NÜWA: Visual Synthesis Pre-training for Neural visUal World creAtion

Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, Nan Duan

Microsoft Research Asia, and Peking University

Arxiv

Presenter: Minho Park

Examples

- Fine-tune with the presented pre-trained model NUWA.

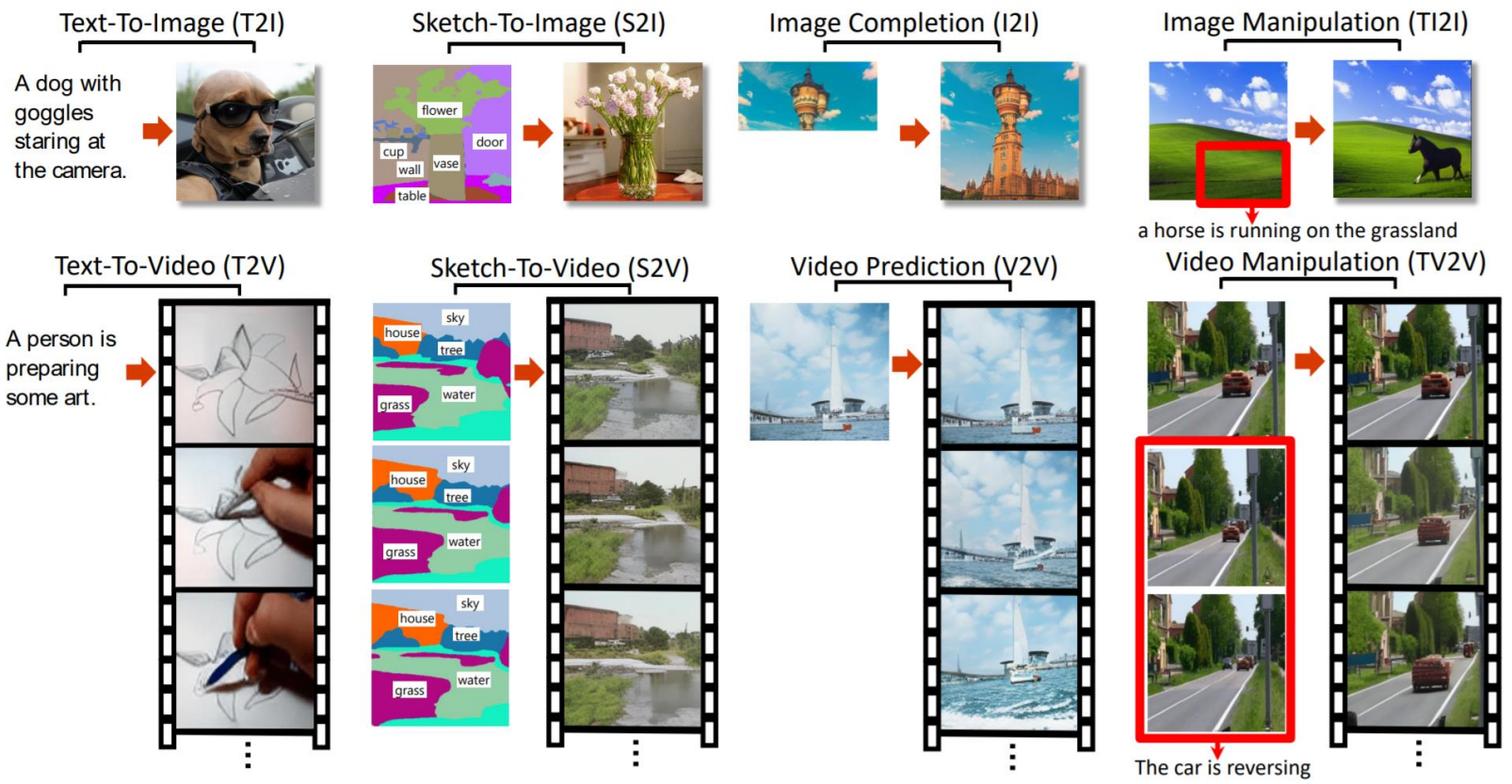
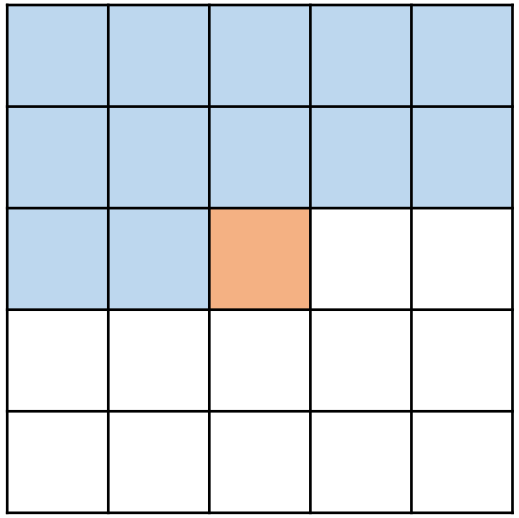


Figure 1. Examples of 8 typical visual generation and manipulation tasks supported by the NUWA model.

