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

Chenfei Wu[1*] Jian Liang[2*] Lei Ji[1] Fan Yang[1] Yuejian Fang[2†]
Daxin Jiang[1] Nan Duan[1†]

[1] Microsoft Research Asia [2] Peking University

Abstract. This paper presents a unified multimodal pre-trained model called NÜWA that can generate new or manipulate existing visual data (i.e., images and videos) for various visual synthesis tasks. To cover language, image, and video at the same time for different scenarios, a 3D transformer encoder-decoder framework is designed, which can not only deal with videos as 3D data but also adapt to texts and images as 1D and 2D data, respectively. A 3D Nearby Attention (3DNA) mechanism is also proposed to consider the nature of the visual data and reduce the computational complexity. We evaluate NÜWA on 8 downstream tasks. Compared to several strong baselines, NÜWA achieves state-of-the-art results on text-to-image generation, text-to-video generation, video prediction, etc. Furthermore, it also shows surprisingly good zero-shot capabilities on text-guided image and video manipulation tasks.

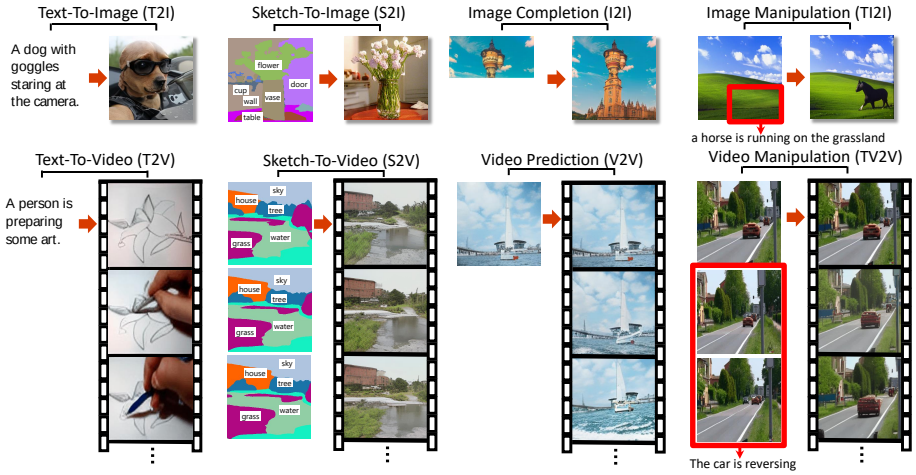


Fig. 1: Examples of 8 typical visual generation and manipulation tasks supported by the NÜWA model.

* Both authors contributed equally to this research. † Corresponding author.

1 Introduction

Nowadays, the Web is becoming more visual than ever before, as images and videos have become the new information carriers and have been used in many practical applications. With this background, visual synthesis is becoming a popular research topic, which aims to generate new or manipulate existing visual data (i.e., images and videos) for various visual scenarios.

Auto-regressive models[35, 21, 39, 29] play an important role in visual synthesis tasks, due to their explicit density modeling and stable training compared with GANs[3, 33, 41, 26]. Earlier visual auto-regressive models like PixelCNN[21], PixelRNN[35], Image Transformer[24], iGPT[4], and Video Transformer[38], performed visual synthesis in a "pixel-by-pixel" manner. However, due to their high computational cost on high-dimensional visual data, such methods can be applied to low-resolution images or videos only and are hard to scale up.

Recently, with the arise of VQ-VAE[22] as a discrete visual tokenization approach, efficient and large-scale pre-training can be applied to visual synthesis tasks for images (e.g., DALL-E[29] and CogView[7]) and videos (e.g., GODIVA[39]). To model the locality of images and videos, sparse attentions are commonly used to reduce computation and improve the performance. Although achieving great success, such solutions still have the following two limitations:

On the one hand, from the pre-training perspective, current works treat images and videos separately and focus on generating either of them. This limits the models to benefit from both image and video data.

On the other hand, from the model perspective, current works use block-sparse attention or axial-sparse attention as the pre-training backbone, both considering only part of visual locality. Block-sparse attention limits attention in a fixed 3D block and axial-sparse attentions limits attention in axes, both failed to fully model the locality of images and videos.

To handle the above issues, we propose NÜWA, with a 3D decoder to share information from both images and videos and a 3D Nearby-sparse Attention (3DNA) to model the full spatial and temporal locality. We verify NÜWA on 8 downstream visual synthesis, as shown in Fig. 1. The main contributions of this work are three-fold:

- We propose NÜWA, a general 3D transformer encoder-decoder framework, which covers language, image, and video at the same time for different visual synthesis tasks. It consists of an adaptive encoder that takes either text or visual sketch as input, and a decoder shared by 8 visual synthesis tasks.
- We propose a 3D Nearby Attention (3DNA) mechanism in the framework to consider the locality characteristic for both spatial and temporal axes. 3DNA not only reduces computational complexity but also improves the visual quality of the generated results.
- Compared to several strong baselines, NÜWA achieves state-of-the-art results on text-to-image generation, text-to-video generation, video prediction, etc. Furthermore, NÜWA shows surprisingly good zero-shot capabilities not only on text-guided image manipulation, but also text-guided video manipulation.

2 Related Works

2.1 Visual Auto-Regressive Models

The method proposed in this paper follows the line of visual synthesis research based on auto-regressive models. Earlier visual auto-regressive models [21, 35, 24, 4, 38] performed visual synthesis in a "pixel-by-pixel" manner. However, due to the high computational cost when modeling high-dimensional data, such methods can be applied to low-resolution images or videos only, and are hard to scale up.

