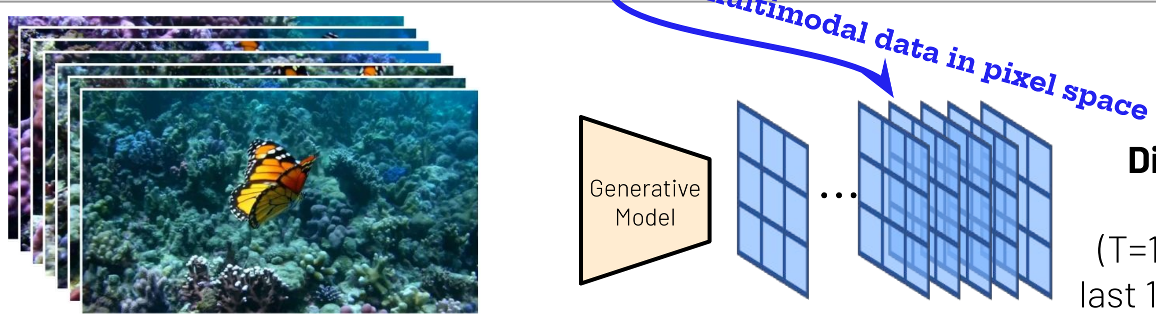
(a). Single-frame Generation



Background: Quadratic complexity in multimodal modeling

- Generative models learn to replicate the distribution of training data for creating new samples.





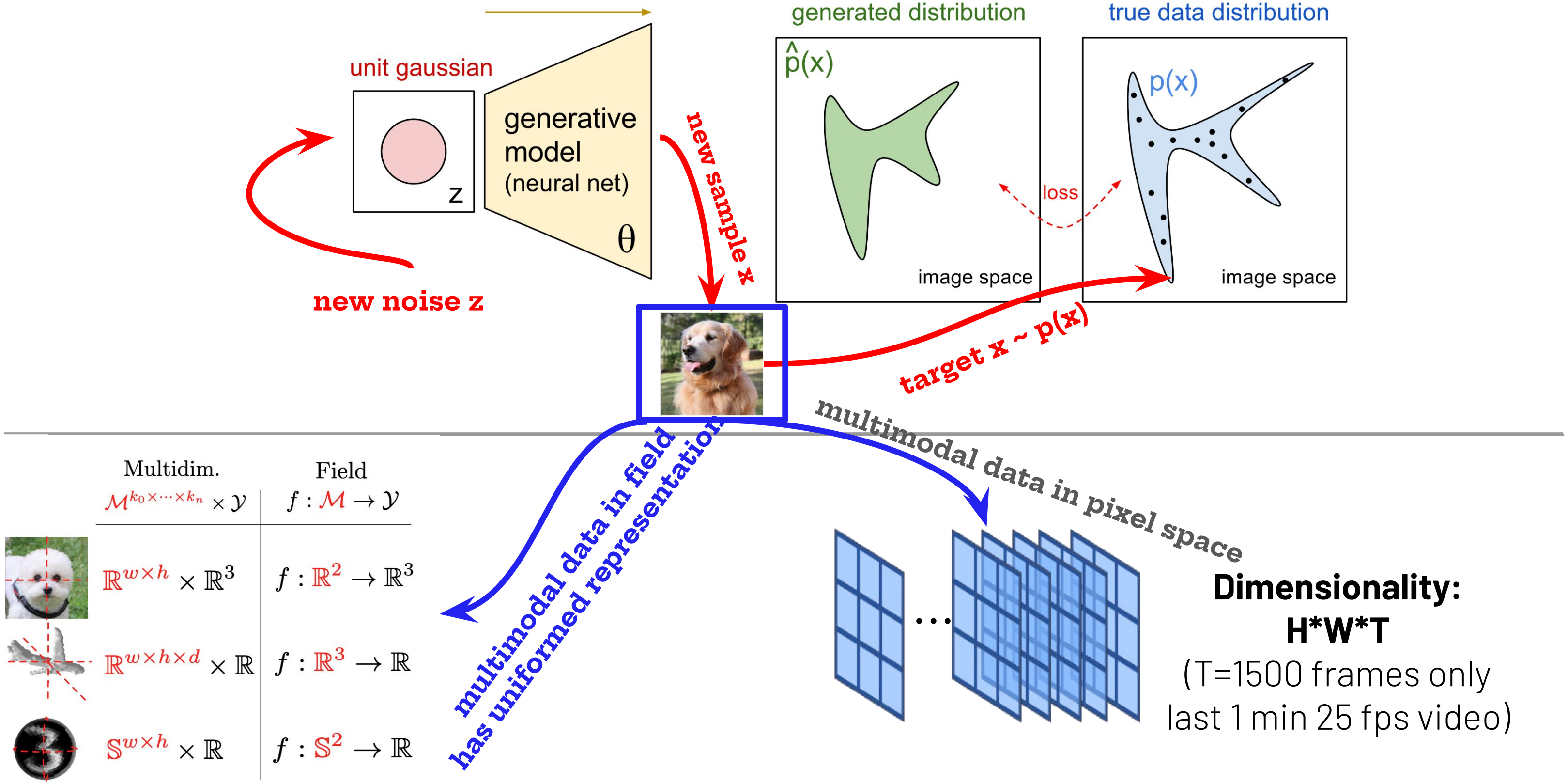
(b). Multi-frame Generation

Dimensionality: H*W*T

(T=1500 frames only last 1 min 25 fps video)

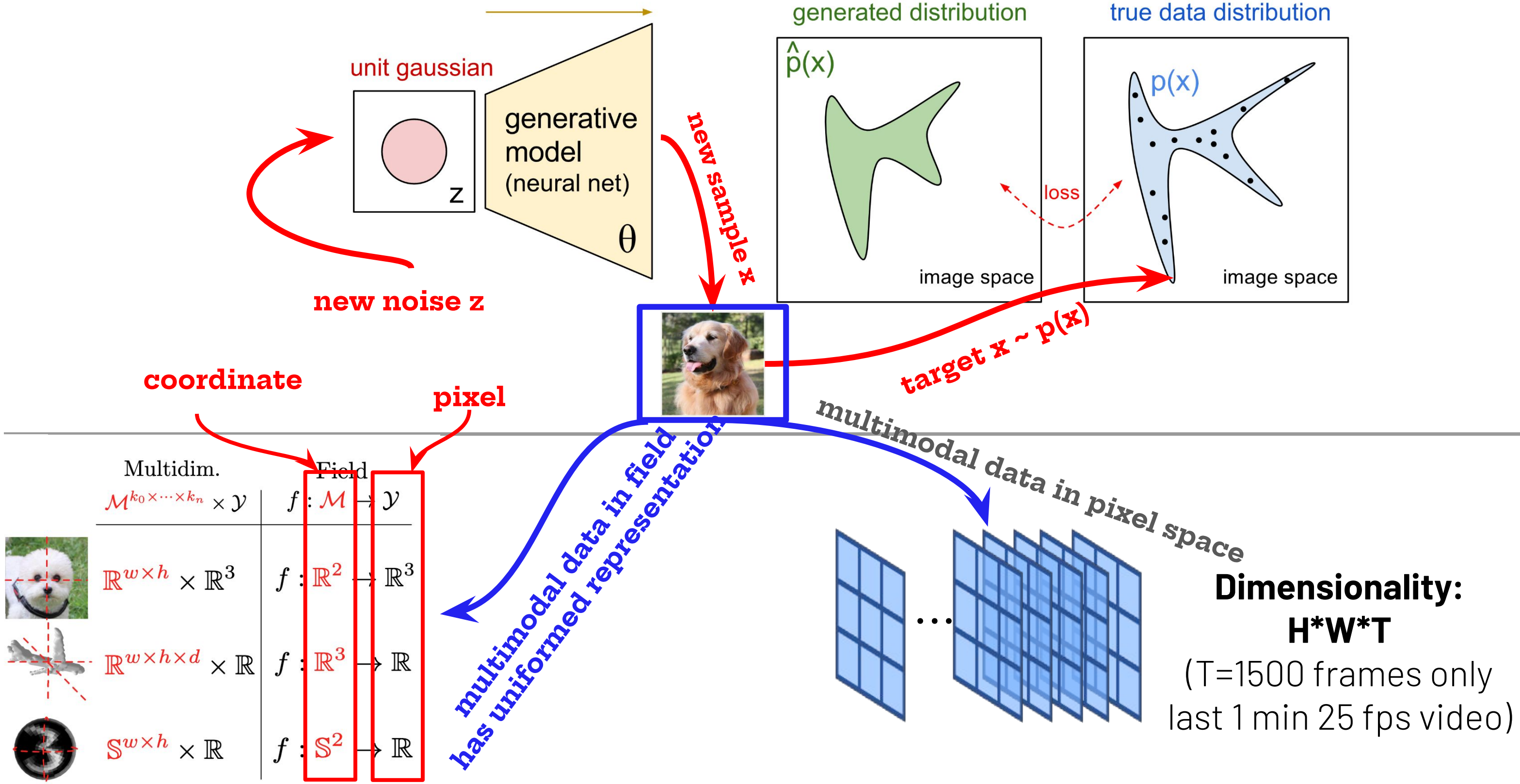
Background: Quadratic complexity in multimodal modeling