Auto-regressive Models

- Generative model (GAN, VAE, and AR)
- PixelCNN / PixelRNN
- Image / Video Transformer, iGPT (Image GPT)
- Problem: high computational cost on high dimensional visual data



$$p(x) = \prod_{i=1}^{n} p(x_i | x_1, \ldots, x_{i-1})$$

VQ-VAE-based Tokenization

- VQ-VAE
- DALL-E and CogView: images
- GODIVA: videos

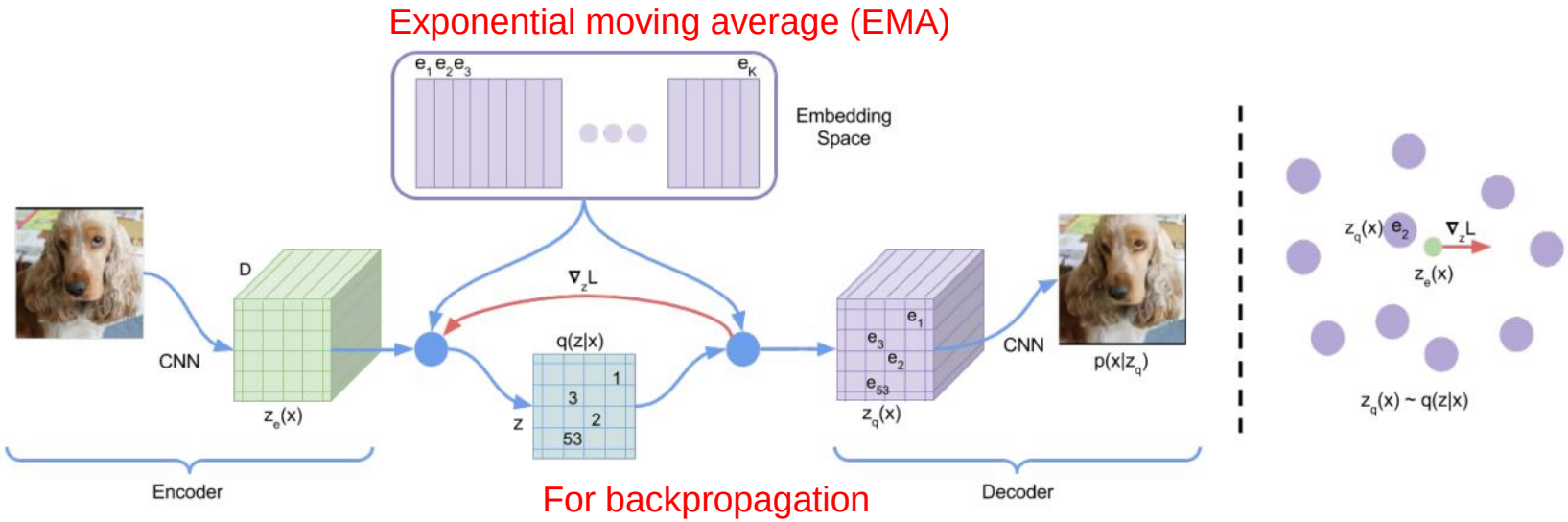


Figure 1: Left: A figure describing the VQ-VAE. Right: Visualisation of the embedding space. The output of the encoder $z(x)$ is mapped to the nearest point e_2. The gradient $\nabla_z L$ (in red) will push the encoder to change its output, which could alter the configuration in the next forward pass.

Deep Foundations, Vector Quantized VAEs, https://www.youtube.com/watch?v=52G6WhBDQ3c

3D(4D) Data Representation

- A unified 3D notation $X \in \mathbb{R}^{h \times w \times s \times d}$.
  - h, w: spatial axis
  - s: temporal axis
  - d: dimension of tokens
- Video: $\mathbb{R}^{H \times W \times s \times d}$
- Image: $\mathbb{R}^{H \times W \times 1 \times C}$
- Text: $\mathbb{R}^{1 \times 1 \times s \times d}$
- 1 is placeholder.