Recently, VQ-VAE-based [22] visual auto-regressive models were proposed for visual synthesis tasks. By converting images into discrete visual tokens, such methods can conduct efficient and large-scale pre-training for text-to-image generation (e.g., DALL-E[29] and CogView[7]), text-to-video generation (e.g., GODIVA[39]), and video prediction (e.g., LVT[27] and VideoGPT[42]), with higher resolution of generated images or videos. However, none of these models was trained by images and videos together. But it is intuitive that these tasks can benefit from both types of visual data.

Compared to these works, NÜWA is a unified auto-regressive visual synthesis model that is pre-trained by the visual data covering both images and videos and can support various downstream tasks. We also verify the effectiveness of different pretraining tasks in Sec. 4.3. Besides, VQ-GAN[9] instead of VQ-VAE is used in NÜWA for visual tokenization, which, based on our experiment, can lead to better generation quality.

2.2 Visual Sparse Self-Attention

How to deal with the quadratic complexity issue brought by self-attention is another challenge, especially for tasks like high-resolution image synthesis or video synthesis.

Similar to NLP, sparse attention mechanisms have been explored to alleviate this issue for visual synthesis. [38, 27] split the visual data into different parts (or blocks) and then performed block-wise sparse attention for the synthesis tasks. However, such methods dealt with different blocks separately and did not model their relationships. [11, 29, 39] proposed to use axial-wise sparse attention in visual synthesis tasks, which conducts sparse attention along the axes of visual data representations. This mechanism makes training very efficient and is friendly to large-scale pre-trained models like DALL-E[29], CogView[7], and GODIVA[39]. However, the quality of generated visual contents could be harmed due to the limited contexts used in self-attention. [24, 28, 5] proposed to use local-wise sparse attention in visual synthesis tasks, which allows the models to see more contexts. But these works were for images only.

Compared to these works, NÜWA proposes a 3D nearby attention that extends the local-wise sparse attention to cover both images to videos. We also verify that local-wise sparse attention is superior to axial-wise sparse attention for visual generation in Sec. 4.3.

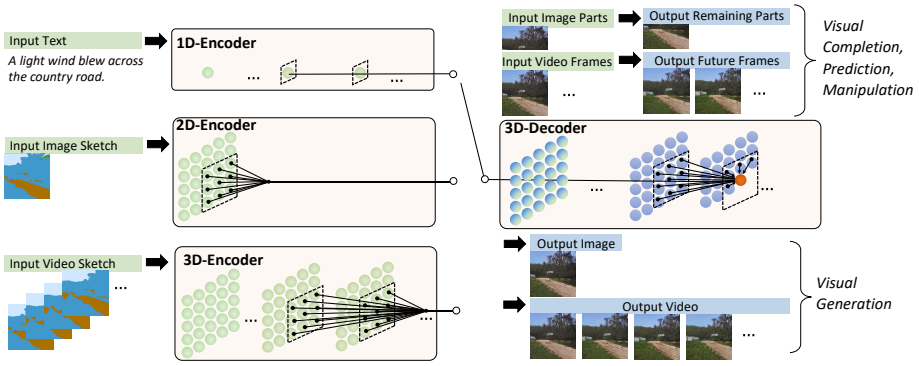


Fig. 2: Overview structure of NÜWA. It contains an adaptive encoder supporting different conditions and a pre-trained decoder benefiting from both image and video data. For image completion, video prediction, image manipulation, and video manipulation tasks, the input partial images or videos are fed to the decoder directly.

3 Method

3.1 3D Data Representation

To cover all texts, images, and videos or their sketches, we view all of them as tokens and define a unified 3D notation $X \in \mathbb{R}^{h \times w \times s \times d}$, where h and w denote the number of tokens in the spatial axis (height and width respectively), s denotes the number of tokens in the temporal axis, and d is the dimension of each token. In the following, we introduce how we get this unified representation for different modalities.

Texts are naturally discrete, and following Transformer[36], we use a lower-cased byte pair encoding (BPE) to tokenize and embed them into $\mathbb{R}^{1 \times 1 \times s \times d}$. We use placeholder 1 because the text has no spatial dimension.

Images are naturally continuous pixels. Input a raw image $I \in \mathbb{R}^{H \times W \times C}$ with height H, width W and channel C, VQ-VAE[22] trains a learnable codebook to build a bridge between raw continuous pixels and discrete tokens, as denoted in Eq. (1)$\sim$(2):

$$z_i = \arg\min_j ||E(I)_i - B_j||^2, \tag{1}$$

$$\hat{I} = G(B[z]), \tag{2}$$

where E is an encoder that encodes I into $h \times w$ grid features $E(I) \in \mathbb{R}^{h \times w \times d_B}$, $B \in \mathbb{R}^{N \times d_B}$ is a learnable codebook with N visual tokens, where each grid of $E(I)$ is searched to find the nearest token. The searched result $z \in \{0, 1, \ldots, N-1\}^{h \times w}$ are embedded by B and reconstructed back to $\hat{I}$ by a decoder G. The training loss of VQ-VAE can be written as Eq. (3):

$$L^V = ||I - \hat{I}||_2^2 + ||sg[E(I)] - B[z]||_2^2 + ||E(I) - sg[B[z]]||_2^2, \tag{3}$$

