AUTOREGRESSIVE VIDEO GENERATION WITHOUT VECTOR QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating a video causally in an autoregressive manner is considered a promising path toward infinite video generation in a flexible context. Prior autoregressive approaches typically rely on vector quantization to convert a video into a discrete-valued space, which could raise challenges in efficiency when modeling long videos. In this work, we propose a novel approach that enables autoregressive video generation without vector quantization. We propose to reformulate the video generation problem as an autoregressive modeling framework integrating temporal *frame-by-frame* prediction and spatial *set-by-set* prediction. Unlike raster-scan prediction in prior autoregressive models or joint distribution modeling of fixed-length tokens in diffusion models, our approach maintains the causal property of GPT-style models for flexible in-context capabilities, while leveraging bidirectional modeling within individual frames for efficiency. We train a novel video autoregressive model with the proposed approach, termed **NOVA**. Our results demonstrate that **NOVA** fully surpasses prior autoregressive video models in data efficiency, inference speed, visual fidelity, and video fluency, even with a much smaller model capacity, *i.e.*, 0.6B parameters. **NOVA** generalizes well across extended video durations and enables diverse zero-shot applications in one unified model. Additionally, with a significantly lower training cost, **NOVA** outperforms state-of-the-art image diffusion models in text-to-image generation tasks. We will release all weights, models, and code to facilitate the reproduction of **NOVA** and further development.

1 INTRODUCTION

Autoregressive large language models (LLMs) (Radford et al. (2019b); Touvron et al. (2023)) have become a foundational architecture in natural language processing (NLP), exhibiting emerging capabilities in in-context learning and long-context reasoning. In autoregressive (AR) vision generation domain, prior approaches (Ramesh et al. (2021); Ding et al. (2021); Yu et al. (2022); Yan et al. (2021); Villegas et al. (2022); Kondratyuk et al. (2023)) typically transform images or video clips (Yan et al. (2021); Hu et al. (2023); Hong et al. (2022); Villegas et al. (2022)) into a discrete-valued token space using vector quantization (Van Den Oord et al. (2017); Esser et al. (2021)), which are then flattened into sequences for token-by-token prediction. However, training high-fidelity vector-quantized tokenizers while maintaining high compression rates is challenging. Hence, the cost of autoregressive inference increases substantially with higher image resolutions or longer video sequences.

In contrast, diffusion-based models (Ho et al. (2020); Nichol & Dhariwal (2021)) can easily benefit from non-quantized high-fidelity tokenizers (Esser et al. (2021)), thereby learning video sequences in a highly-compressed latent space, significantly reducing training and inference costs. However, most diffusion models learn only the joint distribution of a fixed number of frames, lacking the flexibility to generate videos with varied lengths. More importantly, they miss the in-context learning ability of AR models, *i.e.*, integrating multiple tasks into a unified framework like GPT. Therefore, equipping AR models with no-quantized visual tokenizers has become an area of active research.

A most recent step in this direction is MAR (Li et al. (2024b)), which utilizes non-quantized vectors as visual tokens and models the conditional probability between token sets (Chang et al. (2022)), enabling set-by-set AR generation for images. This design preliminarily connects AR to non-quantized tokenizers in the class-to-image task, but its applicability to more complex scenarios (*e.g.*, text-to-image, text-to-video) remains unclear. Furthermore, its set-by-set AR manner is still faced with challenges for in-context learning, *i.e.*, supporting multiple tasks with a single framework.

In this paper, we develop **NOVA**, a generalized AR video generation model based on non-quantized tokenizers, simultaneously taking advantage of **1**) high-fidelity and high-rate visual compression for low training and inference cost, and **2)** in-context learning for integrating multiple generative tasks in a unified model. Concretely, we factorize AR video generation into temporal frame-by-frame prediction and spatial set-by-set prediction. **NOVA** regressively predicts each frame in a casual order across temporal scale, and predicts each token set in a random order across spatial scale. In this way, text-to-video generation can be regarded as a fundamental task that implicitly and comprehensively encompasses various generative tasks, including text-to-image, image-to-video, text&image-to-video, *etc*. Meanwhile, benefit from the non-quantized tokenizer, **NOVA** (0.6B parameters) requires only 192 GPU-days on an A100 (40G) for training. For text-to-video generation, **NOVA** efficiently matches the performance of diffusion models of similar scale, surpassing AR counterparts in data efficiency, inference speed, and video fluency. For text-to-image generation, it outperforms state-of-the-art image diffusion models at significantly lower training costs. Additionally, **NOVA** also demonstrates strong zero-shot generalization across various contexts. We believe that **NOVA** paves the way for future research in effective and efficient AR video generation.

2 RELATED WORKS

2.1 DIFFUSION MODELS FOR VISUAL GENERATION.

Diffusion models (Ho et al. (2020); Song et al. (2020)) have made significant advances in visual generation, including text-to-image tasks (Esser et al. (2024); Betker et al. (2023); Baldridge et al. (2024)) and text-to-video tasks (Brooks et al. (2024); Lab & etc. (2024); Blattmann et al. (2023)). Image diffusion models typically model the joint distribution of fixed-length tokens in pixel (Ho et al. (2020); Nichol et al. (2021); Hoogeboom et al. (2023)) or latent space (Rombach et al. (2022); Esser et al. (2024); Betker et al. (2023); Chen et al. (2023)). Besides, video diffusion models further introduce temporal layers to capture relationships between a fixed number of video frames. After training, additional tasks and modalities are added by incorporating extra inference tricks (Meng et al. (2021)), structure moderation (Blattmann et al. (2023); Esser et al. (2023); Liew et al. (2023)), and adapter layers (Zhang et al. (2023b); Guo et al. (2023)). Although these strategies can be composable, they stand in contrast to the autoregressive approaches (Kondratyuk et al. (2023); Hong et al. (2022); Radford (2018); Touvron et al. (2023)), which trains a single model end-to-end for multi-task learning, offering notable context scalability and zero-shot generalizability across diverse application scenarios, especially in extending video generation duration.