- Generative models learn to replicate the distribution of training data for creating new samples.



Background: Quadratic complexity in multimodal modeling

- Generative models learn to replicate the distribution of training data for creating new samples.



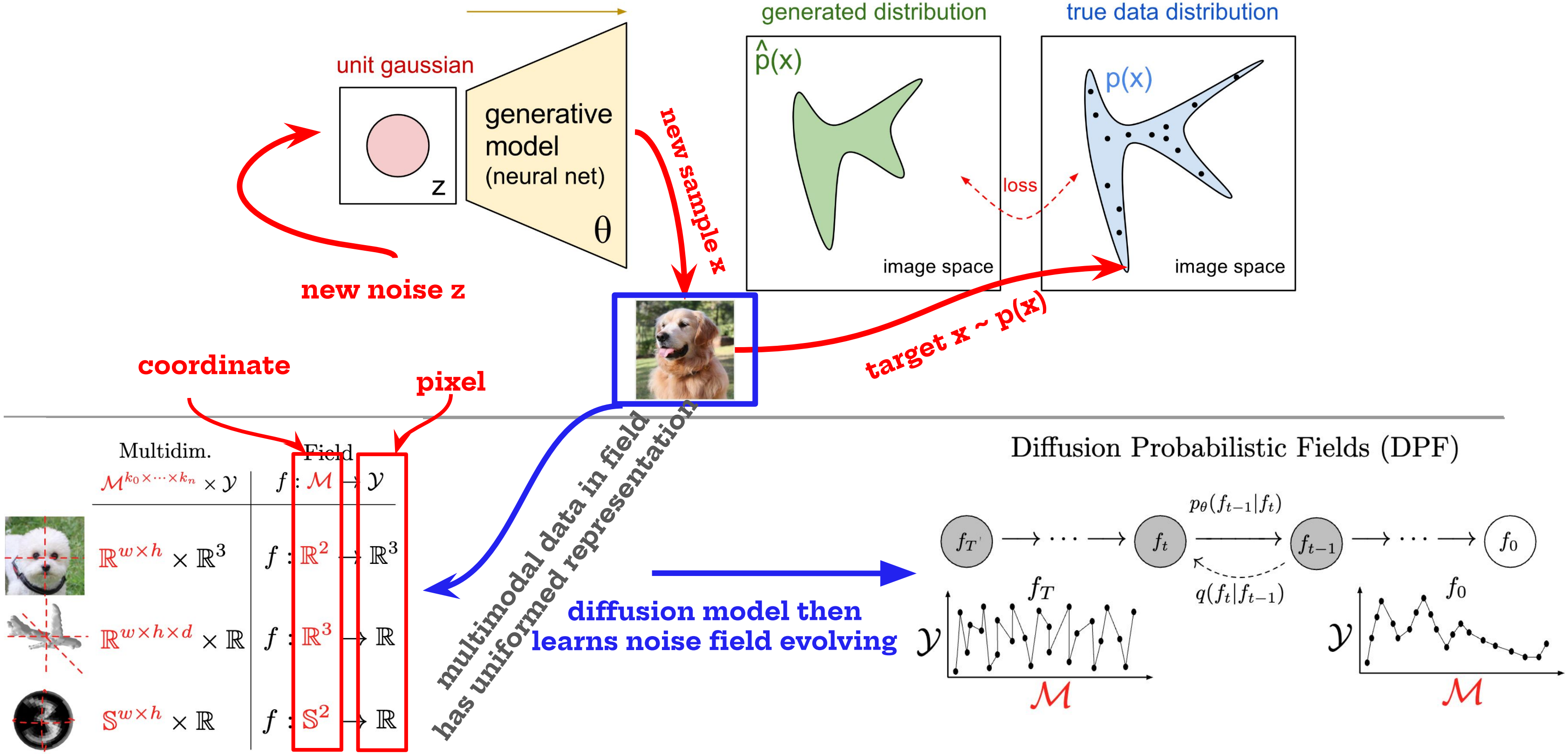
Multidim. $\mathcal{M}^{k_0\times\cdots\times k_n} \times \mathcal{Y}$	Field $f : \mathcal{M} \to \mathcal{Y}$
$\mathbb{R}^{w\times h} \times \mathbb{R}^3$	$f : \mathbb{R}^2 \to \mathbb{R}^3$
$\mathbb{R}^{w\times h\times d} \times \mathbb{R}$	$f : \mathbb{R}^3 \to \mathbb{R}$
$\mathbb{S}^{w\times h} \times \mathbb{R}$	$f : \mathbb{S}^2 \to \mathbb{R}$

Dimensionality: H*W*T

(T=1500 frames only last 1 min 25 fps video)

Background: Quadratic complexity in multimodal modeling

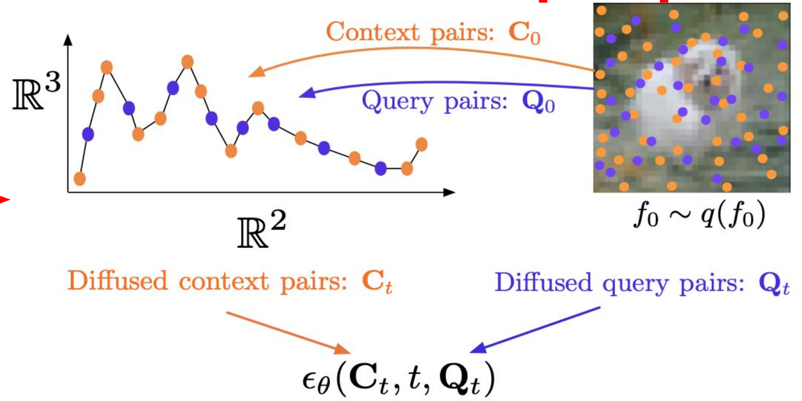
- Generative models learn to replicate the distribution of training data for creating new samples.



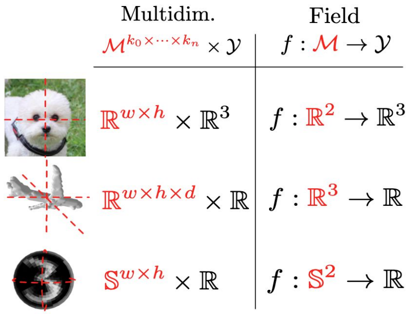
Background: Quadratic complexity in multimodal modeling

- Generative models learn to replicate the distribution of training data for creating new samples.