where $||I-\hat{I}||_2^2$ strictly constraints the exact pixel match between I and $\hat{I}$, which limits the generalization ability of the model. Recently, VQ-GAN[9] enhanced VQ-VAE training by adding a perceptual loss and a GAN loss to ease the exact constraints between I and $\hat{I}$ and focus on high-level semantic matching, as denoted in Eq. (4)∼(5):

$$L^P = ||CNN(I) - CNN(\hat{I})||_2^2, \tag{4}$$

$$L^G = logD(I) + log(1 - D(\hat{I})). \tag{5}$$

After the training of VQ-GAN, $B[z] \in \mathbb{R}^{h \times w \times 1 \times d}$ is finally used as the representation of images. We use placeholder 1 since images have no temporal dimensions.

Videos can be viewed as a temporal extension of images, and recent works like VideoGPT[42] and VideoGen[45] extend convolutions in the VQ-VAE encoder from 2D to 3D and train a video-specific representation. However, this fails to share a common codebook for both images and videos. In this paper, we show that simply using 2D VQ-GAN to encode each frame of a video can also generate temporal consistency videos and at the same time benefit from both image and video data. The resulting representation is denoted as $\mathbb{R}^{h \times w \times s \times d}$, where s denotes the number of frames.

For image sketches, we consider them as images with special channels. An image segmentation matrix $\mathbb{R}^{H \times W}$ with each value representing the class of a pixel can be viewed in a one-hot manner $\mathbb{R}^{H \times W \times C}$ where C is the number of segmentation classes. By training an additional VQ-GAN for image sketch, we finally get the embedded image representation $\mathbb{R}^{h \times w \times 1 \times d}$. Similarly, for video sketches, the representation is $R^{h \times w \times s \times d}$.

3.2 3D Nearby Self-Attention

In this section, we define a unified 3D Nearby Self-Attention (3DNA) module based on the previous 3D data representations, supporting both self-attention and cross-attention. We first give the definition of 3DNA in Eq. (6), and introduce detailed implementation in Eq. (7)∼(11):

$$Y = 3DNA(X, C; W), \tag{6}$$

where both $X \in \mathbb{R}^{h \times w \times s \times d^{in}}$ and $C \in \mathbb{R}^{h' \times w' \times s' \times d^{in}}$ are 3D representations introduced in Sec. 3.1. If $C = X$, 3DNA denotes the self-attention on target X and if $C \neq X$, 3DNA is cross-attention on target X conditioned on C. W denotes learnable weights.

We start to introduce 3DNA from a coordinate (i, j, k) under X. By a linear projection, the corresponding coordinate (i', j', k') under C is $\left(\lfloor i\frac{h'}{h} \rfloor, \lfloor j\frac{w'}{w} \rfloor, \lfloor k\frac{s'}{s} \rfloor\right)$. Then, the local neighborhood around (i', j', k') with a width, height and temporal extent $e^w, e^h, e^s \in \mathbb{R}^+$ is defined in Eq. (7),

$$N^{(i,j,k)} = \left\{ C_{abc} \middle| \, |a - i'| \leq e^h, |b - j'| \leq e^w, |c - k'| \leq e^s \right\}, \tag{7}$$

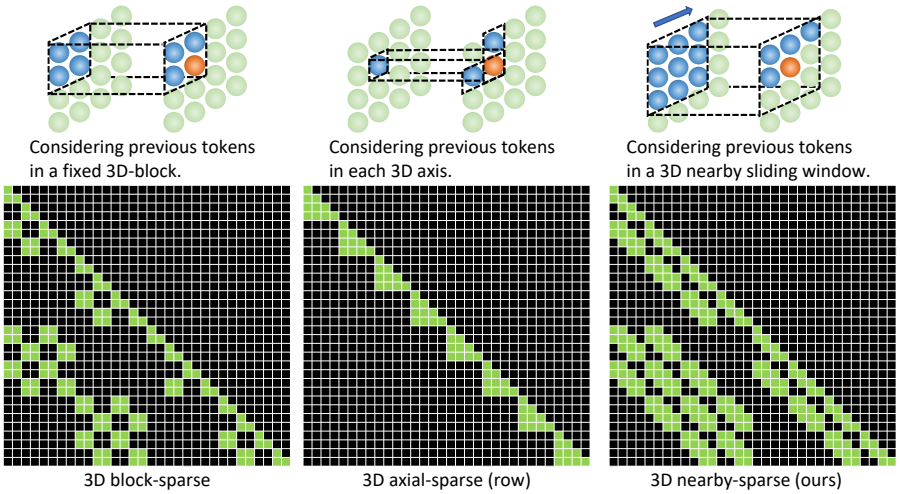


Fig. 3: Comparisons between different 3D sparse attentions. All samples assume that the size of the input 3D data is $4 \times 4 \times 2 = 32$. The illustrations in the upper part show which tokens (blue) need to be attended to generate the target token (orange). The matrices of the size 32×32 in the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

where $N^{(i,j,k)} \in \mathbb{R}^{e^h \times e^w \times e^s \times d^{in}}$ is a sub-tensor of condition C and consists of the corresponding nearby information that (i, j, k) needs to attend. With three learnable weights $W_Q, W_K, W_V \in \mathbb{R}^{d^{in} \times d^{out}}$, the output tensor for the position (i, j, k) is denoted in Eq. (8)∼(11):

$$Q^{(i,j,k)} = XW^Q \tag{8}$$

$$K^{(i,j,k)} = N^{(i,j,k)}W^K \tag{9}$$

$$V^{(i,j,k)} = N^{(i,j,k)}W^V \tag{10}$$

$$y_{ijk} = softmax\left(\frac{(Q^{(i,j,k)})^\intercal K^{(i,j,k)}}{\sqrt{d^{in}}}\right)V^{(i,j,k)} \tag{11}$$

where the (i, j, k) position queries and collects corresponding nearby information in C. This also handles $C = X$, then (i, j, k) queries the nearby position of itself.