2.2 AUTOREGRESSIVE MODELS FOR VISUAL GENERATION

Raster-scan Autoregressive Models are typically implemented on the discrete-valued RGB pixels (Kalchbrenner et al. (2017); Reed et al. (2017)) or latent space (Esser et al. (2021); Van Den Oord et al. (2017)), analogous to their language counterparts (Radford et al. (2019a); Anil et al. (2023)). Recent studies involve scalable autoregressive transformers to generate token sequences in the raster-scan order for image generation (Ramesh et al. (2021); Ding et al. (2021; 2022); Yu et al. (2022); Sun et al. (2024b)), and video generation (Yan et al. (2021); Kondratyuk et al. (2023); Nash et al. (2022)). Specifically, VAR (Tian et al. (2024)) introduces next-scale prediction to progressively process the token-by-token sequence across multiple resolutions, leading to improved image quality.

Masked Autoregressive Models further develop a masked generative models (Chang et al. (2022)) to introduce a generalized autoregressive concept. They introduce a bidirectional transformer and predict randomly masked tokens by attending to unmasked conditions. This makes up for the suboptimal modeling and inefficient inference of sequentially line-by-line strategy, which inspires a series of subsequent works in text-to-image (Chang et al. (2023)) and text-to-video generation (Hong et al. (2022); Yu et al. (2023); Villegas et al. (2022)). Particularly, MAR (Li et al. (2024b)) decouples discrete tokenizers from autoregressive models and utilizes a diffusion procedure for per-token probability distributions. It is fully validated in the class-to-image field, holding great potential in the text-to-image domain. However, its application to text-to-video generation intuitively requires a masked autoregressive process across entire video frames, challenging multi-context learning and

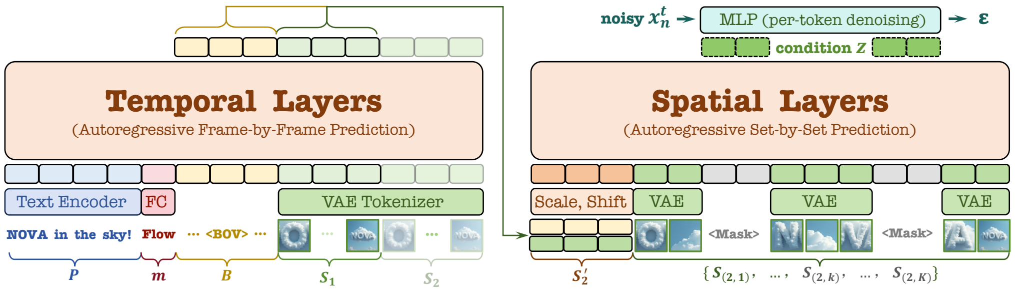


Figure 1: Overview of our proposed NOVA. With text inputs and video flow, NOVA performs an autoregressive modeling framework integrating temporal frame-by-frame prediction and spatial set-by-set prediction. Finally, we implement diffusion denoising in a continuous-values space.

training efficiency. In contrast, our NOVA model breaks down video generation into frame-by-frame temporal predictions combined with spatial set-by-set predictions. This allows each frame to act as a meta causal unit, enabling extended video duration and zero-shot generalizability across various contexts. Besides, the subsequent spatial set-of-tokens prediction unlocks the power of bidirectional modeling patterns, enhancing inference efficiency while preserving visual quality and fidelity.

3 METHODOLOGY

We first review two categories of autoregressive video generation in Sec. 3.1. In Sec. 3.2-3.4, we introduce framework pipeline and implementation details of our NOVA, illustrated in Figure 1.

3.1 RETHINKING AUTOREGRESSIVE MODELS FOR VIDEO GENERATION

Table 1: Symbology Settings.

N, n	The number of all video tokens.
F, f	The number of all video frames.
K, k	The number of sets in an image.

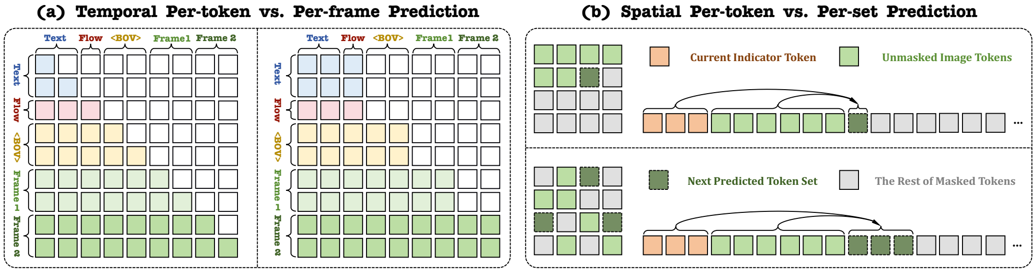
As mentioned above, we regard text-to-video generation and autoregressive (AR) model as the basic task and means, respectively. We briefly retrospect related technical background. There exist two types of AR video generation approaches: **(1) Token-by-token generation via raster-scan order.** These studies (Kondratyuk et al. (2023)) perform causal per-token prediction within video frame sequence, which is defined as follows:

$$p\left(C, x_1, \ldots, x_N\right) = \prod_{n}^{N} p\left(x_n \,|\, C, x_1, \ldots, x_{n-1}\right), \tag{1}$$

where C indicates various condition contexts, *e.g.*, label, text, image, and *etc*. Note that x_n denotes n-th token of N video raster-scale tokens. In contrast, **(2) Masked set-by-set generation in a random order** treats all tokens inside each video frame equally, and implements per-set prediction in a bidirectional transformer decoder (Yu et al. (2023)). However, such a generalized AR are trained by extensive synchronous modeling on large fixed-length video frames, potentially resulting in poor context scalability and coherence issues on extended video duration. Hence, NOVA attempts to unbind per-set generation inside one video image from pre-frame prediction across video sequence, accommodating temporal causal and spatial paradigms as a generalized AR framework.