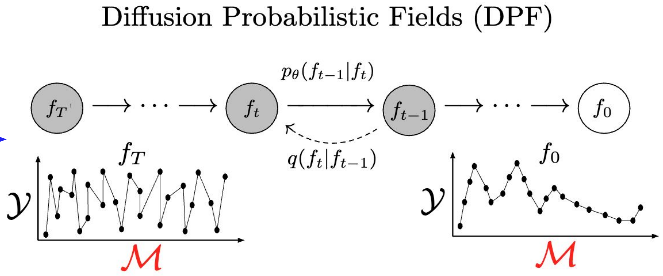
sparse representation



DPF only models a portion of the data sample

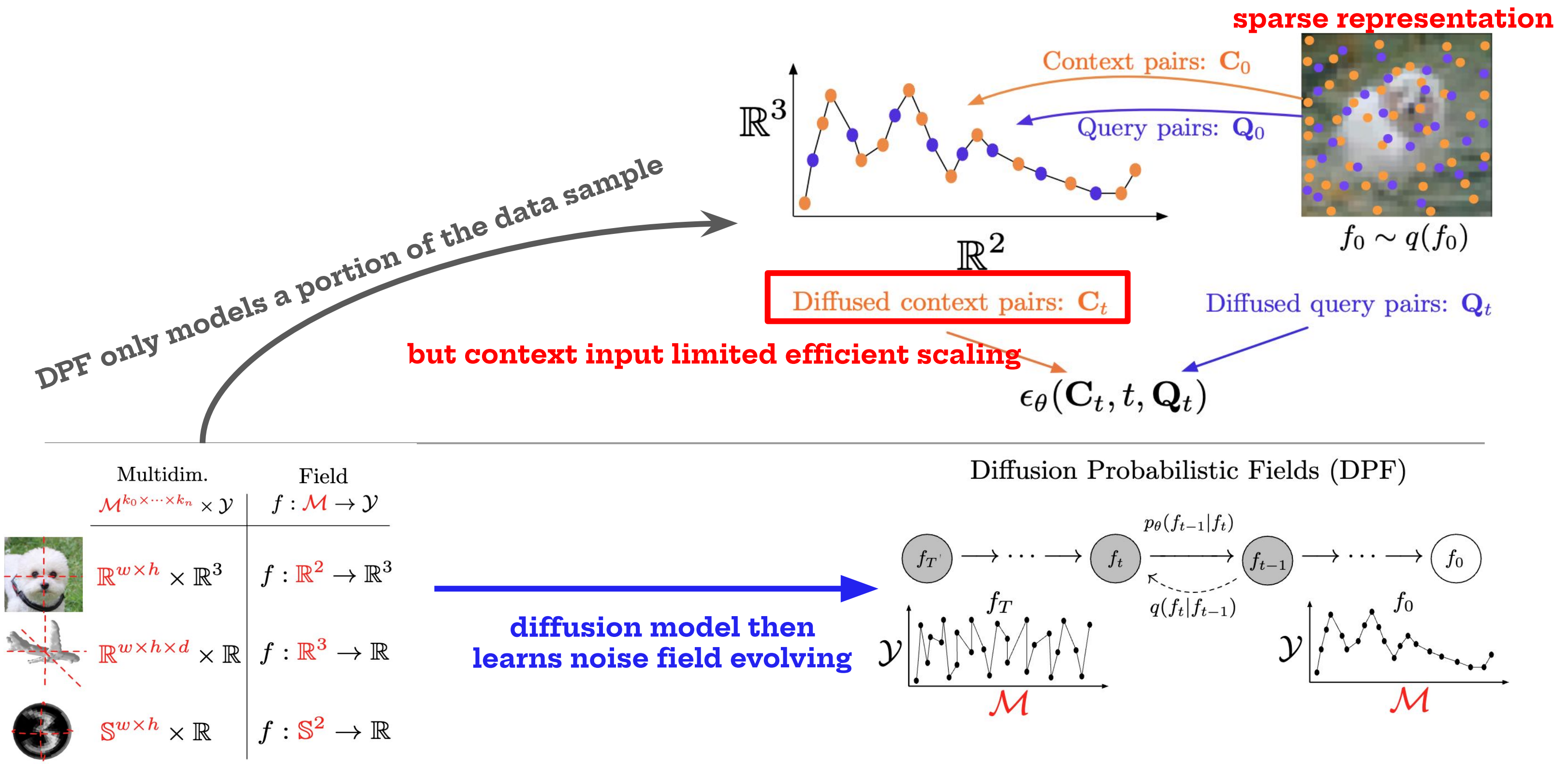


diffusion model then learns noise field evolving



Background: Quadratic complexity in multimodal modeling

- Generative models learn to replicate the distribution of training data for creating new samples.



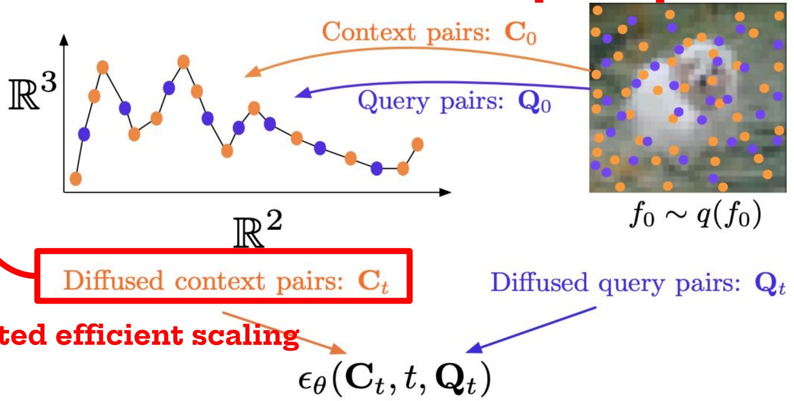
Background: Quadratic complexity in multimodal modeling

- Generative models learn to replicate the distribution of training data for creating new samples.

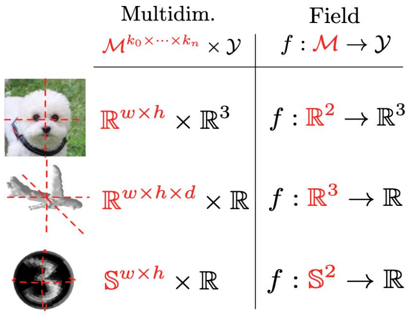
CelebA-HQ 64^2	FID
# context pairs = 1028 (50%)	103.86
# context pairs = 2048 (75%)	91.12
# context pairs = 4096 (100%)	74.89

DPF comprises the compute for modeling sparse representation performance with more context

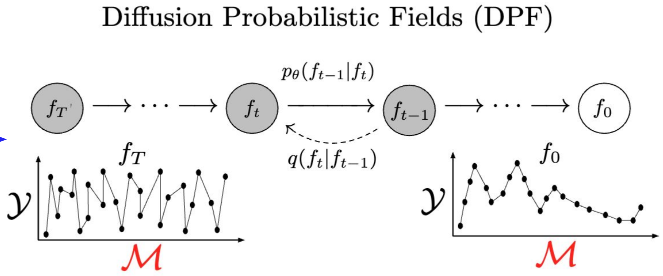
sparse representation



but context input limited efficient scaling



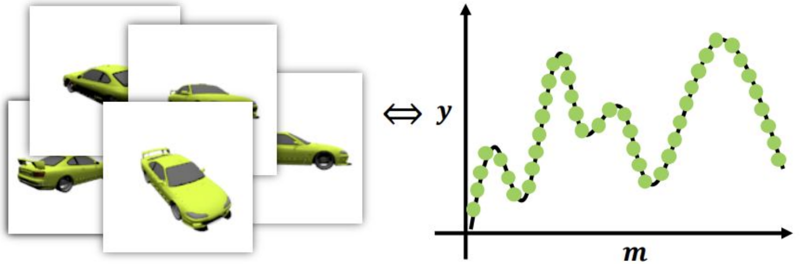
diffusion model then learns noise field evolving



Intuition: Efficient Representation and Context

- Scaling generative filed models by using **efficient representation** of field and context input.