Fig. 3 shows comparisons between different 3D sparse attentions. Assume we have 3D data with the size of $4 \times 4 \times 2$, the idea of 3D block-sparse attention is to split the 3D data into several fixed blocks and handle these blocks separately. There are many ways to split blocks, such as splitting in time, space, or both. The 3D block-sparse example in Fig. 3 considers the split of both time and space. The 3D data is divided into 4 parts, each has the size of $2 \times 2 \times 2$. To generate the orange token, 3D block-sparse attention considers previous tokens inside the fixed 3D block. Although 3D block-sparse attention considers both spatial and temporal axes, this spatial and temporal information is limited and fixed in the

3D block especially for the tokens along the edge of the 3D block. Only part of nearby information is considered since some nearby information outside the 3D block is invisible for tokens inside it. The idea of 3D axial-sparse attention is to consider previous tokens along the axis. Although 3D axis-sparse attention considers both spatial and temporal axes, this spatial and temporal information is limited along the axes. Only part of nearby information is considered and some nearby information that does not in the axis will not be considered in the 3D axis attention. In this paper, we propose a 3D nearby-sparse, which considers the full nearby information and dynamically generates the 3D nearby attention block for each token. The attention matrix also shows the evidence as the attended part (blue) for 3D nearby-sparse is more smooth than 3D block-sparse and 3D axial-sparse.

3.3 3D Encoder-Decoder

In this section, we introduce 3D encode-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 $C \in \mathbb{R}^{h' \times w' \times s' \times d^{in}}$, the positional encoding for both Y and C are updated by three different learnable vocabularies considering height, width, and temporal axis, respectively in Eq. (12)∼(13):

$$Y_{ijk} := Y_{ijk} + P_i^h + P_j^w + P_k^s \tag{12}$$

$$C_{ijk} := C_{ijk} + P_i^{h'} + P_j^{w'} + P_k^{s'} \tag{13}$$

Then, the condition C is fed into an encoder with a stack of L 3DNA layers to model the self-attention interactions, with the lth layer denoted in Eq. (14):

$$C^{(l)} = 3DNA(C^{(l-1)}, C^{(l-1)}), \tag{14}$$

Similarly, the decoder is also a stack of L 3DNA layers. The decoder calculates both self-attention of generated results and cross-attention between generated results and conditions. The lth layer is denoted in Eq. (15).

$$Y_{ijk}^{(l)} = 3DNA(Y_{<i,<j,<k}^{(l-1)}, Y_{<i,<j,<k}^{(l-1)}) + 3DNA(Y_{<i,<j,<k}^{(l-1)}, C^{(L)}), \tag{15}$$

where $< i, < j, < k$ denote the generated tokens for now. The initial token $V_{0,0,0}^{(1)}$ is a special $< bos >$ token learned during the training phase.

3.4 Training Objective

We train our model on three tasks, Text-to-Image (T2I), Video Prediction (V2V) and Text-to-Video (T2V). The training objective for the three tasks are cross-entropys denoted as three parts in Eq. (16), respectively:

$$\mathcal{L} = -\sum_{t=1}^{h \times w} log\ p_\theta \left(y_t | y_{<t}, C^{text}; \theta\right) - \sum_{t=1}^{h \times w \times s} log\ p_\theta \left(y_t | y_{<t}, c; \theta\right) - \sum_{t=1}^{h \times w \times s} log\ p_\theta \left(y_t | y_{<t}, C^{text}; \theta\right) \tag{16}$$

For T2I and T2V tasks, C^{text} denotes text conditions. For the V2V task, since there is no text input, we instead get a constant 3D representation c of the special word "None". θ denotes the model parameters.

4 Experiments

Based on Sec. 3.4 we first pre-train NÜWA on three datasets: Conceptual Captions[16] for text-to-image (T2I) generation, which includes 2.9M text-image pairs, Moments in Time[20] for video prediction (V2V), which includes 727K videos, and VATEX dataset[37] for text-to-video (T2V) generation, which includes 241K text-video pairs. In the following, we first introduce implementation details in Sec. 4.1 and then compare NÜWA with state-of-the-art models in Sec. 4.2, and finally conduct ablation studies in Sec. 4.3 to study the impacts of different parts.

4.1 Implementation Details

In Sec. 3.1, we set the sizes of 3D representations for text, image, and video as follows. For text, the size of 3D representation is $1 \times 1 \times 77 \times 1280$. For image, the size of 3D representation is $21 \times 21 \times 1 \times 1280$. For video, the size of 3D representation is $21 \times 21 \times 10 \times 1280$, where we sample 10 frames from a video with 2.5 fps. Although the default visual resolution is 336×336, we pre-train different resolutions for a fair comparison with existing models. For the VQ-GAN model used for both images and videos, the size of grid feature $E(I)$ in Eq. (1) is 441×256, and the size of the codebook B is $12,288$.