3.2 TEMPORAL AUTOREGRESSIVE MODELING VIA FRAME-BY-FRAME PREDICTION

Inspired by (Zhuo et al. (2024)), we first utilize a pre-trained language model (Javaheripi et al. (2023)) to encode text prompts into a token sequence before the training and inference stages. To better control video dynamics, we adopt OpenCV (cv2) (Bradski (2000)) to compute the optical flow between adjacent frames, and transform their average value across the entire video via a fully-connected layer (FC). Besides, we employ open-source 3D variational autoencoder (VAE) (Lab & etc. (2024)) with 4 temporal strides to translate multiple frames into the latent space, followed by a patch embedding layer with 4 spatial strides to align with the channel dimensions of the subsequent



(a) Temporal Per-token vs. Per-frame Prediction

(b) Spatial Per-token vs. Per-set Prediction

Figure 2: Overview of our block-wise temporal and spatial generalized autoregressive attention. Different from per-token generation, NOVA regressively predicts each frame in a casual order across the temporal scale, and predicts each token set in a random order across the spatial scale.

transformer. Notably, next-token prediction in early AR models seems counter-intuitive for undirected visual patches within a single image and suffers from high latency during inference. In contrast, video frames can naturally be viewed as a causal sequence, with each frame acting as a meta unit for AR generation. Therefore, we implement block-wise causal masking attention depicted on Figure 2(a), ensuring that each frame can only attend to the text prompts, video flow, and its preceding frames, while allowing all current frame tokens to be visible to each other as follows:

$$p\left(P, m, B, S_1, ..., S_F\right) = \prod_{f}^{F} p\left(S_f \,|\, P, m, B, S_1, ..., S_{f-1}\right), \tag{2}$$

where P, m indicate text prompts and video flow respectively. Here, S_f denotes the overall tokens of f-th video frame, and B represent learnable begin-of-video (BOV) embeddings for predicting initial video frame, the number of which corresponds to the patch number of one single frame. Note that we add 1-D and 2-D sine-cosine embeddings (Vaswani et al. (2017)) with video frame features to indicate time and position information respectively, which are convenient for temporal and spatial extrapolation. From equation 2, we can reformulate text-to-image and image-to-video generation as $p\left(S_1 \,|\, P, m, B\right)$ and $p\left(S_f \,|\, \varnothing, m, B, S_1, ..., S_{f-1}\right)$. This generalized causal process can synchronously model the condition contexts for each video frame, greatly enhancing training efficiency, and allow kv-cache technology for fast decoding procedure during inference.

3.3 SPATIAL AUTOREGRESSIVE MODELING VIA SET-BY-SET PREDICTION

Inspired by (Chang et al. (2022); Li et al. (2024b)), we define each token set with multiple tokens from random directions as a meta causal token, facilitating a generalized AR process with efficient parallel decoding. Notably, we tried to utilize the temporal layers' outputs targeting one frame as indicator features to assist the spatial layers, gradually decoding all randomly masked token sets within the corresponding image. However, this approach resulted in image structure collapse and inconsistent video fluency over the increasing number of frames. We hypothesize that this occurs because the indicator features from adjacent frames are similar, making it difficult to accurately learn continuous and imperceptible motion changes without explicit modeling. Besides, the indicator features derived from the ground-truth contextual frame during training contribute to weak robustness and stability of spatial AR layers against cumulative inference errors.

To address this issue, we introduce a Scaling and Shift Layer that reformulates cross-frame motion changes by learning relative distribution variations within a unified space, rather than directly modeling the unreferenced distribution of the current frame. Notably, we select the BOV-attended output of the temporal layers as the anchor feature set, as it serves as the initial feature set with significantly less noise accumulation than subsequent frame feature sets. Specifically, we first translate the features from current frame set into dimension-wise variance and mean parameters γ and β via multi-layer perception (MLP). After that, we affine the normalized features from the anchor set into indicator features S'_f via channel-wise scale and shift operation. Specially, we explicitly set $\gamma = 1$ and $\beta = 0$ for the first frame. With unmasked token features, we predict randomly masked visual tokens in a set-by-set order through a bidirectional paradigm, which can be formulated as follows:

$$p\left(S'_f, S_{(f,1)}, ..., S_{(f,K)}\right) = \prod_{k}^{K} p\left(S_{(f,k)} \,|\, S'_f, S_{(f,1)}, ..., S_{(f,k-1)}\right), \tag{3}$$

where S'_f denotes the indicator features for generating f-th video frame, and $S_{(f,k)}$ denotes k-th token set of f-th video frame. We add 2-D sine-cosine embeddings with masked and unmasked tokens to indicate their relative position. This generalized spatial AR prediction leverages powerful bidirectional patterns within single-image tokens and achieves efficient inference with parallel masked decoding. *Notably, we incorporate post-norm layers before the residual connections in both temporal and spatial AR layers.* Our empirical findings show that this design effectively addresses architectural and optimization challenges that previously hindered stable training in generalized video generation.

3.4 DIFFUSION PROCEDURE DENOISING FOR PER-TOKEN PREDICTION

During training, we import *diffusion loss* (Li et al. (2024b)) to estimate per-token probability in a continuous-valued space. For example, we define one ground-truth token as x_n and corresponding NOVA's output as z_n. The loss function can be formulated as a denoising criterion:

$$\mathcal{L}(x_n \,|\, z_n) = \mathbb{E}_{\varepsilon,t}\left[\left\|\epsilon - \epsilon_\theta\left(x_n^t \mid t, z_n\right)\right\|^2\right]. \tag{4}$$

Here ϵ is a Gaussian vector sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and noisy data $x_n^t = \sqrt{\bar{\alpha}_t} x_n + \sqrt{1 - \bar{\alpha}_t}\epsilon$, where $\bar{\alpha}_t$ is a noise schedule (Nichol & Dhariwal (2021)) indexed by a time step t. The noise estimator ϵ_θ is multiple MLP blocks parameterized by θ. The notation $\epsilon_\theta(x_n^t \,|\, t, z_n)$ means that this network takes x_n^t as the input, and is conditional on both t and z_n. We follow (Li et al. (2024b)) to sample t by 4 times during training for each image.