Method

- A unified multimodal pre-trained model called NUWA

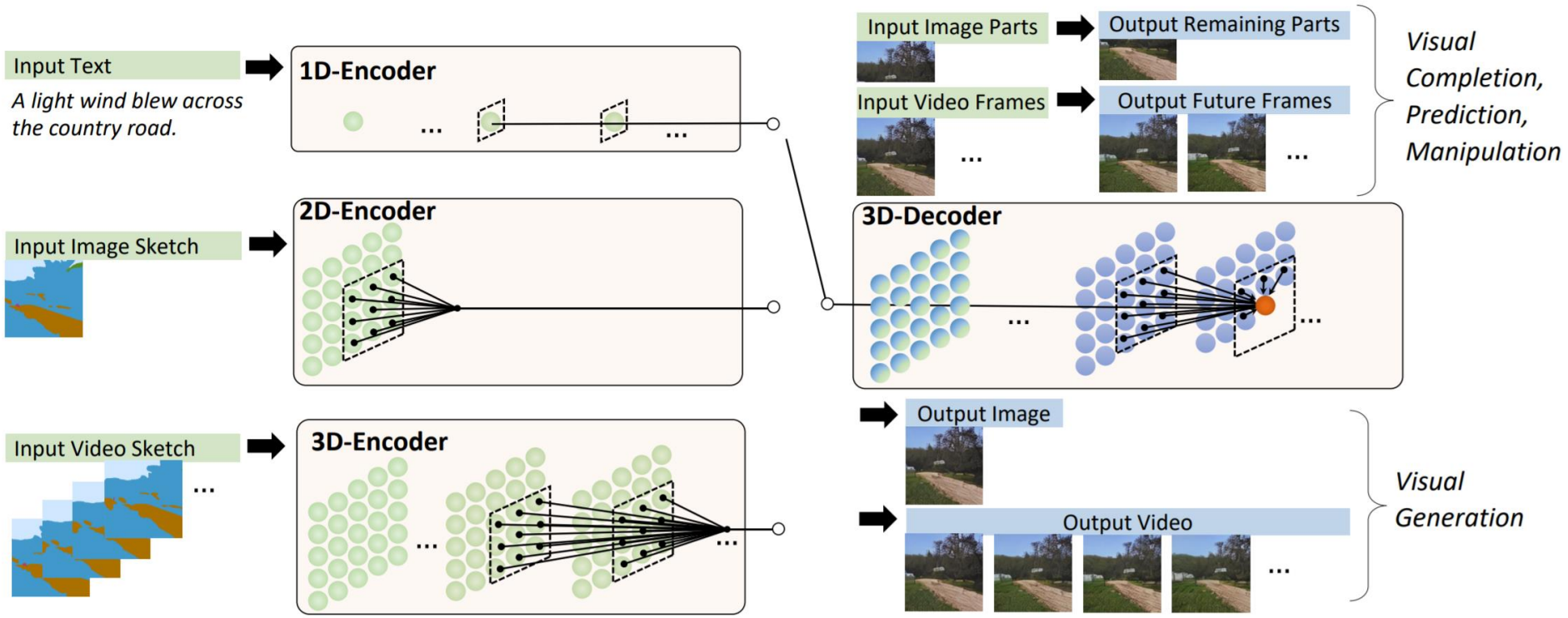


Figure 2. Overview structure of NUWA.

Train VQ-VAE

- $z_i = \underset{j}{\operatorname{argmin}} \left\| E(I)_i - B_j \right\|$
- $\hat{I} = G(B[z])$
  - $E(I) \in \mathbb{R}^{h \times w \times d_B}$
  - $B \in \mathbb{R}^{N \times d_B}$: learnable codebook
- Training loss of VQ-VAE

$$L^V = \left\| I - \hat{I} \right\|_2^2 + \left\| sg[E(I)] - B[z] \right\|_2^2 + \beta \left\| E(I) - sg\left[B[z]\right] \right\|_2^2$$

Stop gradient: standard technique for stability

- Adding a perceptual loss and a GAN loss (VQ-GAN)

$$L^P = \left\| CNN(I) - CNN\left(\hat{I}\right) \right\|_2^2$$

$$L^G = \log D(I) + \log\left(1 - D\left(\hat{I}\right)\right)$$

Focus on high-level semantic

3D Nearby Self-Attention

- $Y = 3DNA(X, C; W)$
  - $X \in \mathbb{R}^{h \times w \times s \times d^{in}}, C \in \mathbb{R}^{h' \times w' \times s' \times d^{in}}$: 3D representations
  - W: learnable weights
- $C = X, C \neq X$: self- and cross-attention on target X conditioned on C.
- Local neighborhood around $(i', j', k') = (\left\lfloor i\frac{h'}{h} \right\rfloor, \left\lfloor j\frac{w'}{w} \right\rfloor, \left\lfloor k\frac{s'}{s} \right\rfloor)$ with extents $e^w, e^h, e^s \in \mathbb{R}^+$ is (not exponential)

$$N^{(i,j,k)} = \{C_{abc} \big| |a - i'| \leq e^h, |b - j'| \leq e^w, |c - k'| \leq e^s\} \in \mathbb{R}^{e^h \times e^w \times e^s \times d^{in}}$$

- $W^Q, W^K, W^V \in \mathbb{R}^{d^{in} \times d^{in}}$

3D Nearby Self-Attention

$$Q^{(i,j,k)} = XW^Q$$
$$K^{(i,j,k)} = N^{(i,j,k)}W^K$$
$$V^{(i,j,k)} = N^{(i,j,k)}W^V$$

$$y_{ijk} = softmax\left(\frac{Q^{(i,j,k)}\left(K^{(i,j,k)}\right)^T}{\sqrt{d^{in}}}\right)V^{(i,j,k)}$$

- Reduce the complexity: $O\left((hws)^2\right) \rightarrow O\left((hws)\left(e^h e^w e^s\right)\right)$
- Shows superior performance.

3D Encoder-Decoder

- 3D encoder-decoder built based on 3DNA
- To generate a target $Y \in \mathbb{R}^{h \times w \times s \times d^{out}}$ under the condition of $C \in \mathbb{R}^{h' \times w' \times s' \times d^{in}}$

1. Positional encoding

$$Y_{ijk} \coloneqq Y_{ijk} + P_i^h + P_j^w + P_k^s$$
$$C_{ijk} \coloneqq C_{ijk} + P_i^{h'} + P_j^{w'} + P_k^{s'}$$