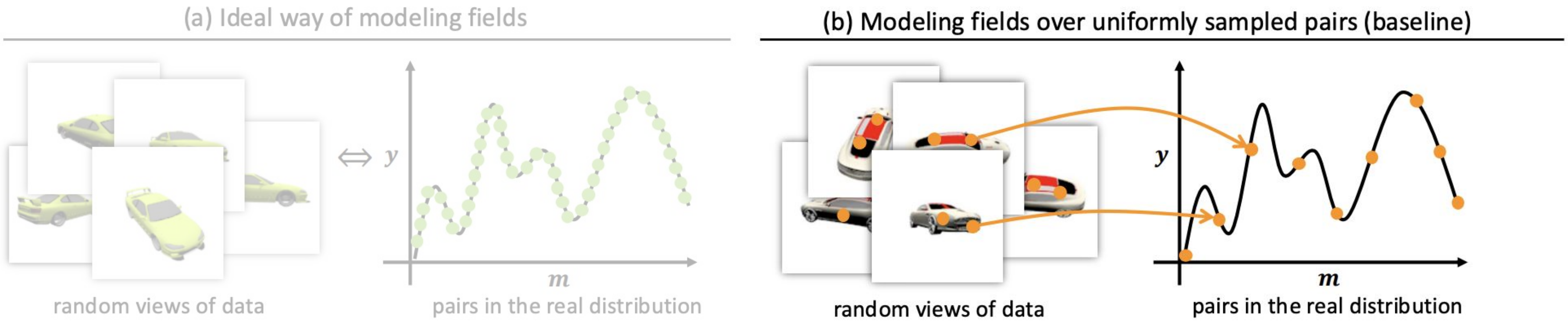
(a) Ideal way of modeling fields



Standard generative (diffusion) model suffers from dense data modeling

Intuition: Efficient Representation and Context

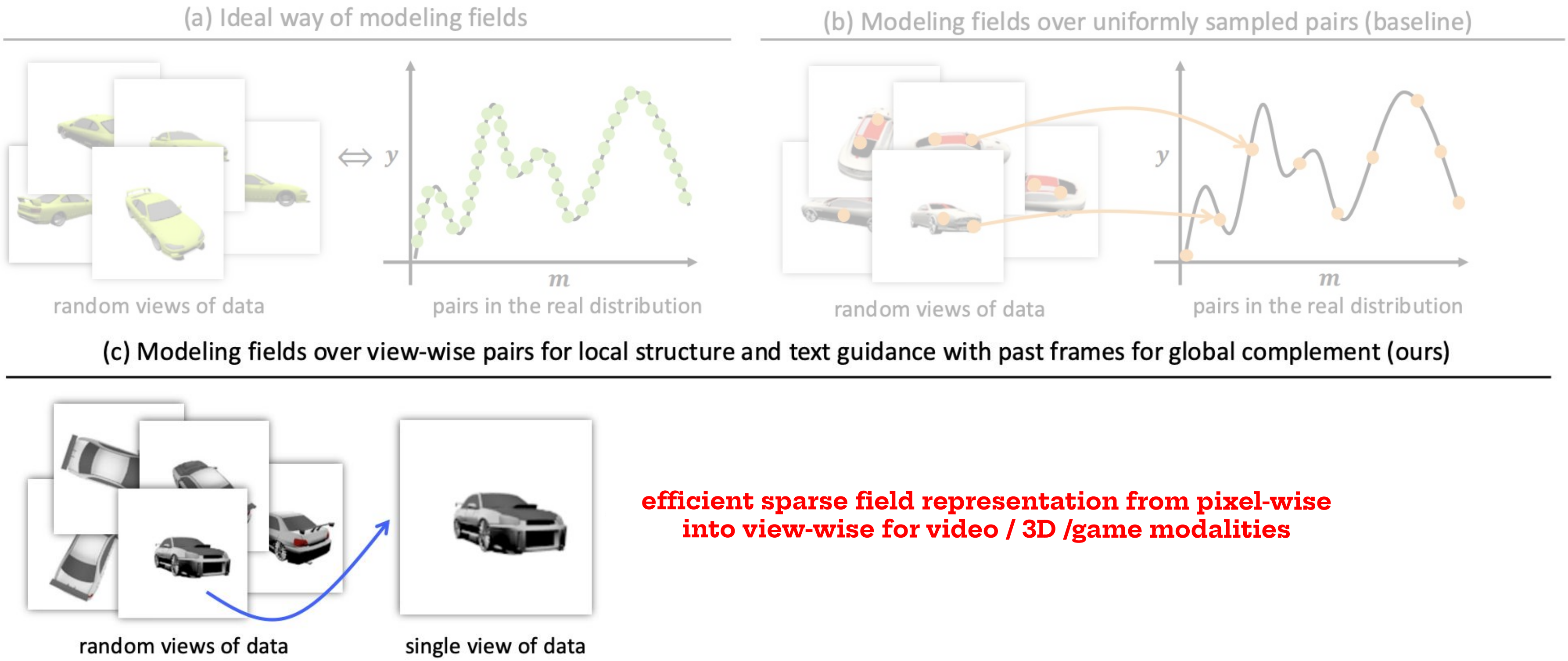
- Scaling generative filed models by using **efficient representation** of field and context input.



Field model is flexible by using sparse representation but suffers from low-quality

Intuition: Efficient Representation and Context

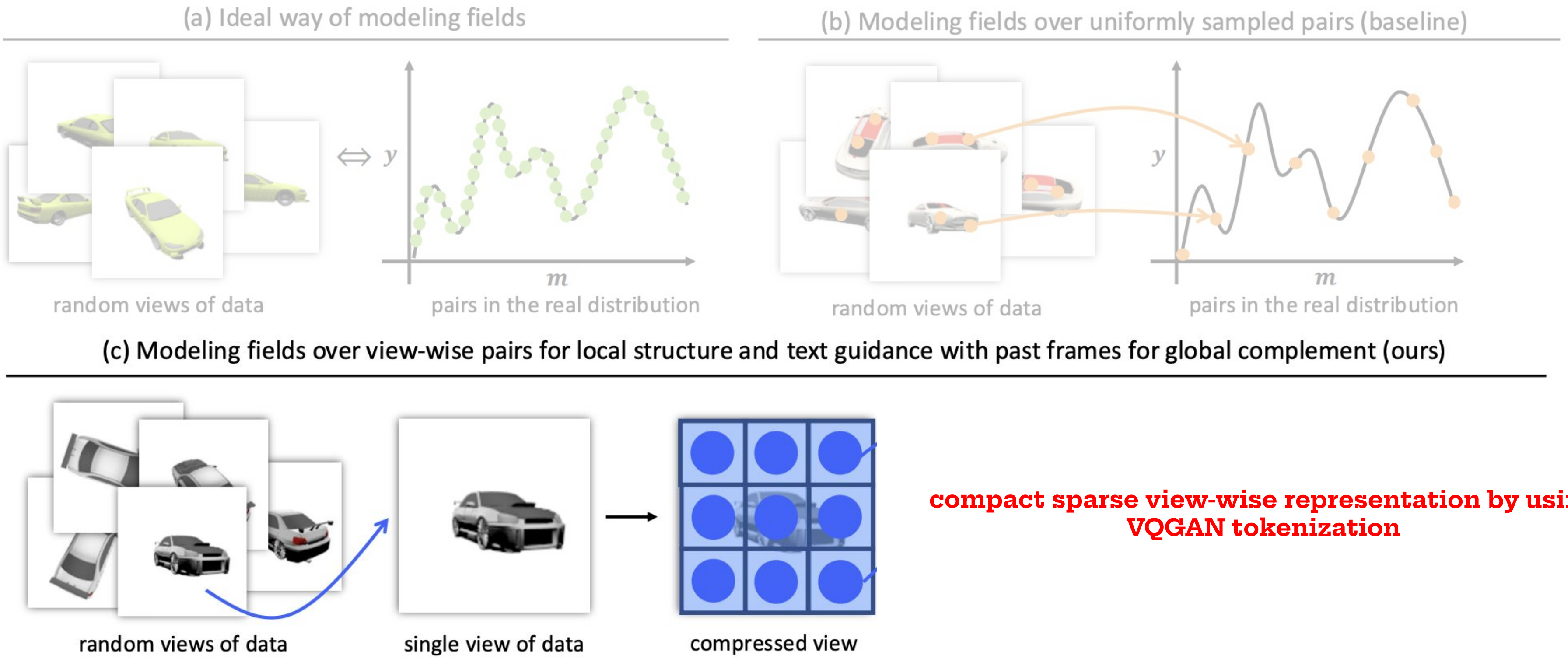
- Scaling generative filed models by using **efficient representation** of field and context input.