During inference, we sample x_n^T from a random Gaussian noise $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and denoise it step-by-step by sequentially sampling x_n^T to x_n^0 via $x_n^{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_n^t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(x_n^t | t, z_n\right)\right) + \sigma_t\epsilon$, where σ_t is the noise level at time step t, and ϵ is sampled from the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

4 EXPERIMENT

4.1 EXPERIMENT SETUP

Datasets. We involve several diverse, curated, and high-quality datasets to facilitate the training of our NOVA. For text-to-image training, we construct 16M image-text pairs sourced from DataComp (Gadre et al. (2024)), COYO (Byeon et al. (2022)), Unsplash (team (2020)), and JourneyDB (Sun et al. (2024a)), while for text-to-video training, we select 12M video-text pairs on a subset OpenSora Plan (Lab & etc. (2024)) of Panda-70M (Chen et al. (2024b)) and internal video-text pairs. We further collect 1M of high-resolution video-text pairs from Pexels (team (2014)) to fine-tune our final video generation model. Following (Diao et al. (2024)), we train a caption engine based on EMU2 (17B) (Sun et al. (2023)) model to create high-quality descriptions for our image and video datasets. The maximum text length is set to 256.

Architectures. We mostly follow (Li et al. (2024b)) to build NOVA's spatial AR layer and denoising MLP block, including a layer sequence of LayerNorm (Lei Ba et al. (2016)), AdaLN (Huang & Belongie (2017)), linear layer, SiLU activation (Elfwing et al. (2018)), and another linear layer. We configure the temporal encoder, spatial encoder, and decoder with 16 layers each, all using 1024 dimensions, while the denoising MLP consists of 3 blocks with 1280 dimensions. The spatial layers adopt the encoder-decoder architecture of MAR (Li et al. (2024b)), similar to MAE (He et al. (2022)). Specifically, the encoder processes the visible patches for reconstruction. The decoder further processes visible and masked patches for generation. In total, NOVA comprises approximately 600 million parameters. To capture the latent features of images, we employ a pre-trained and frozen VAE from (Zheng et al. (2024)), which achieves 4× compression in the temporal dimension and 8×8 compression in the spatial dimension. And we adopt the masking and diffusion schedulers from (Li et al. (2024b); Nichol & Dhariwal (2021)), using a masking ratio between 0.7 and 1.0 during training, and progressively reducing it from 1.0 to 0 following a cosine schedule (Chang et al. (2023))

Table 2: **Text-to-image evaluation on T2I-CompBench.** We highlight the best values in blue , and the second-best values in green . The baseline data come from Huang et al. (2023).

Model	ModelSpec		Attribute Binding			Object Relationship		Complex ↑	Mean ↑	A100 days ↓
	#Param.	**Data**	**Color ↑**	**Shape ↑**	**Texture ↑**	**Spatial ↑**	**Non-Spatial ↑**			
Stable-v1.4	1B	2B	0.3765	0.3576	0.4156	0.1246	0.3079	0.3080	0.3150	-
Stable-v2	1B	2B	0.5065	0.4221	0.4922	0.1342	0.3096	0.3386	0.3672	-
Composable-v2	1B	2B	0.4063	0.3299	0.3645	0.0800	0.2980	0.2898	0.2947	-
Structured-v2	1B	2B	0.4990	0.4218	0.4900	0.1386	0.3111	0.3355	0.3660	-
AttnExct-v2	1B	2B	0.6400	0.4517	0.5963	0.1455	0.3109	0.3401	0.4141	-
GORS	1B	2B	0.6603	0.4785	0.6287	0.1815	0.3193	0.3328	0.4335	-
DALL·E-v2	6.5B	650M	0.5750	0.5464	0.6374	0.1283	0.3043	0.3696	0.4268	-
SDXL	2.6B	-	0.6369	0.5408	0.5637	0.2032	0.3027	0.3662	0.4441	-
Lumina-Next	2B	14M	0.5083	0.3330	0.4315	0.1823	0.3027	0.3662	0.3540	288
PixArt-α	0.6B	25M	0.6886	0.5582	0.7044	0.2082	0.3179	0.4117	0.4815	753
NOVA (T2I)	0.6B	16M	0.7164	0.5608	0.6861	0.2915	0.3094	0.4092	0.4955	127
NOVA (T2V)	0.6B	29M	0.6429	0.4657	0.6210	0.1824	0.3053	0.3487	0.4276	267

during inference. In line with common practice (Ho et al. (2020)), we train with a 1000-step noise schedule but default to 100 steps for inference.

Training details. NOVA is trained with sixteen 8-A100 (40G) nodes. We utilize the AdamW optimizer (Loshchilov et al. (2017)) ($\beta_1 = 0.9, \beta_2 = 0.95$) with a weight decay of 0.02 and a base learning rate of 1e-4 in all experiments. The peak learning rate is adjusted for different batch sizes during training using the scaling rule (Goyal (2017)) : $\text{lr} = \text{base\_lr} \times \text{batchsize}/256$. We train text-to-image models from scratch and then load these weights to train text-to-video models.

Evaluation. We use T2I CompBench (Huang et al. (2023)) to assess the alignment between the generated images and text condition. We adhere to the guidelines by generating 10 image samples for each of the 1800 compositional text prompts, which are divided into three main categories and six subcategories. Each sample has a resolution of 512×512. We use VBench (Huang et al. (2024)) to evaluate the capacity of text-to-video generation across 16 dimensions. For a given text prompt, we randomly generate 5 samples, each with a video size of 29×768×480.