Different sparse extents are used for different modalities in Sec. 3.2. For text, we set $(e^w, e^h, e^s) = (1, 1, \infty)$, where ∞ denotes that the full text is always used in attention. For image and image sketches, $(e^w, e^h, e^s) = (3, 3, 1)$. For video and video sketches, $(e^w, e^h, e^s) = (3, 3, 3)$.

We pre-train on 64 A100 GPUs for two weeks with the layer L in Eq. (14) set to 24, an Adam [13] optimizer with a learning rate of 1e-3, a batch size of 128, and warm-up 5% of a total of 50M steps. The final pre-trained model has a total number of 870M parameters.

4.2 Comparison with state-of-the-art

Text-to-Image (T2I) fine-tuning: We compare NÜWA on the MSCOCO[16] dataset quantitatively in Tab. 1 and qualitatively in Fig. 4. Following DALL-E[29], we use k blurred FID score (FID-k) and Inception Score (IS)[31] to evaluate the quality and variety respectively, and following GODIVA[39], we use CLIPSIM metric, which incorporates a CLIP[25] model to calculate the semantic similarity between input text and the generated image. For a fair comparison, all the models use the resolution of 256×256. We generate 60 images for each text and select the best one by CLIP[25]. In Tab. 1, NÜWA significantly outperforms CogView[7] with FID-0 of 12.9 and CLIPSIM of 0.3429. Although XMC-GAN[44] reports a significant FID score of 9.3, we find NÜWA generates more realistic images compared with the exact same samples in XMC-GAN's paper (see Fig. 4). Especially in the last example, the boy's face is clear and the balloons are correctly generated.

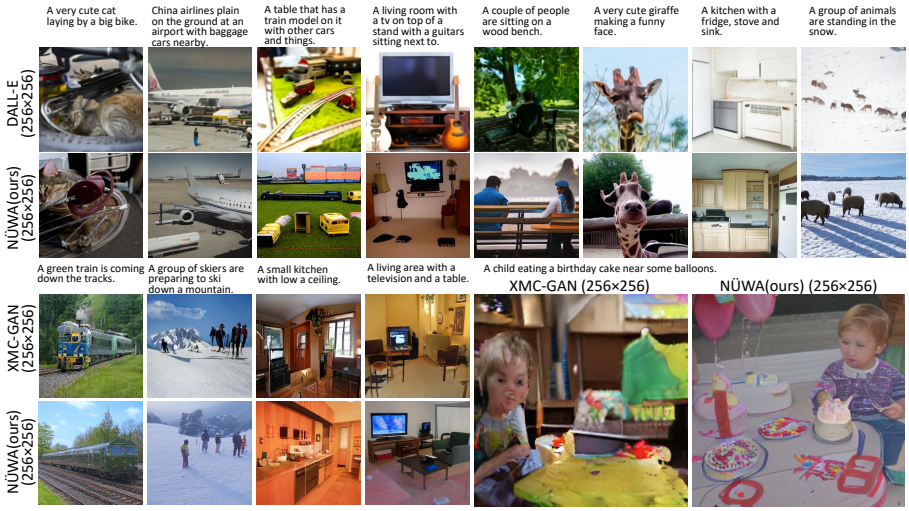


Fig. 4: Qualitative comparison with state-of-the-art models for Text-to-Image (T2I) task on MSCOCO dataset.

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[41]	35.2	44.0	72.0	108.0	100.0	23.3	0.2772
DM-GAN[46]	26.0	39.0	73.0	119.0	112.3	**32.2**	0.2838
DF-GAN[32]	26.0	33.8	55.9	91.0	97.0	18.7	0.2928
DALL-E[29]	27.5	28.0	45.5	83.5	85.0	17.9	-
CogView[7]	27.1	19.4	**13.9**	19.4	23.6	18.2	0.3325
XMC-GAN[44]	**9.3**	-	-	-	-	30.5	-
NÜWA(scratch)	*Full Attention*						
	17.1	16.5	16.3	18.5	20.9	22.7	0.3257
NÜWA(scratch)	*Axial Attention*						
	18.7	18.4	19.2	20.3	21.3	22.8	0.3253
NÜWA(scratch)	*3D Nearby Attention (ours)*						
	16.9	15.6	16.5	18.9	**20.2**	23.1	0.3276
NÜWA(finetune)	*Pretrain on CC Dataset*						
	14.2	15.2	16.9	20.5	24.7	25.8	0.3424
NÜWA(finetune)	*Pretrain on CC, Moments and Vatex Dataset*						
	12.9	**13.8**	15.7	19.3	24	27.2	**0.3429**
NÜWA(zeroshot)	*Pretrain on CC, Moments and Vatex Dataset*						
	22.6	18.6	17.2	**17.4**	24.8	24.5	0.3331

To validate the effectiveness of our proposed 3DNA, we train NÜWA from scratch and Tab. 1 shows 3DNA outperforms axial attentions and full attentions. To validate the effectiveness of joint pretraining both images and videos, we pretrain NÜWA on pure image dataset (CC) and mixed image and video dataset (CC, Moments, Vatex). Tab. 1 shows Text-to-Video task also helps Text-to-Image task. This is interesting as videos provides external motion knowledge to help the model better build the connection between text and image.