2. Self-attention with a stack of L 3DNA layers

$$C^{(l)} = 3DNA\left(C^{(l-1)}, C^{(l-1)}\right)$$

3. Self- and cross-attention with a stack of L 3DNA layers

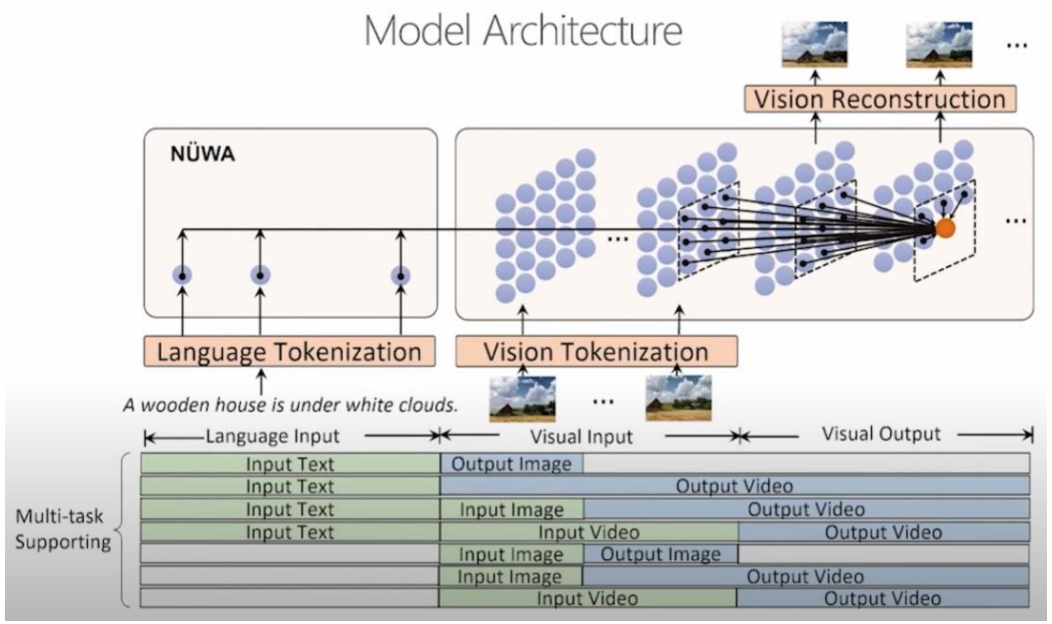
$$Y_{ijk}^{(l)} = 3DNA\left(Y_{<i,j,k}^{(l-1)}, Y_{<i,j,k}^{(l-1)}\right) + 3DNA\left(Y_{<i,j,k}^{(l-1)}, C^{(L)}\right)$$

masked

Training Objective

- Train on three tasks, Text-to-Image (T2I), Video Prediction (V2V), and Text-to-Video (T2V).

$$\mathcal{L} = \underbrace{-\sum_{t=1}^{h\times w} \log p_\theta(y_t|y_{<t}, C^{text}; \theta)}_{\text{Text-to-Image (T2I)}} \underbrace{-\sum_{t=1}^{h\times w\times s} \log p_\theta(y_t|y_{<t}, \text{"None"}; \theta)}_{\text{Video Prediction (V2V)}} \underbrace{-\sum_{t=1}^{h\times w\times s} \log p_\theta(y_t|y_{<t}, C^{text}; \theta)}_{\text{Text-to-Video (T2V)}}$$



Microsoft Research, Research talk: NUWA: Neural visual world creation with multimodal pretraining, https://youtu.be/jhmJ5qb-JAU

Implementation Details

- A unified multimodal pre-trained model called NUWA

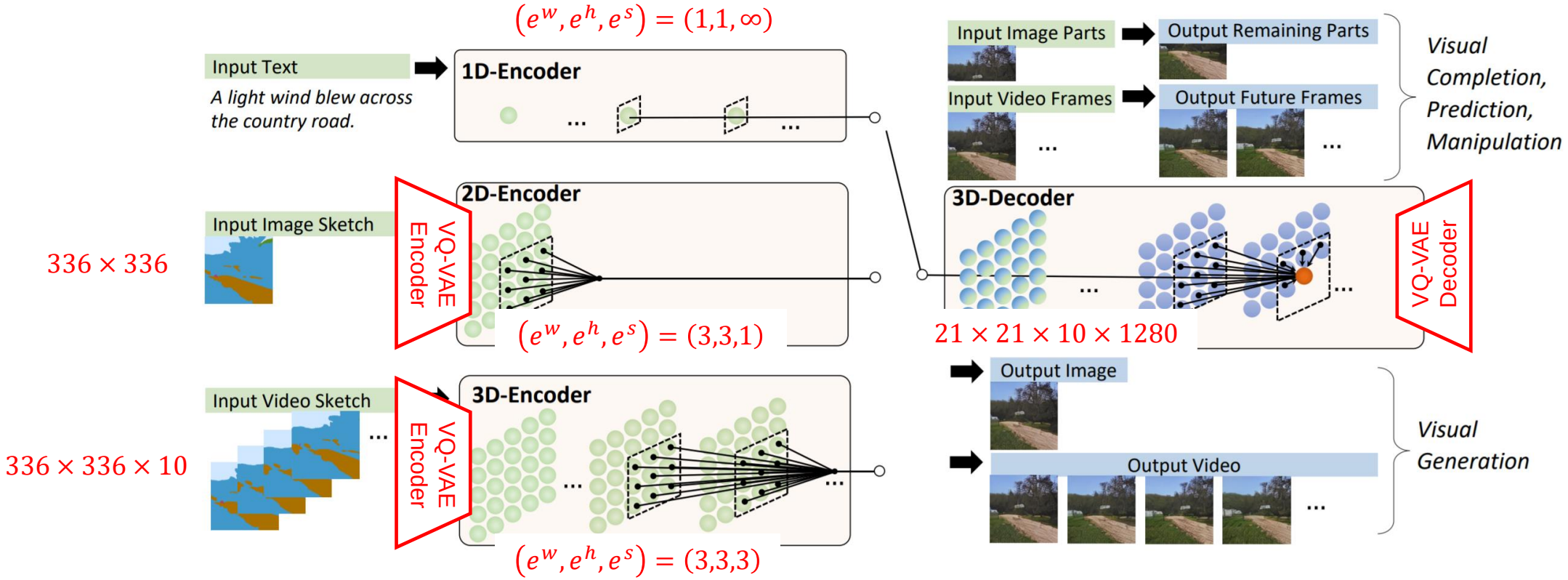


Figure 2. Overview structure of NUWA.