4.2 MAIN RESULTS

NOVA outperforms existing text-to-image models with superior performance and efficiency. In Table 2, we compare NOVA with several recent state-of-the-art models, including Stable-v1.4 (Rombach et al. (2022)), Stable-v2 (Rombach et al. (2022)), Composable-v2 (Liu et al. (2022a)), Structured-v2 (Feng et al. (2022)), AttnExct-v2 (Chefer et al. (2023)), GORS (Huang et al. (2023)), DALL·E-v2 (Ramesh et al. (2022)), SDXL (Podell et al. (2023)), Lumina-Next (Zhuo et al. (2024)), PixArt-α (Chen et al. (2023)) from T2I-CompBench (Huang et al. (2023)). After text-to-image training, NOVA achieves state-of-the-art (SOTA) performance across 4 out of 7 standard metrics and ranks second in two others. This validate its strong capabilities in compositional text-to-image generation, particularly in accurately capturing color binding and spatial relationships between objects. Notably, these results are achieved with the smallest model capacity and the second smallest data scale, requiring only 16% training overhead of the best competitor PixArt-α. *Last but not least, we evaluate our text-to-video NOVA model outperforms most specialized text-to-image generators*, *e.g.*, SD-v2 and DALL·E-v2. This underscores the robustness and versatility of our model in multi-context scenarios, with text-to-video generation as the fundamental training task.

NOVA rivals diffusion text-to-video models and significantly suppresses the AR counterpart. We emphasize that the current version of our NOVA is designed to generate videos at 29 frames and can extend video length through the *pre-filling* of recently generated frames. We conduct a quantitative comparison between NOVA and leading open-source and proprietary text-to-video models. As shown in Table 3, NOVA remarkably outperforms the counterpart CogVideo (Hong et al. (2022)) across various text-to-video evaluation metrics, despite being significantly smaller in size (0.6B vs. 9B). Additionally, we also compared NOVA with the latest diffusion models at a similar data scale, including closed-source Gen-2 (VID (2023)), Kling (Kuaishou (2024)), Gen-3 (Runway

Table 3: **Text-to-video evaluation on VBench.** We have classified existing video generation methods into different categories for clarity. The baseline data are sourced from Huang et al. (2024).

Models	#Param.	Data	Latency	Total Score	Quality Score	Semantic Score	Aesthetic Quality	Object Class	Multiple Objects	Human Action	Spatial Relationship	Scene	Appearance Style	Overall Consistency
Closed-source models														
Gen-2	-	-	-	80.58	82.47	73.03	66.96	90.92	55.47	89.20	66.91	48.91	19.34	26.17
Kling (2024-07)	-	-	-	81.85	83.39	75.68	61.21	87.24	68.05	93.40	73.03	50.86	19.62	26.42
Gen-3	-	-	-	82.32	84.11	75.17	63.34	87.81	53.64	96.4	65.09	54.57	24.31	26.69
Diffusion models (w/ Stable Diffusion weights)														
LaVie	3B	25M	-	77.08	78.78	70.31	54.94	91.82	33.32	96.8	34.09	52.69	23.56	26.41
Show-1	4B	10M	-	78.93	80.42	72.98	57.35	93.07	45.47	95.60	53.50	47.03	23.06	27.46
AnimateDiff-v2	1B	10M	-	80.27	82.90	69.75	67.16	90.90	36.88	92.60	34.60	50.19	22.42	27.04
VideoCrafter-v2.0	2B	10M	-	80.44	82.20	73.42	63.13	92.55	40.66	95.00	35.86	55.29	25.13	28.23
T2V-Turbo (VC2)	2B	10M	-	81.01	82.57	74.76	63.04	93.96	54.65	95.20	38.67	55.58	24.42	28.16
Diffusion models														
OpenSora-v1.1	1B	10M	48s	75.66	77.74	67.36	50.12	86.76	40.97	84.20	52.47	38.63	23.50	26.37
OpenSoraPlan-v1.1	1B	4.5M	60s	78.00	80.91	66.38	56.85	76.30	40.35	86.80	53.11	27.17	22.90	26.52
OpenSora-v1.2	1B	32M	55s	79.76	81.35	73.39	56.85	82.22	51.83	91.20	68.56	42.44	23.95	26.85
CogVideoX	2B	35M	90s	80.91	82.18	75.83	60.82	83.37	62.63	98.00	69.90	51.14	24.80	26.66
Autoregressive models														
CogVideo	9B	5.4M	560s	67.01	72.06	46.83	38.18	73.4	18.11	78.20	18.24	28.24	22.01	7.70
NOVA	0.6B	13M	12s	75.84	77.11	70.74	55.79	84.84	53.90	84.80	58.71	50.87	20.76	25.22

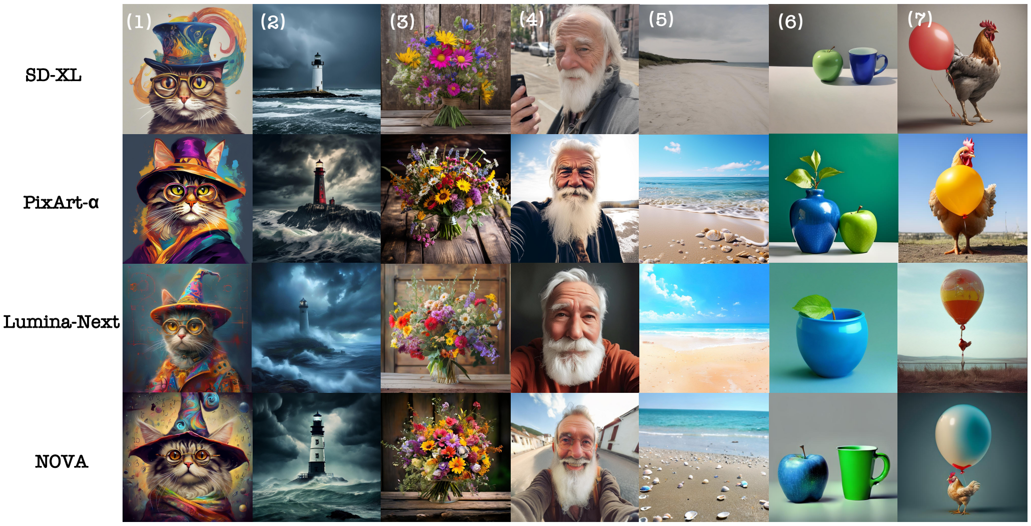
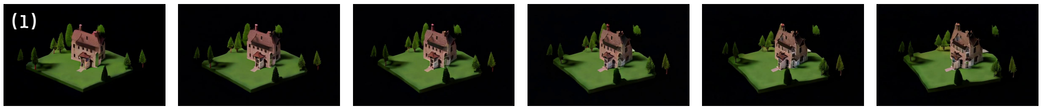