Fig. 5: Quantitative comparison with state-of-the-art models for Text-to-Video (T2V) task on Kinetics dataset.

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) [15]	42.6	82.13	14.65	0.2853
SC (128×128) [1]	74.7	33.51	7.34	0.2915
TFGAN (128×128)[1]	76.2	31.76	7.19	0.2961
NÜWA(scratch)	77.4	29.32	7.08	0.3007
NÜWA(finetune)	**77.9**	**28.46**	**7.05**	**0.3012**

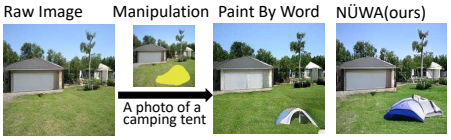
Table 3: Quantitative comparison with state-of-the-art models for Video Prediction (V2V) task on BAIR (64×64) dataset.

Model	Cond.	FVD↓
DVD-GAN-FP[6]	1	110
Video Transformer (S)[38]	1	106±3
TriVD-GAN-FP[17]	1	103
CCVS[19]	1	99±2
Video Transformer (L)[38]	1	94±2
NÜWA(scratch)	1	87.6
NÜWA(finetune)	1	**86.9**

Text-to-Video (T2V) fine-tuning: We compare NÜWA on the Kinetics[12] dataset quantitatively in Tab. 2 and qualitatively in Fig. 5. Following TFGAN[1], we evaluate the visual quality on FID-img and FID-vid metrics and semantic consistency on the accuracy of the label of generated video.To ensure NÜWA is trained with the same size information of Kinetics as other methods, images are first bilinear interpolated into 128×128 before resized into 336×336. As shown in Tab. 2, NÜWA achieves the best performance. In Fig. 5, we also

Fig. 6: Quantitative comparison with state-of-the-art models for Sketch-to-Image (S2I) task on MSCOCO-Stuff.

Fig. 7: Qualitative comparison with the state-of-the-art model for Image Completion (I2I) task in a zero-shot manner.



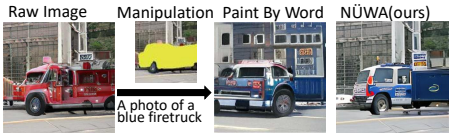


Fig. 8: Quantitative comparison with state-of-the-art models for text-guided image manipulation (TI2I) in a zero-shot manner.

show the strong zero-shot ability for generating unseen text, such as "playing golf at swimming pool" or "running on the sea".

Video Prediction (V2V) fine-tuning: We compare NÜWA on BAIR Robot Pushing[8] dataset quantitatively in Tab. 3. Cond. denotes the number of frames given to predict future frames. For a fair comparison, all the models use 64×64 resolutions. Although given only one frame as condition (Cond.), NÜWA still significantly pushes the state-of-the-art FVD[34] score from 94±2 to 86.9.

Sketch-to-Image (S2I) fine-tuning: We compare NÜWA on MSCOCO stuff[16] qualitatively in Fig. 6. NÜWA generates realistic buses of great varieties compared with Taming-Transformers[9] and SPADE[23]. Even the reflection of the bus window is clearly visible.

Image Completion (I2I) zero-shot evaluation: We compare NÜWA in a zero-shot manner qualitatively in Fig. 7. Given the top half of the tower, compared with Taming Transformers[9], NÜWA shows richer imagination of what could be for the lower half of the tower, including buildings, lakes, flowers, grass, trees, mountains, etc.

Text-Guided Image Manipulation (TI2I) zero-shot evaluation: We compare NÜWA in a zero-shot manner qualitatively in Fig. 8. Compared with Paint By Word[2], NÜWA shows strong manipulation ability, generating high-quality text-consistent results while not changing other parts of the image. For example, in the third row, the blue firetruck generated by NÜWA is more realistic, while the behind buildings show no change. This is benefited from real-

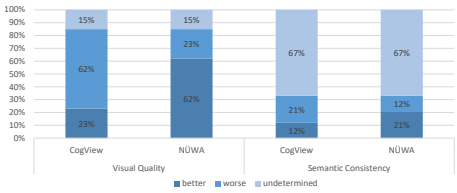


Fig. 9: Human evaluation on MSCOCO dataset for Text-to-Image (T2I) task.

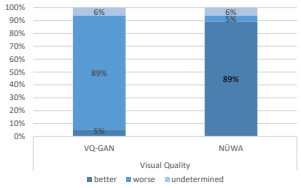


Fig. 10: Human evaluation on MSCOCO dataset for Image Completion (I2I) task.

world visual patterns learned by multi-task pre-training on various visual tasks. Another advantage is the inference speed of NÜWA, practically 50 seconds to generate an image, while Paint By Words requires additional training during inference, and takes about 300 seconds to converge.

Sketch-to-Video (S2V) fine-tuning and **Text-Guided Video Manipulation (TV2V) zero-shot evaluation**: Since there are no current benchmarks for these two tasks,we thus arrange them in Ablation Study in Section 4.3.

Human Evaluation Fig. 9 presents human comparison results between CogView[7] and our NÜWA on the MSCOCO dataset for Text-to-Image (T2I) task. We randomly selected 2000 texts and ask annotators to compare the generated results between two models including both visual quality and semantic consistency. The annotators are asked to choose among three options: better, worse, or undetermined. NÜWA achieves 62% votes for visual quality and 21% votes for semantic consistency. Fig. 10 shows another human comparison between VQ-GAN[9] and our NÜWA model on the MSCOCO dataset for the Image Completion (I2I) task.