efficient sparse field representation from pixel-wise into view-wise for video / 3D /game modalities

Intuition: Efficient Representation and Context

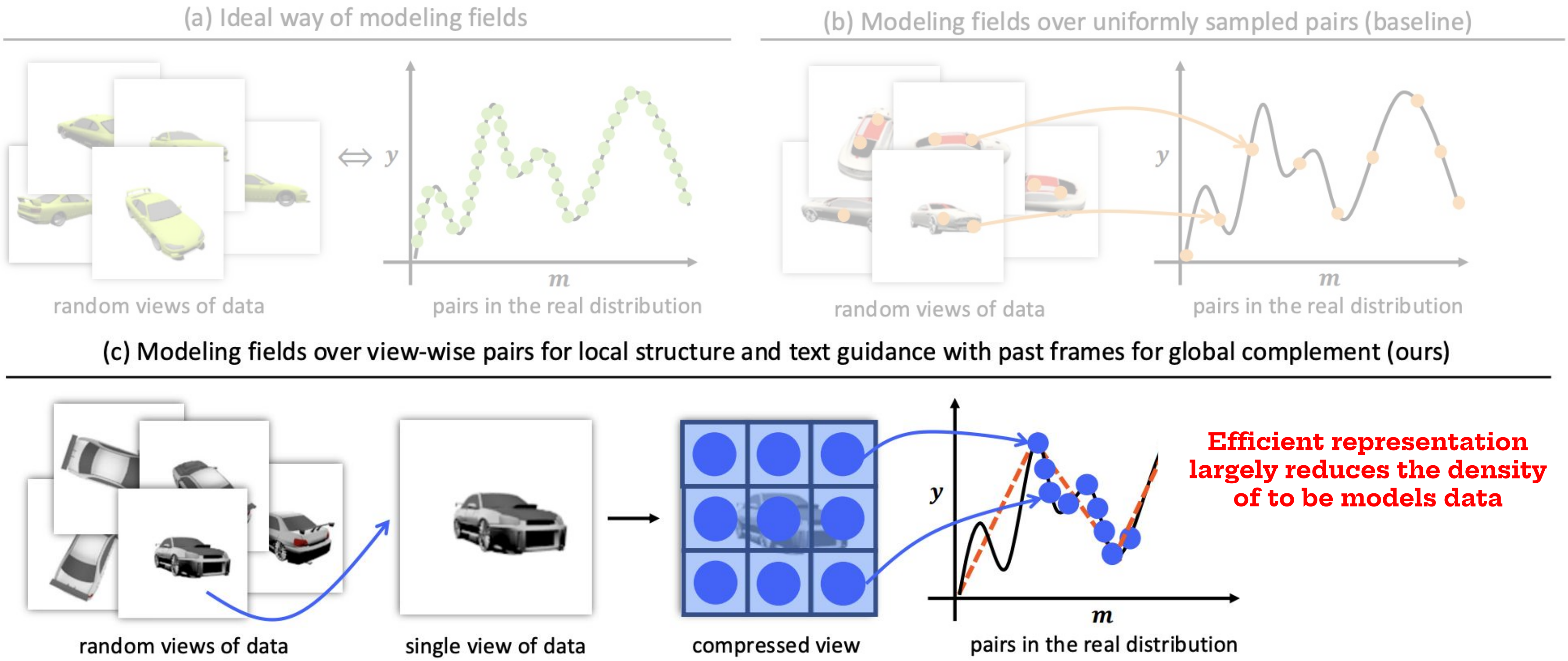
- Scaling generative filed models by using **efficient representation** of field and context input.



compact sparse view-wise representation by using VQGAN tokenization

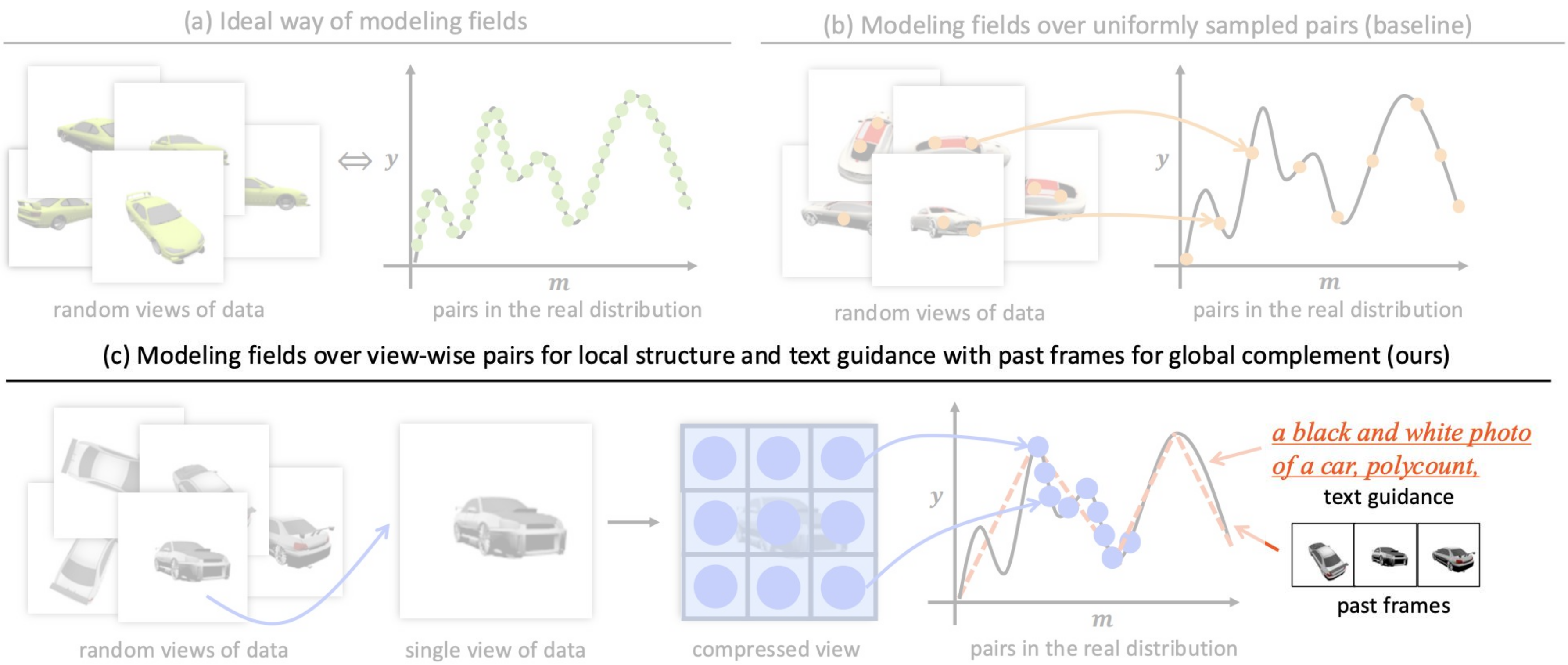
Intuition: Efficient Representation and Context

- Scaling generative filed models by using **efficient representation** of field and context input.