Figure 3: **Text-to-image generation.** Text prompts from left to right: “A digital artwork of a cat styled in a whimsical fashion...”, “A solitary lighthouse standing tall against a backdrop of stormy seas and dark, rolling clouds”, “A vibrant bouquet of wildflowers on a rustic wooden table”, “A selfie of an old man with a white beard”, “A serene, expansive beach with no people”, “A blue apple and a green cup.” and “A balloon on the bottom of a chicken.”

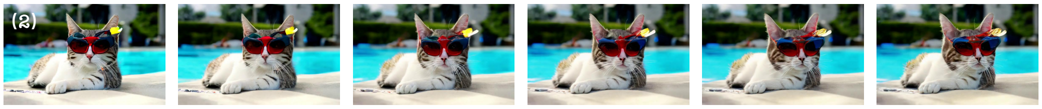
(2024)), open-source LaVie (Wang et al. (2023)), Show-1 (Zhang et al. (2023a)), AnimateDiff-v2 (Guo et al. (2024)), VideoCrafter-v2.0 (Chen et al. (2024a)), T2V-Turbo (VC2) (Li et al. (2024a)), OpenSora-v1.1 (Zheng et al. (2024)), OpenSoraPlan-v1.1 (Lab & etc. (2024)), OpenSora-v1.2 (Lab & etc. (2024)), and CogVideoX (Yang et al. (2024)). The results underscore the effectiveness of text-to-image pre-training within our generalized causal process. Notably, we have narrowed the gap between autoregressive and diffusion methods in modeling large-scale video-text pairs, enhancing both the quality and instruction-following capabilities of video generation. Besides, NOVA shows a significant speed advantage over existing models in terms of inference latency.

4.3 QUALITATIVE RESULTS

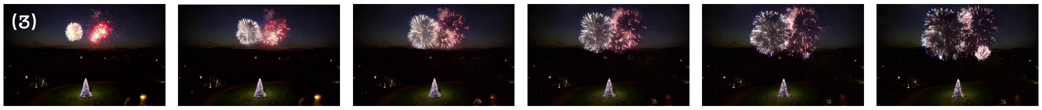
High-Fidelity Image and High-Fluency Video. We present a qualitative comparison of current leading image generation methods in Fig. 3. NOVA demonstrates strong visual quality and fidelity across a range of prompt styles, and excels in color attribute binding and spatial object relationships. We present text-to-video visualizations in Fig. 4, which highlight NOVA’s ability to capture multi-view perspectives, smooth object motion, and stable scene transitions based on the provided text prompts.



Text prompt : A 3D model of a 1800s victorian house.

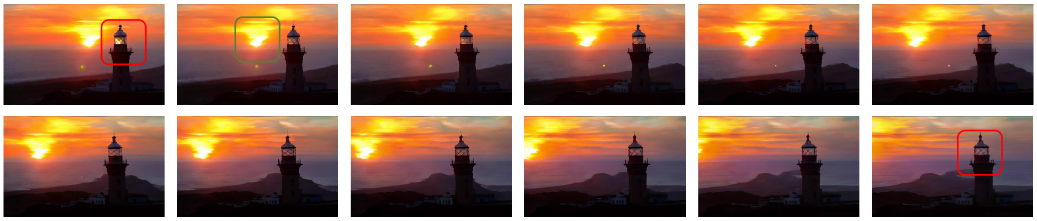


Text prompt : A cat wearing sunglasses and working as a lifeguard at a pool.



Text prompt : A drone view of celebration with Christmas tree and fireworks, starry sky, background.

Figure 4: **Text-to-video generation.** We highlight the keywords in red color. NOVA follows the text prompts and vividly captures the motion of subjects (i.e., 3D model, cat and fireworks).



A lighthouse ... a vibrant sunset sky... to a warm palette of oranges and yellows ... has a classic design with a domed top and a lantern room where the light would be housed, ...to intensify.

Figure 5: **Zero-shot video extrapolation.** We highlight the subjects in red and green respectively. The top images are generated, while the bottom images are extrapolated.