4.3 Ablation Study

The above part of Tab. 4 shows the effectiveness of different VQ-VAE (VQ-GAN) settings. We experiment on ImageNet[30] and OpenImages[14]. R denotes raw resolution, D denotes the number of discrete tokens. The compression rate is denoted as Fx, where x is the quotient of $\sqrt{R}$ divided by $\sqrt{D}$. Comparing the first two rows in Tab. 4, VQ-GAN shows significantly better Fréchet Inception Distance (FID)[10] and Structural Similarity Matrix (SSIM) scores than VQ-VAE. Comparing Row 2-3, we find that the number of discrete tokens is the key factor leading to higher visual quality instead of compress rate. Although Row 2 and Row 4 have the same compression rate F16, they have different FID scores of 6.04 and 4.79. So what matters is not only how much we compress the original image, but also how many discrete tokens are used for representing an image. This is in line with cognitive logic, it's too ambiguous to represent human faces with just one token. And practically, we find that 16^2 discrete tokens usually lead to poor performance, especially for human faces, and 32^2 tokens show the best performance. However, more discrete tokens mean more computing, especially for videos. We finally use a trade-off version for our pre-training: 21^2 tokens. By training on the Open Images dataset, we further improve the FID score of the 21^2 version from 4.79 to 4.31.

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

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

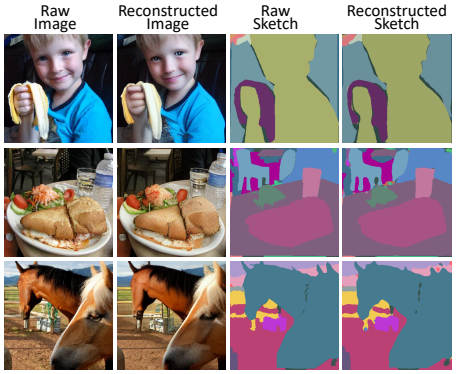


Fig. 11: Reconstruction samples of VQ-GAN and VQ-GAN-Seg.

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**

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

Model	Encoder	Decoder	FID-vid↓	DetectedPA↑
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**

The below part of Tab. 4 shows the performance of VQ-GAN for sketches. VQ-GAN-Seg on MSCOCO[16] is trained for Sketch-to-Image (S2I) task and VQ-GAN-Seg on VSPW[18] is trained for Sketch-to-Video (S2V) task. All the above backbone shows good performance in Pixel Accuracy (PA) and Frequency Weighted Intersection over Union (FWIoU), which shows a good quality of 3D sketch representation used in our model. Fig. 11 also shows some reconstructed samples of 336×336 images and sketches.

Tab. 5 shows the effectiveness of multi-task pre-training for the Text-to-Video (T2V) generation task. We study on a challenging dataset, MSR-VTT[40], with natural descriptions and real-world videos. Compared with training only on a single T2V task (Row 1), training on both T2V and T2I (Row 2) improves the CLIPSIM from 0.2314 to 0.2379. This is because T2I helps to build a connection between text and image, and thus helpful for the semantic consistency of the T2V task. In contrast, training on both T2V and V2V (Row 3) improves the FVD score from 52.98 to 51.81. This is because V2V helps to learn a common unconditional video pattern, and is thus helpful for the visual quality of the T2V task. The default setting, training on three tasks, achieves the best performance.

Tab. 6 shows the effectiveness of 3D nearby attention for the Sketch-to-Video (S2V) task on the VSPW[18] dataset. We study on the S2V task because both the encoder and decoder of this task are fed with 3D video data. To evaluate the semantic consistency for S2V, we propose a new metric called Detected PA, which uses a semantic segmentation model[43] to segment each frame of the generated video and then calculate the pixel accuracy between the generated segments and

Fig. 12: Samples of different manipulations on the same video.

input video sketch. The default NÜWA setting with both nearby encoder and nearby decoder, achieves the best FID-vid and Detected PA. The performance drops if either encoder or decoder is replaced by full attention, showing that focusing on nearby conditions and nearby generated results is better than simply considering all the information.

We compare nearby-sparse and axial-sparse in two-folds. Firstly, the computational complexity of nearby-sparse is $O\left((hws)\left(e^h e^w e^s\right)\right)$ and axis-sparse attention is $O\left((hws)\left(h+w+s\right)\right)$. For generating long videos (larger s), nearby-sparse will be more computational efficient. Secondly, nearby-sparse has better performance because it attends to "nearby" locations containing interactions between both spatial and temporal axes, while axis-sparse handles different axis separately and only consider interactions on the same axis.

Fig. 12 shows a new task "Text-Guided Video Manipulation (TV2V)" proposed in this paper. TV2V aims to change the future of a video starting from a selected frame guided by text. All samples start to change the future of the video from the second frame. The first row shows the original video frames, where a diver is swimming in the water. After feeding "The diver is swimming to the surface" into NÜWA's encoder and providing the first video frame, NÜWA successfully generates a video with the diver swimming to the surface in the second row. The third row shows another successful sample that lets the diver swim to the bottom. What if we want the diver flying to the sky? The fourth row shows that NÜWA can make it as well, where the diver is flying upward, like a rocket.