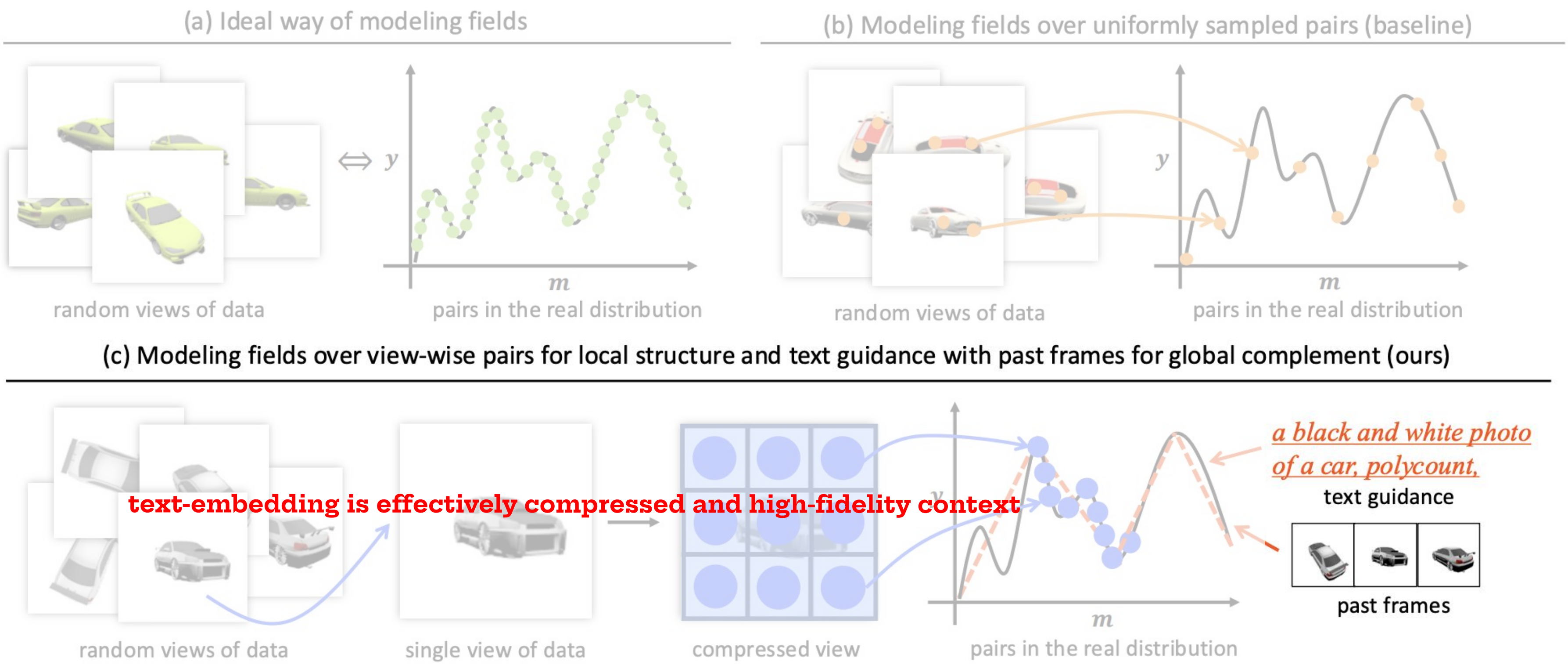
Intuition: Efficient Representation and Context

- Scaling generative filed models by using efficient representation of field and **context input**.



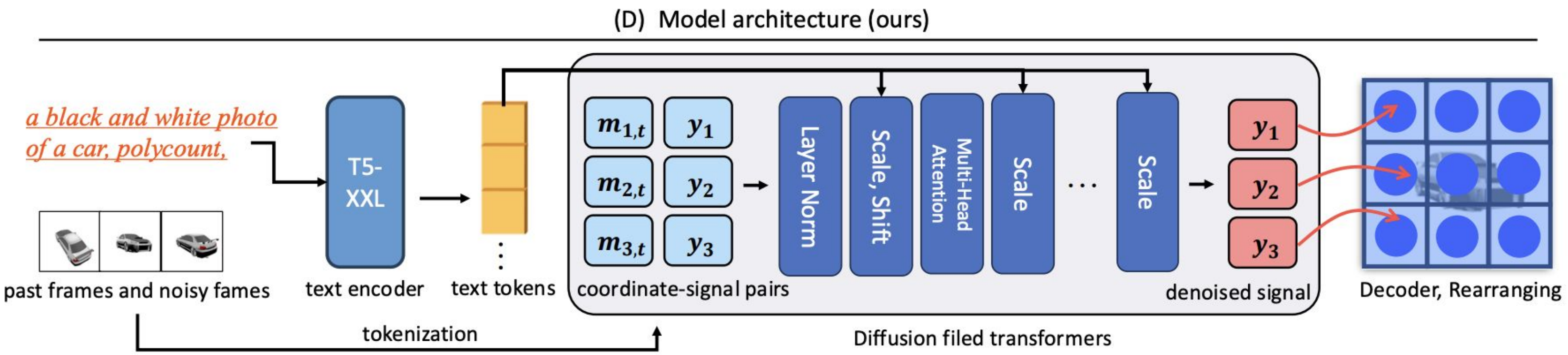
Intuition: Efficient Representation and Context

- Scaling generative filed models by using efficient representation of field and **context input**.



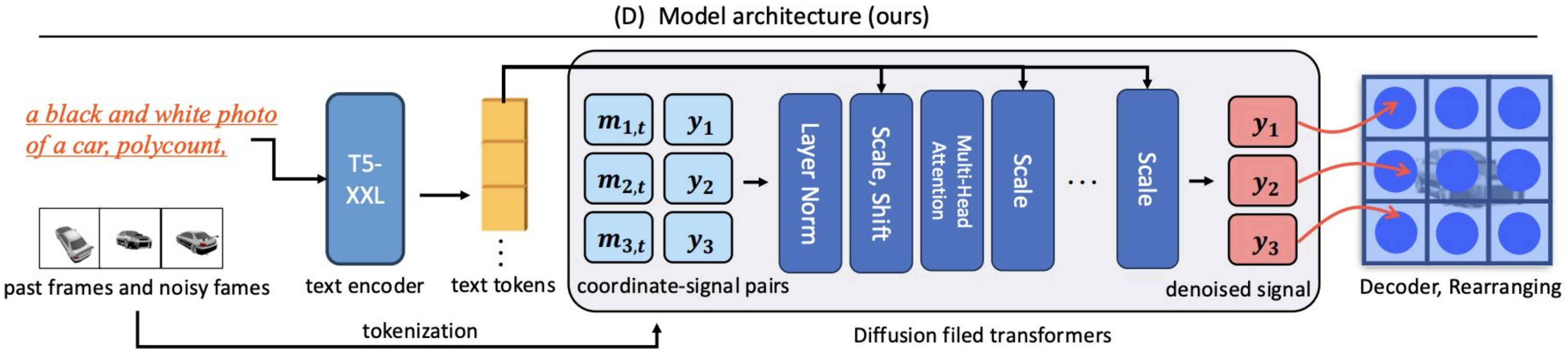
Method: Field-DiT with view-wise and global geometry modeling