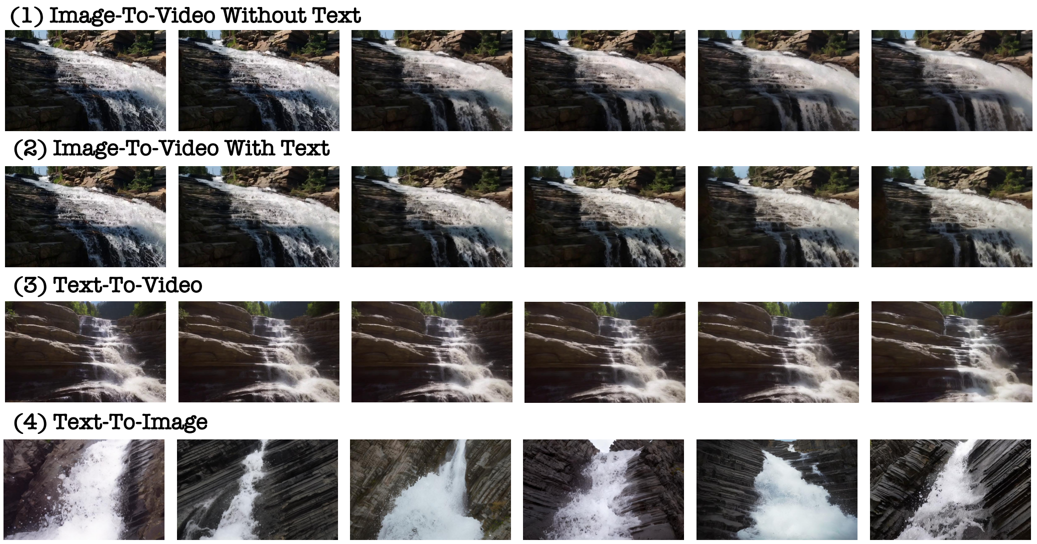
Zero-Shot Generalization on Video Extrapolation. By pre-filling generated frames, NOVA can produce videos that surpass the training length. For example, by shifting both the text and BOV embeddings, we generate 5-second videos that are up to twice the original length, as shown in Fig. 5. We observed that during video extrapolation, NOVA consistently preserves temporal consistency of subject across frames. For instance, when the prompt describes a dome top and a lantern room, the model accurately represents the lighting within the house and captures the transition of a sunset. This further underscores the advantages of causal modeling in long-context video generation tasks.

Zero-Shot Generalization on Multiple Contexts. By pre-filling the reference image, NOVA can generate videos from images, either with or without accompanying text. In Figure 6, we provide a qualitative example. We show that NOVA can simulate realistic motions without text prompts. Moreover, when text is included, perspective movements appear more natural. This indicates that NOVA is able to capture the fundamental physics, such as interaction forces and fluid dynamics.

4.4 ABLATION STUDY

Effectiveness of temporal autoregressive modeling. To highlight the advantages of temporal autoregressive modeling, we have facilitated spatial autoregressive to finish video generation task. Specifically, we modify the attention mask of the temporal layer to bidirectional attention, and randomly predict the entire video sequence using set by set prediction. We observe less subject movement in videos under the same training iterations (Fig. 7). Additionally, in zero-shot generalization across various contexts or video extrapolation, the network output exhibited more artifacts and temporal inconsistencies. Furthermore, this approach is not compatible with kv-cache acceleration during inference, leading to a linear increase in latency with the number of video frames. This further demonstrates the superiority of causal modeling over multitask approaches for video generation.

Effectiveness of Scaling and Shift Layer. To capture cross-frame motion changes, we employ a simple yet effective scaling and shifting layer to explicitly model the relative distribution from the



Text : A cascade of water rushes down a rocky incline, frothing and churning as it descends, is surrounded by rugged, layered rock formations.

Figure 6: **Zero-shot image-to-video generation.** It is evident that NOVA successfully maintains temporal consistency in objects, both with and without text. Such as ensuring ”water continues to flow smoothly.” (Zoom in to view in red box). This highlights NOVA’s capability for zero-shot multitasking.

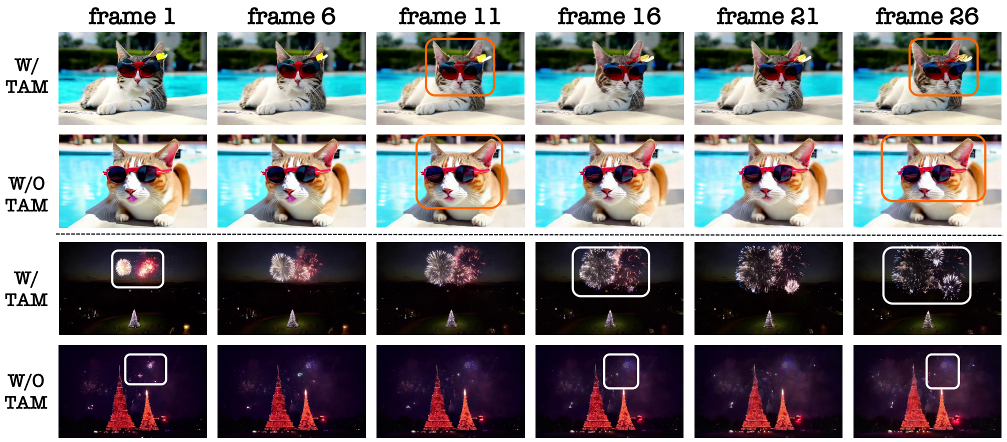
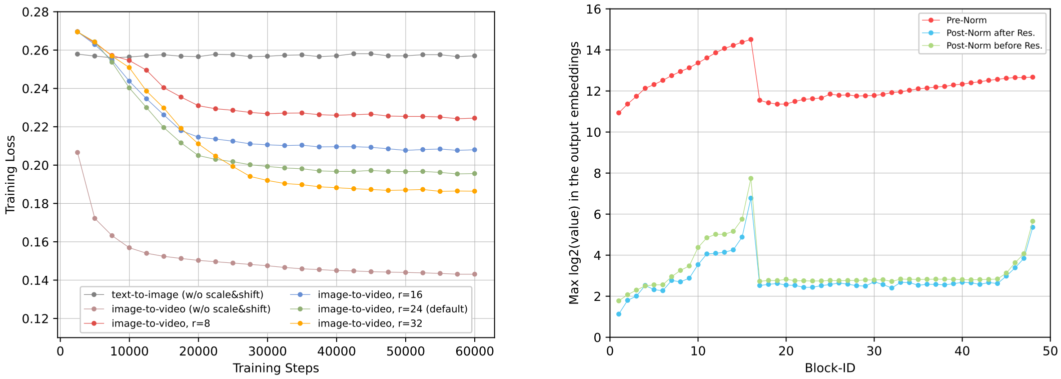


Figure 7: **Temporal autoregressive modeling (TAM) for video generation.** We highlight the subtle changes in frames generated from the same prompt. Compared to spatial-only autoregressive method, the inclusion of TAM enables NOVA to more accurately capture the dynamics of subject movement.