Experiments (T2I)

- Qualitative results

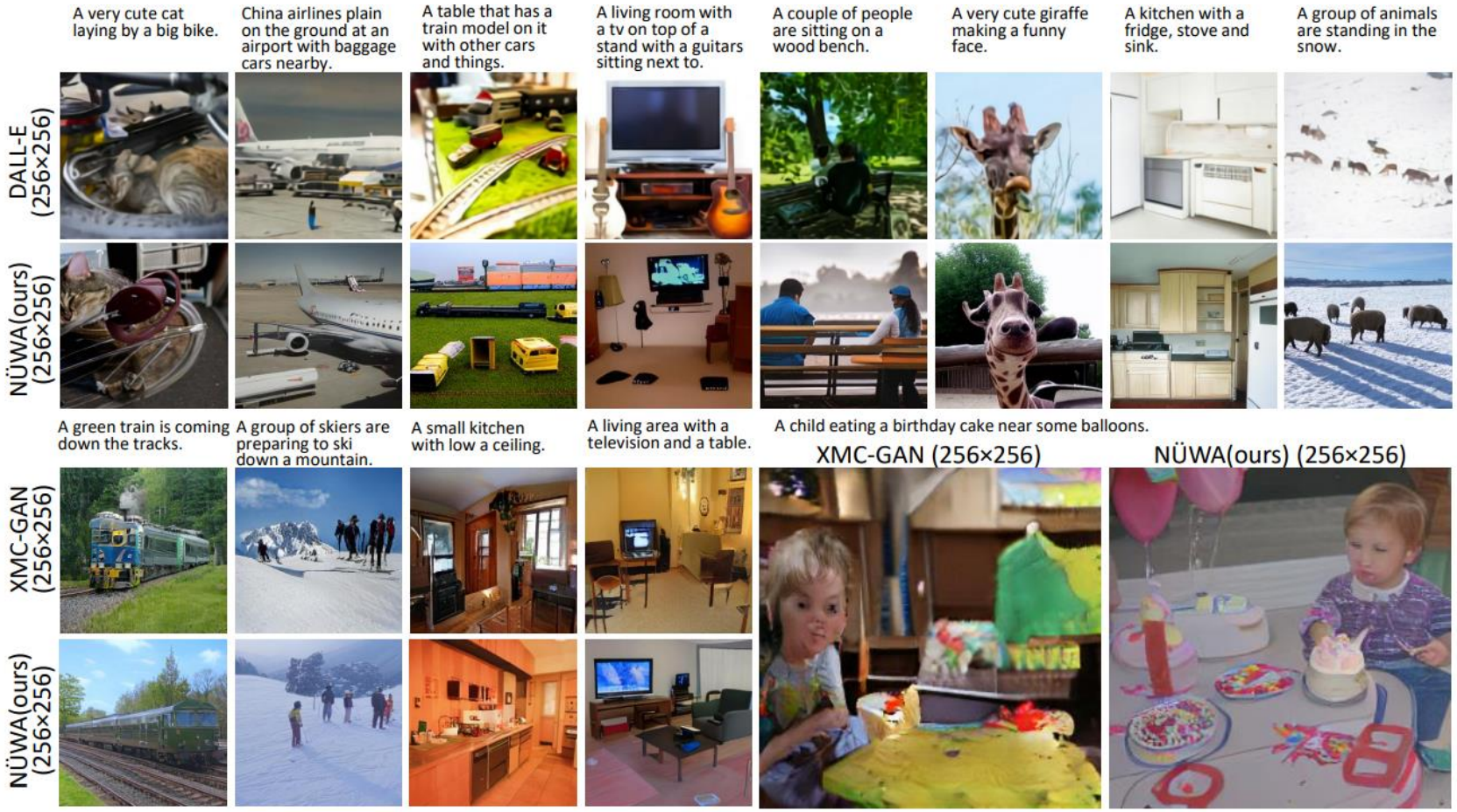


Figure 3. Qualitative comparison with state-of-the-art models for Text-to-Image (T2I) task on MSCOCO dataset.

Experiments (T2V)

- Qualitative results

Figure 4. Quantitative comparison with state-of-the-art models for Text-to-Video (T2V) task on Kinetics dataset.

Experiments (T2I, T2V)

- Quantitative results

Table 1. Qualitative comparison with the state-of-the-art models for Text-to-Image (T2I) task on the MSCOCO (256×256) dataset.