- During training, Field-DiT learns to denoise a few consecutive (8) frames of the data that is modality-invariant, with text-embedding and coordinate embedding as conditions



Method: Field-DiT with view-wise and global geometry modeling

- During training, Field-DiT learns to denoise a few consecutive (8) frames of the data that is modality-invariant, with text-embedding and coordinate embedding as conditions



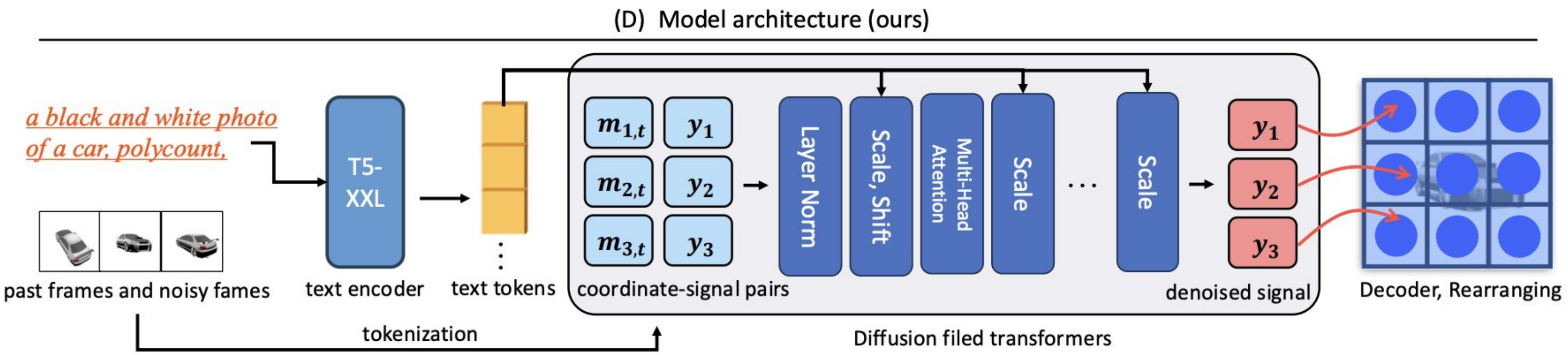
$$\mathbf{Q} = \Big\{ \underbrace{\{(\mathbf{m}_i, \mathbf{y}_{(i,t)}) | i = 1, 2, \ldots, H{\cdot}W\}}_{\text{pairs from the } n\text{-th view}} \ \Big| \ n = 1, 2, \ldots, N \Big\}$$

$$= \Big\{ \{(\mathbf{m}_{(i,n)}, \mathbf{y}_{(i,n,t)} = \sqrt{\bar{\alpha}}\mathbf{y}_{(i,n,0)} + \sqrt{1-\bar{\alpha}_t}\epsilon_i) | i{=}1, 2, \ldots, H{\cdot}W\} \ \Big| \ n{=}1, 2, \ldots, N \Big\}.$$

the stochastic variable (noise) keeps consistent across diffusion views of the same data

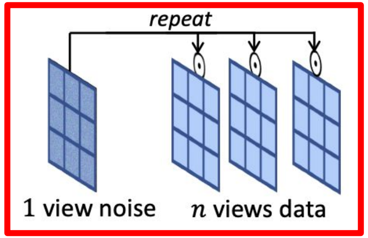
Method: Field-DiT with view-wise and global geometry modeling

- During training, Field-DiT learns to denoise a few consecutive (8) frames of the data that is modality-invariant, with text-embedding and coordinate embedding as conditions



$$\mathbf{Q} = \Big\{ \underbrace{\{(\mathbf{m}_i, \mathbf{y}_{(i,t)}) | i = 1, 2, \ldots, H \cdot W\}}_{\text{pairs from the } n\text{-th view}} \;\Big|\; n = 1, 2, \ldots, N \Big\}$$

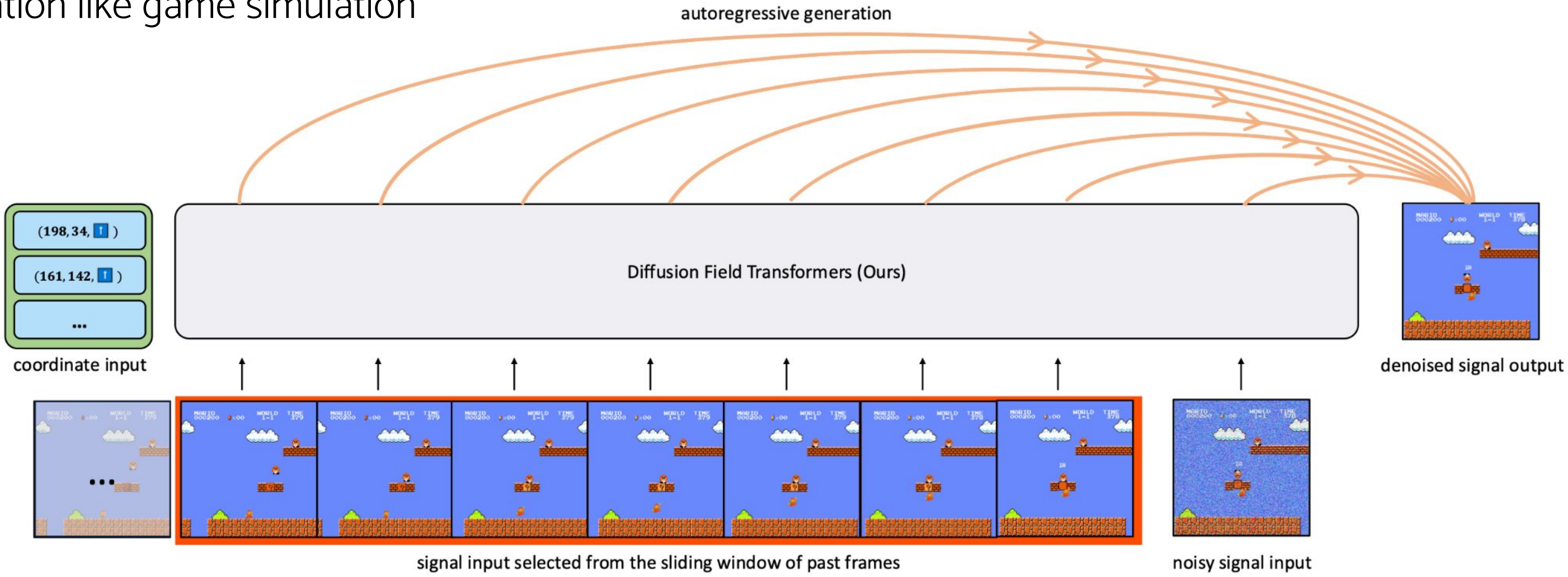
$$= \Big\{ \{(\mathbf{m}_{(i,n)}, \mathbf{y}_{(i,n,t)} = \sqrt{\bar{\alpha}} \mathbf{y}_{(i,n,0)} + \sqrt{1 - \bar{\alpha}_t \epsilon_i}) | i{=}1, 2, \ldots, H \cdot W\} \;\Big|\; n{=}1, 2, \ldots, N \Big\}.$$



the stochastic variable (noise) keeps consistent across diffusion views of the same data