BOV-attended feature space. In Figure 8(a), we demonstrate that this approach significantly reduces the drift between text-to-image and image-to-video generation losses. As we gradually decrease the inner rank of the MLP, the training difficulty increases, leading to a more comprehensive and robust learning process for the network. However, extremely low rank values pose challenges for motion modeling, as they significantly limit the layer’s representation capability (Figure 9). The rank is set to 24 by default in all text-to-video experiments, resulting in more accurate motion predictions.

Effectiveness of Post-Norm Layer. Training large-scale image and video generation models (Ding et al. (2021); Team (2024)) from scratch often poses significant challenges with mixed precision, which is also observed in other visual recognition methods (Liu et al. (2022b)). As shown in



(a) Parameter decomposition for Scaling and Shift layer.

(b) Normalization layer position.

Figure 8: **Ablation studies on NOVA's architecture components.** We carefully examine the two key stability factors in large-scale video generation training, as illustrated in (a) and (b).

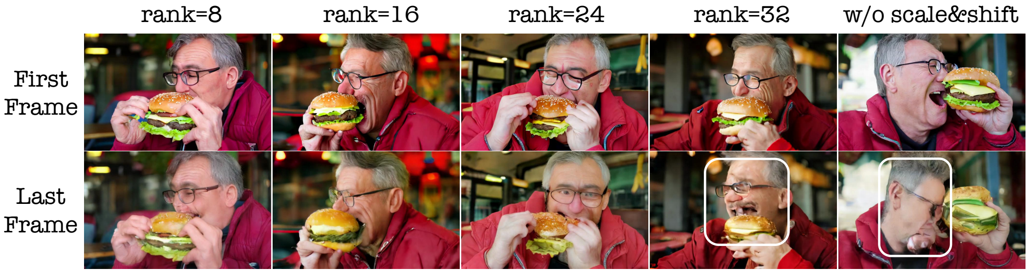


Figure 9: **Visualization of decomposition ranks in the Scaling and Shift layer.** The first row displays the results of the first frame, while the second row presents the results of the last frame.

Figure 8(b), the training process with pre-normalization (Vaswani et al. (2017)) suffers from numerical overflow and variance instability. We attempted various regularization techniques on the residual branch, such as stochastic depth (Huang et al. (2016)) and residual dropout (Vaswani et al. (2017)), but found them to be less effective in alleviating this issue. Inspired by (Liu et al. (2022b)), we introduce post-normalization and empirically discover that it can effectively mitigate the accumulation of model weights compared to pre-normalization, resulting in a smoother and more stable training process.

5 CONCLUSION

In this paper, we present **NOVA**, an innovative autoregressive model designed for both text-to-image and text-to-video generation. **NOVA** delivers exceptional image quality and video fluency while significantly minimizing training and inference overhead. Our key designs—temporal frame-by-frame prediction, spatial set-by-set generation, and continuous-space modeling—enhance autoregressive properties across various contexts and establish robust bidirectional modeling patterns for improved efficiency. Extensive experiments demonstrate that **NOVA** achieves near-commercial quality in image generation, alongside promising fidelity and fluency in video generation. For the first time, **NOVA** provides valuable insights into autoregressive generation for the AIGC community and startups, empowering them to develop high-quality, cost-effective models. As a first step, we would continue scalable experiments with larger models and data scales to explore NOVA's limits in future work.

REFERENCES

Runway. gen-2: Generate novel videos with text, images or video clips, 2023. URL `https://runwayml.com/research/gen-2/`. Accessed: 2023-10-01.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024a.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13320–13331, 2024b.

Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *arXiv preprint arXiv:2406.11832*, 2024.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021.

Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35: 16890–16902, 2022.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7346–7356, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

P Goyal. Accurate, large minibatch sg d: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 15979–15988, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pp. 13213–13232. PMLR, 2023.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 646–661. Springer, 2016.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 2023.

Nal Kalchbrenner, Aäron Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. In *International Conference on Machine Learning*, pp. 1771–1779. PMLR, 2017.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

Kuaishou. Kling ai. `https://klingai.com/`, 2024. Accessed: 2024-10-01.

PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, April 2024. URL `https://doi.org/10.5281/zenodo.10948109`.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *ArXiv e-prints*, pp. arXiv–1607, 2016.

Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *arXiv preprint arXiv:2405.18750*, 2024a.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024b.

Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*, 2023.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pp. 423–439. Springer, 2022a.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12009–12019, 2022b.

Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5, 2017.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Charlie Nash, Joao Carreira, Jacob Walker, Iain Barr, Andrew Jaegle, Mateusz Malinowski, and Peter Battaglia. Transframer: Arbitrary frame prediction with generative models. *arXiv preprint arXiv:2203.09494*, 2022.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Alec Radford. Improving language understanding by generative pre-training. 2018.

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, 2019a.

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, 2019b.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Scott Reed, Aäron Oord, Nal Kalchbrenner, Sergio Gómez Colmenarejo, Ziyu Wang, Yutian Chen, Dan Belov, and Nando Freitas. Parallel multiscale autoregressive density estimation. In *International conference on machine learning*, pp. 2912–2921. PMLR, 2017.

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*, pp. 10684–10695, 2022.

Runway. Gen-3 alpha: A new frontier for video generation, 2024. URL `https://runwayml.com/research/introducing-gen-3-alpha/`. Accessed: 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in Neural Information Processing Systems*, 36, 2024a.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024b.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. *arXiv: 2312.13286*, 2023.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Pexels team. Pexels, royalty-free stock footage website, 2014.

Unsplash team. Unsplash dataset, 2020.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv: 2307.09288*, 2023.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pp. 5998–6008, 2017.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2022.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023.

David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023a.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023b.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024. URL `https://github.com/hpcaitech/Open-Sora`.

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *arXiv preprint arXiv:2406.18583*, 2024.