Model	FID-0↓	FID-1	FID-2	FID-4	FID-8	IS↑	CLIPSIM↑
AttnGAN [47]	35.2	44.0	72.0	108.0	100.0	23.3	0.2772
DM-GAN [52]	26.0	39.0	73.0	119.0	112.3	**32.2**	0.2838
DF-GAN [36]	26.0	33.8	55.9	91.0	97.0	18.7	0.2928
DALL-E [33]	27.5	28.0	45.5	83.5	85.0	17.9	-
CogView [9]	27.1	19.4	**13.9**	19.4	**23.6**	18.2	0.3325
XMC-GAN [50]	**9.3**	-	-	-	-	30.5	-
NÜWA	12.9	**13.8**	15.7	**19.3**	24	27.2	**0.3429**

Table 2. Quantitative comparison with state-of-the-art models for Text-to-Video (T2V) task on Kinetics dataset.

Model	Acc↑	FID-img↓	FID-vid↓	CLIPSIM↑
T2V (64×64) [21]	42.6	82.13	14.65	0.2853
SC (128×128) [2]	74.7	33.51	7.34	0.2915
TFGAN (128×128) [2]	76.2	31.76	7.19	0.2961
NÜWA (128×128)	**77.9**	**28.46**	**7.05**	**0.3012**

Experiments (S2I, I2I)

- Qualitative results

Figure 5. Quantitative comparison with state-of-the-art models for Sketch-to-Image (S2I) task on MSCOCO stuff dataset.

Figure 6. Qualitative comparison with the state-of-the-art model for Image Completion (I2I) task in a zero-shot manner.

Experiments (TI2I: Image Manipulate)

- Qualitative results

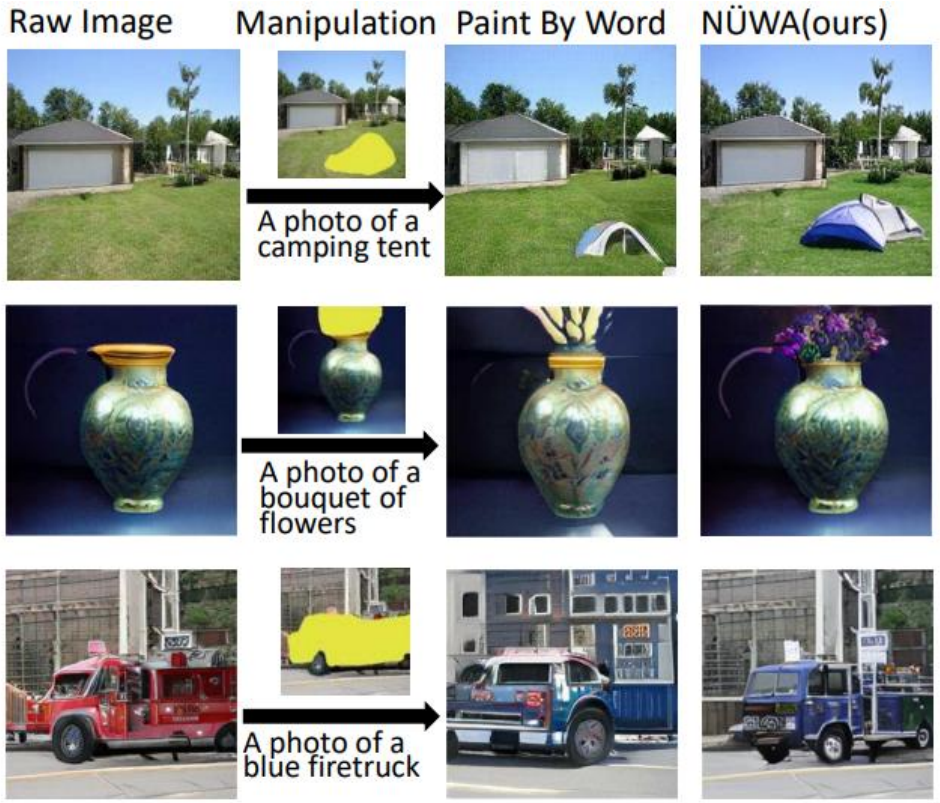


Figure 7. Quantitative comparison with state-of-the-art models for text-guided image manipulation (TI2I) in a zero-shot manner.

Experiments (TI2I: Image Manipulate)