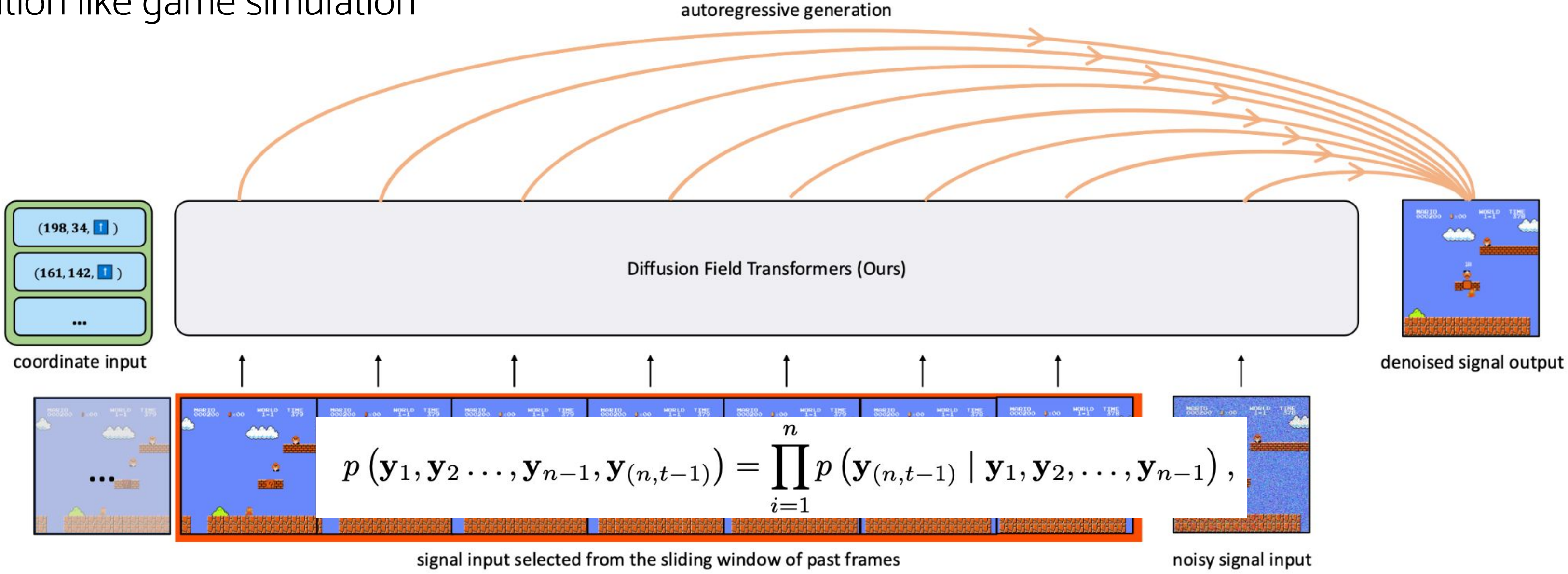
Method: Field-DiT with view-wise and global geometry modeling

- During training, Field-DiT learns to denoise a few consecutive (8) frames of the data that is modality-invariant, with text-embedding and coordinate embedding as conditions
- Historic frames conditioning is proposed with fix-length sliding window for extreme long-context generation like game simulation



Method: Field-DiT with view-wise and global geometry modeling

- During training, Field-DiT learns to denoise a few consecutive (8) frames of the data that is modality-invariant, with text-embedding and coordinate embedding as conditions
- Historic frames conditioning is proposed with fix-length sliding window for extreme long-context generation like game simulation



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

Experimental Results

*See paper for experimental settings

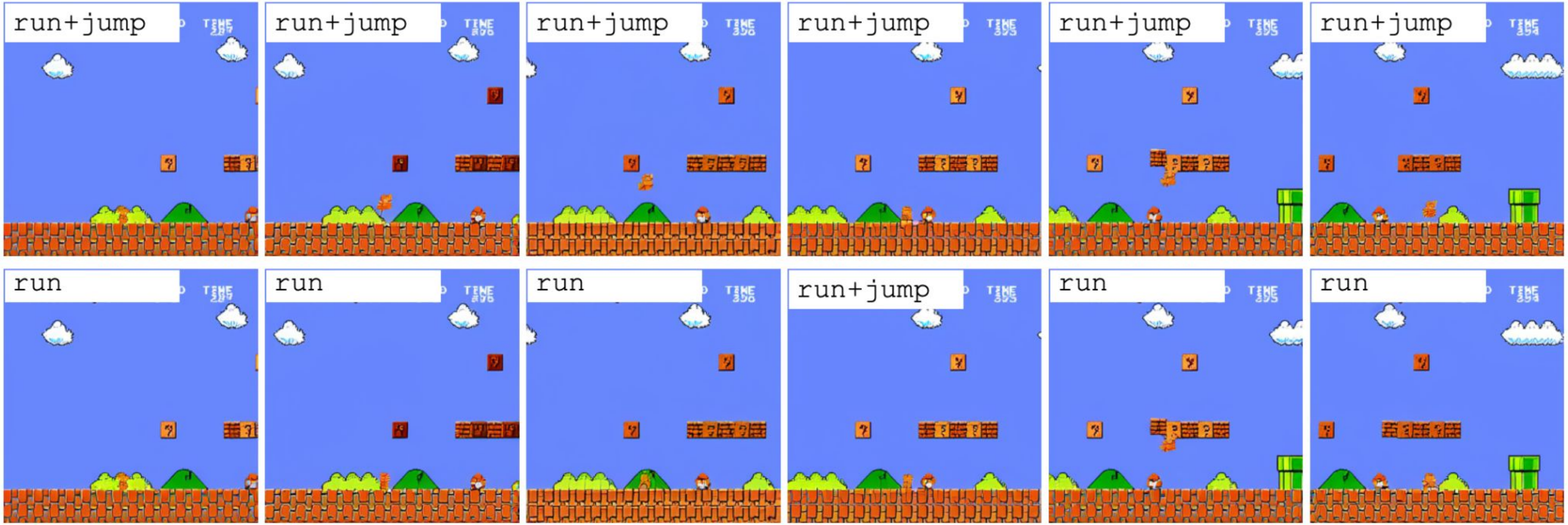
- Field-DiT outperforms the diffusion field baseline and related modality-invariant modeling method.

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**

Experimental Results

*See paper for experimental settings

- Field-DiT outperforms the diffusion field baseline and related modality-invariant modeling method.
- Field-DiT simulates game from action input with the same architecture as the video generation.



Experimental Results

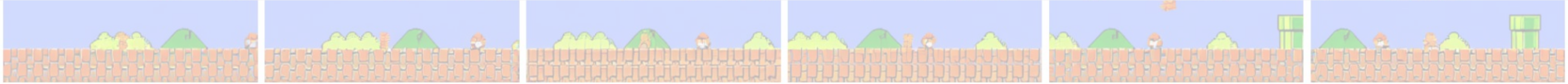
*See paper for experimental settings

- Field-DiT outperforms the diffusion field baseline and related modality-invariant modeling method.
- Field-DiT simulates game from action input with the same architecture as the video generation.

Field-DiT replicates the game with time-invariant accuracy

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.



Experimental Results

*See paper for experimental settings

- Field-DiT outperforms the diffusion field baseline and related modality-invariant modeling method.
- Field-DiT simulates game from action input with the same architecture as the video generation.

Experimental Results

***See paper for experimental settings**

- Field-DiT outperforms the diffusion field baseline and related modality-invariant modeling method.
- Field-DiT simulates game from action input with the same architecture as the video generation.

She has wavy hair and high cheekbones. To begin with, this female talks for a short time, and she then talks for a short time, next she talks for a short time, in the end, she talks for a short time.

Experimental Results

*See paper for experimental settings

- Field-DiT outperforms the diffusion field baseline and related modality-invariant modeling method.
- Field-DiT simulates game from action input with the same architecture as the video generation.

Takeaways

- Sparsity trade-offs the continuous long-context modeling for efficiency but language context naturally comprises the global geometry.
- Unifying different modalities through modeling providing unique priors that are unattainable from single modality modeling (e.g. video prior for game simulation)

Thank you :)
For more details,
please come to our **poster session**!

Code and models:

https://github.com/MKFMIKU/Field-DiT