5 Conclusion

In this paper, we present NÜWA as a unified pre-trained model that can generate new or manipulate existing images and videos for 8 visual synthesis tasks. Several contributions are made here, including (1) a general 3D encoder-decoder framework covering texts, images, and videos; (2) a nearby-sparse attention mechanism that considers the nearby characteristic of both spatial and temporal axes; (3) comprehensive experiments on 8 synthesis tasks. This is our first step towards building an AI platform to enable visual world and help creators.

Acknowledgements The paper is supported by the National Key Research and Development Project (Grant No.2020AAA0106600).

References

1. Balaji, Y., Min, M.R., Bai, B., Chellappa, R., Graf, H.P.: Conditional GAN with Discriminative Filter Generation for Text-to-Video Synthesis. In: IJCAI. pp. 1995–2001 (2019)
2. Bau, D., Andonian, A., Cui, A., Park, Y., Jahanian, A., Oliva, A., Torralba, A.: Paint by word. arXiv preprint arXiv:2103.10951 (2021)
3. Brock, A., Donahue, J., Simonyan, K.: Large Scale GAN Training for High Fidelity Natural Image Synthesis. arXiv:1809.11096 [cs, stat] (Feb 2019)
4. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International Conference on Machine Learning. pp. 1691–1703. PMLR (2020)
5. Child, R., Gray, S., Radford, A., Sutskever, I.: Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509 (2019)
6. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What does bert look at? an analysis of bert's attention. arXiv preprint arXiv:1906.04341 (2019)
7. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., Tang, J.: CogView: Mastering Text-to-Image Generation via Transformers. arXiv:2105.13290 [cs] (May 2021)
8. Ebert, F., Finn, C., Lee, A.X., Levine, S.: Self-Supervised Visual Planning with Temporal Skip Connections. In: CoRL. pp. 344–356 (2017)
9. Esser, P., Rombach, R., Ommer, B.: Taming Transformers for High-Resolution Image Synthesis. arXiv:2012.09841 [cs] (Jun 2021)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
11. Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T.: Axial Attention in Multidimensional Transformers. arXiv preprint arXiv:1912.12180 (2019)
12. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
13. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A.: The open images dataset v4. International Journal of Computer Vision **128**(7), 1956–1981 (2020)
15. Li, Y., Min, M., Shen, D., Carlson, D., Carin, L.: Video generation from text. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. pp. 740–755. Springer (2014)
17. Luc, P., Clark, A., Dieleman, S., Casas, D.d.L., Doron, Y., Cassirer, A., Simonyan, K.: Transformation-based adversarial video prediction on large-scale data. arXiv preprint arXiv:2003.04035 (2020)
18. Miao, J., Wei, Y., Wu, Y., Liang, C., Li, G., Yang, Y.: VSPW: A Large-scale Dataset for Video Scene Parsing in the Wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4133–4143 (2021)
19. Moing, G.L., Ponce, J., Schmid, C.: CCVS: Context-aware Controllable Video Synthesis. arXiv preprint arXiv:2107.08037 (2021)

20. Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C.: Moments in time dataset: One million videos for event understanding. IEEE transactions on pattern analysis and machine intelligence **42**(2), 502–508 (2019)
21. van den Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., Kavukcuoglu, K.: Conditional image generation with pixelcnn decoders. arXiv preprint arXiv:1606.05328 (2016)
22. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. arXiv preprint arXiv:1711.00937 (2017)
23. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2337–2346 (2019)
24. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, \., Shazeer, N., Ku, A., Tran, D.: Image transformer. arXiv preprint arXiv:1802.05751 (2018)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs] (Feb 2021)
26. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
27. Rakhimov, R., Volkhonskiy, D., Artemov, A., Zorin, D., Burnaev, E.: Latent Video Transformer. arXiv preprint arXiv:2006.10704 (2020)
28. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. arXiv preprint arXiv:1906.05909 (2019)
29. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-Shot Text-to-Image Generation. arXiv:2102.12092 [cs] (Feb 2021)
30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. arXiv:1409.0575 [cs] (Jan 2015)
31. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29**, 2234–2242 (2016)
32. Tao, M., Tang, H., Wu, S., Sebe, N., Jing, X.Y., Wu, F., Bao, B.: Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. arXiv preprint arXiv:2008.05865 (2020)
33. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1526–1535 (2018)
34. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
35. Van Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: International Conference on Machine Learning. pp. 1747–1756. PMLR (2016)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, \., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
37. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings

of the IEEE/CVF International Conference on Computer Vision. pp. 4581–4591 (2019)
38. Weissenborn, D., Täckström, O., Uszkoreit, J.: Scaling autoregressive video models. In: ICLR (2020)
39. Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., Duan, N.: GODIVA: Generating Open-DomaIn Videos from nAtural Descriptions. arXiv:2104.14806 [cs] (Apr 2021)
40. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5288–5296 (2016)
41. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1316–1324 (2018)
42. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: VideoGPT: Video Generation using VQ-VAE and Transformers. arXiv preprint arXiv:2104.10157 (2021)
43. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. pp. 173–190. Springer (2020)
44. Zhang, H., Koh, J.Y., Baldridge, J., Lee, H., Yang, Y.: Cross-modal contrastive learning for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 833–842 (2021)
45. Zhang, Y., Yan, W., Abbeel, P., Srinivas, A.: VideoGen: Generative Modeling of Videos using VQ-VAE and Transformers (Sep 2020)
46. Zhu, M., Pan, P., Chen, W., Yang, Y.: Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5802–5810 (2019)