- Qualitative results

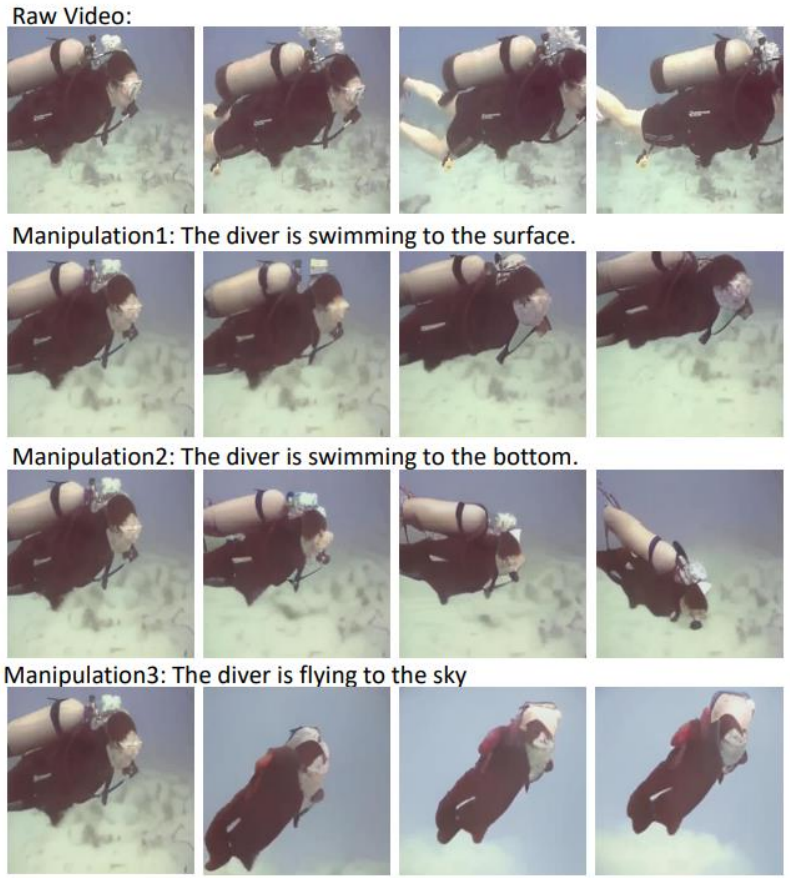


Figure 9. Samples of different manipulations on the same video.

Ablation Study

- VQ-GAN

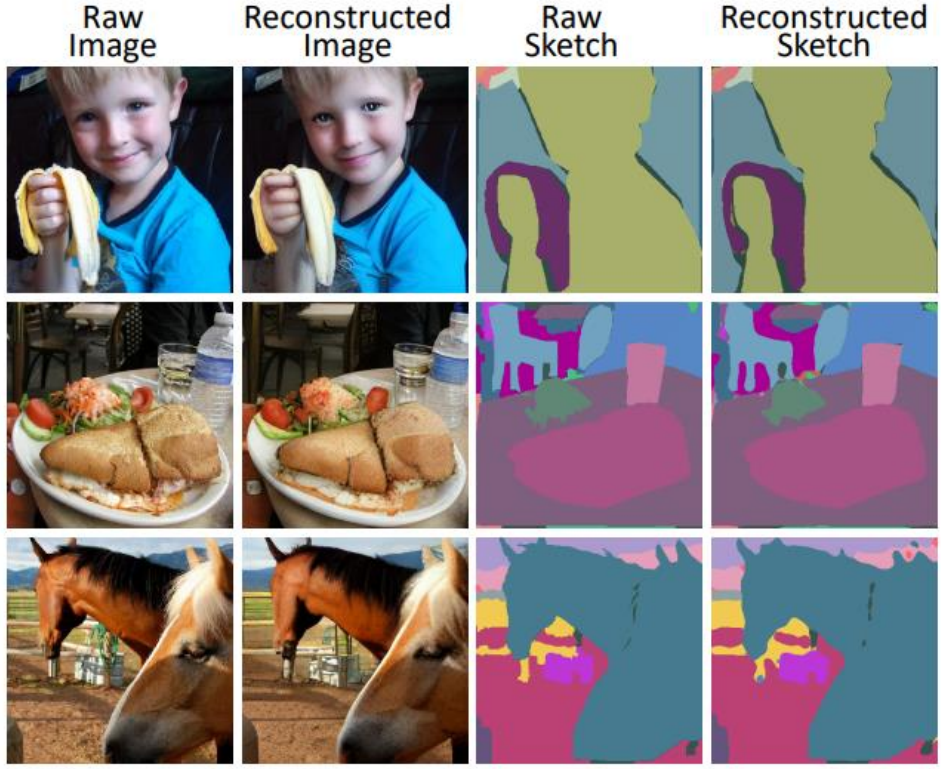


Figure 8. Reconstruction samples of VQ-GAN and VQ-GAN-Seg.

Table 4. Effectiveness of different VQ-VAE (VQ-GAN) settings.

Model	Dataset	$R \rightarrow D$	Rate	SSIM	FID
VQ-VAE	ImageNet	$256^2 \rightarrow 16^2$	F16	0.7026	13.3
VQ-GAN	ImageNet	$256^2 \rightarrow 16^2$	F16	0.7105	6.04
VQ-GAN	ImageNet	$256^2 \rightarrow 32^2$	F8	0.8285	2.03
VQ-GAN	ImageNet	$336^2 \rightarrow 21^2$	F16	0.7213	4.79
VQ-GAN	OpenImages	$336^2 \rightarrow 21^2$	F16	0.7527	4.31
Model	Dataset	$R \rightarrow D$	Rate	PA	FWIoU
VQ-GAN-Seg	MSCOCO	$336^2 \rightarrow 21^2$	F16	96.82	93.91
VQ-GAN-Seg	VSPW	$336^2 \rightarrow 21^2$	F16	95.36	91.82

Ablation Study

Effectiveness of multi-task pre-training

Table 5. Effectiveness of multi-task pre-training for Text-to-Video (T2V) generation task on MSRVTT dataset.

Model	Pre-trained Tasks	FID-vid↓	CLIPSIM↑
NÜWA-TV	T2V	52.98	0.2314
NÜWA-TV-TI	T2V+T2I	53.92	0.2379
NÜWA-TV-VV	T2V+V2V	51.81	0.2335
NÜWA	T2V+T2I+V2V	**47.68**	**0.2439**

Effectiveness of 3DNA

Table 6. Effectiveness of 3D nearby attention for Sketch-to-Video (S2V) task on VSPW dataset.

Model	Encoder	Decoder	FID-vid↓	Detected PA↑
NÜWA-FF	Full	Full	35.21	0.5220
NÜWA-NF	Nearby	Full	33.63	0.5357
NÜWA-FN	Full	Nearby	32.06	0.5438
NÜWA-AA	Axis	Axis	29.18	0.5957
NÜWA	Nearby	Nearby	**27.79**	**0.6085**

Comparisons between 3D Sparse Attention

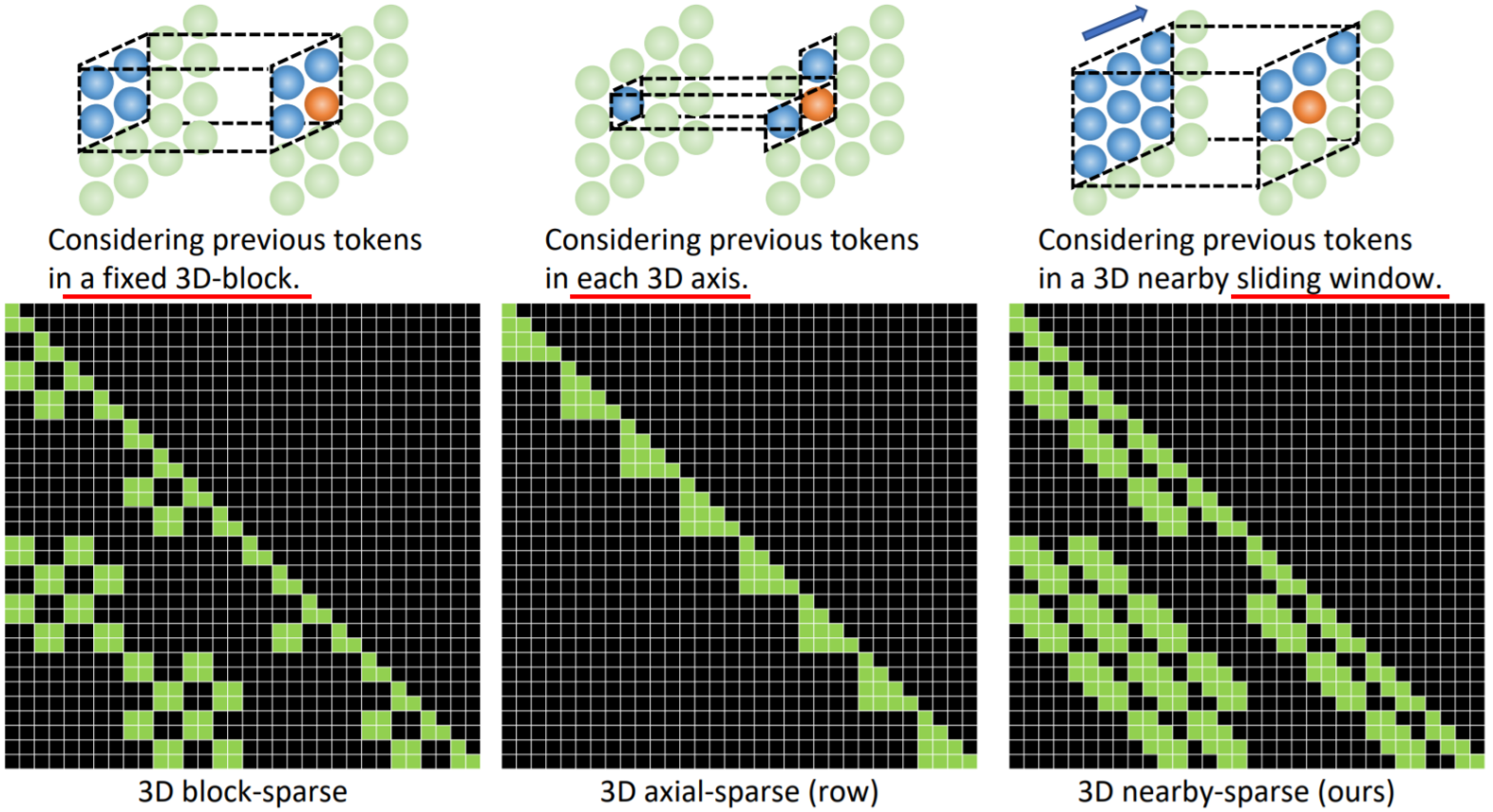


Figure 10. Comparisons between different 3D sparse attentions. All samples assume that the size of the input 3D data is $4 \times 4 \times 2 = 32$. The illustrations in the upper part show which tokens (blue) need to be attended to generate the target token (orange). The matrices of the size 32×32 in the lower part show the attention masks in sparse attention (black denotes masked tokens).

Table 7. Complexity of different 3D sparse attention.

Module	Complexity
3D full	$O\left((hws)^2\right)$
3D block-sparse [31, 44]	$O\left(\left(\frac{hws}{b}\right)^2\right)$
3D axial-sparse [15, 33, 45]	$O\left((hws)\,(h+w+s)\right)$
3D nearby-sparse (ours)	$O\left((hws)\left(e^h e^w e^s\right)\right)$

References

- Wu, Chenfei, et al. "N\" UWA: Visual Synthesis Pre-training for Neural visUal World creAtion." *arXiv preprint arXiv:2111.12417* (2021).
- Deep Foundations, Vector Quantized VAEs, https://www.youtube.com/watch?v=52G6WhBDQ3c
- Microsoft Research, Research talk: NUWA: Neural visual world creation with multimodal pretraining, https://youtu.be/jhmJ5qb-JAU