GenTron: Diffusion Transformers for Image and Video Generation

Shoufa Chen[1,2*] Mengmeng Xu[2*] Jiawei Ren[2] Yuren Cong[2] Sen He[2]
Yanping Xie[2] Animesh Sinha[2] Ping Luo[1] Tao Xiang[2] Juan-Manuel Perez-Rua[2]
[1]The University of Hong Kong [2]Meta

Figure 1. **GenTron: Transformer based diffusion model for high-quality text-to-image/video generation.**

Abstract

*In this study, we explore Transformer-based diffusion models for image and video generation. Despite the dominance of Transformer architectures in various fields due to their flexibility and scalability, the visual generative domain primarily utilizes CNN-based U-Net architectures, particularly in diffusion-based models. We introduce **GenTron**, a family of **Gen**erative models employing **Tr**ansformer-based diffusi**on**, to address this gap. Our initial step was to adapt Diffusion Transformers (DiTs) from class to text conditioning, a process involving thorough empirical exploration of the conditioning mechanism. We then scale GenTron from approximately 900M to over 3B parameters, observing improvements in visual quality. Furthermore, we extend GenTron to text-to-video generation, incorporating novel motion-free guidance to enhance video quality. In human evaluations against SDXL, GenTron achieves a 51.1% win rate in visual quality (with a 19.8% draw rate), and a 42.3% win rate in text alignment (with a 42.9% draw rate). GenTron notably performs well in T2I-CompBench, highlighting its compositional generation ability. We hope GenTron could provide meaningful insights and serve as a valuable reference for future research. Please refer to the website[1] and the arXiv version for the most up-to-date results: https://arxiv.org/abs/2312.04557.*

[1] https://www.shoufachen.com/gentron_website/
[*] Equal contribution.

1. Introduction

Diffusion models have recently shown remarkable progress in content creation, impacting areas such as image generation [27, 55, 57], video generation [5, 7, 28, 60], and editing [14, 34]. Among these notable developments, the CNN-based U-Net architecture has emerged as the predominant backbone design, a choice that stands in contrast to the prevailing trend in natural language processing [8, 19, 62] and computer visual perception [20, 38, 69] domains, where attention-based transformer architectures [63] have empowered a renaissance and become increasingly dominant. To provide a comprehensive understanding of Transformers in diffusion generation and to bridge the gap in architectural choices between visual generation and the other two domains — visual perception and NLP — a thorough investigation of visual generation using Transformers is of substantial scientific value.

We focus on diffusion models with Transformers in this work. Specifically, our starting point is the foundational work known as DiT [45], which introduced a *class*-conditioned latent diffusion model that employs a Transformer to replace the traditionally used U-Net architecture. We overcome the limitation of the original DiT model, which is constrained to handling only a restricted number (*e.g.*, 1000) of predefined classes, by utilizing language embeddings derived from open-world, free-form *text* captions instead of predefined one-hot class embeddings. Along the way, we comprehensively investigate conditioning strategies, including (1) conditioning architectures: adaptive layer norm (adaLN) [46] *vs*. cross-attention [63]; and (2) text encoding methods: a generic large language model [13] *vs*. the language tower of multimodal models [51], or the combination of both of them. We additionally carry out comparative experiments and offer detailed empirical analyses to evaluate the effectiveness of these conditioning strategies.

Next, we explore the scaling-up properties of GenTron. The Transformer architectures have been demonstrated to possess significant scalability in both visual perception [11, 17, 53, 69] and language [8, 19, 49, 50, 62] tasks. For example, the largest dense language model has 540B parameters [12], and the largest vision model has 22B [17] parameters. In contrast, the largest diffusion transformer, DiT-XL [45], only has about 675M parameters, trailed far behind both the Transformers utilized in other domains (*e.g.*, NLP) and recent diffusion arts with convolutional U-Net architectures [15, 47]. To compensate for this considerable lagging, we scale up GenTron in two dimensions, the number of transformer blocks and hidden dimension size, following the scaling strategy in [69]. As a result, our largest model, *GenTron-G/2*, has more than 3B parameters and achieves significant visual quality improvement compared with the smaller one.

Furthermore, we have advanced GenTron from a T2I to a T2V model by inserting a temporal self-attention layer into each transformer block, making the first attempt to use transformers as the exclusive building block for video diffusion models. We also discuss existing challenges in video generation and introduce our solution, the **motion-free guidance (MFG)**. Specifically, This approach involves intermittently *disabling* motion modeling during training by setting the temporal self-attention mask to an identity matrix. Besides, MFG seamlessly integrates with the joint image-video strategy [9, 16, 30, 65], where images are used as training samples whenever motion is deactivated. Our experiments indicate that this approach clearly improves the visual quality of generated videos.

In human evaluations, GenTron outperforms SDXL, achieving a 51.1% win rate in visual quality (with a 19.8% draw rate), and a 42.3% win rate in text alignment (with a 42.9% draw rate). Furthermore, when compared to previous studies, particularly as benchmarked against T2I-CompBench [31]–a comprehensive framework for evaluating open-world compositional T2I generation–GenTron demonstrates superior performance across various criteria. These include attribute binding, object relationships, and handling of complex compositions.

Our **contributions** are summarized as follows: (1) We have conducted a thorough and systematic investigation of transformer-based T2I generation with diffusion models. This study encompasses various conditioning choices and aspects of model scaling. (2) In a pioneering effort, we explore a purely transformer-based diffusion model for T2V generation. We introduce *motion-free guidance*, an innovative technique that efficiently fine-tunes T2I generation models for producing high-quality videos. (3) Experimental results indicate a clear preference for GenTron over SDXL in human evaluations. Furthermore, GenTron demonstrates superior performance compared to existing methods in the T2I-CompBench evaluations.

2. Related Work

Diffusion models for T2I and T2V generation. Diffusion models [27, 43] are a type of generative model that creates data samples from random noise. Later, latent diffusion models [45, 47, 55] are proposed for efficient T2I generation. These designs usually have *1)* a pre-trained Variational Autoencoder [36] that maps images to a compact latent space, *2)* a conditioner modeled by cross-attention [28, 55] to process text as conditions with a strength control [26], and *3)* a backbone network, U-Net [56] in particular, to process image features. The success of diffusion on T2I generation tasks underscores the promising potential for text-to-video (T2V) generation [35, 42, 48, 60]. VDM [30] and Imagen Video [28] extend the image diffusion architecture on the temporal dimension with promising initial results. To

avoid excessive computing demands, video latent diffusion models [5, 9, 25, 41, 64, 72] implement the video diffusion process in a low-dimensional latent space.

Transformer-based Diffusion. Recently, Transformer-based Diffusion models have attracted increasing research interest. Among these, U-ViT [3] treats all inputs as tokens by integrating transformer blocks with a U-net architecture. In contrast, DiT [45] employs a simpler, non-hierarchical transformer structure. MDT [22] and MaskDiT [71] enhance DiT's training efficiency by incorporating the mask strategy [24]. Dolfin [66] is a transformer-based model for layout generation. Concurrently to this work, PixArt-α [10] demonstrates promising outcomes in Transformer-based T2I diffusion. It's trained using a three-stage decomposition process with high-quality data. Our work diverges from PixArt-α in key aspects. Firstly, while PixArt-α emphasizes training efficiency, our focus is on the design choice of conditioning strategy and scalability in T2I Transformer diffusion models. Secondly, we extend our exploration beyond image generation to video diffusion. We propose an innovative approach in video domain, which is not covered by PixArt-α.

3. Method

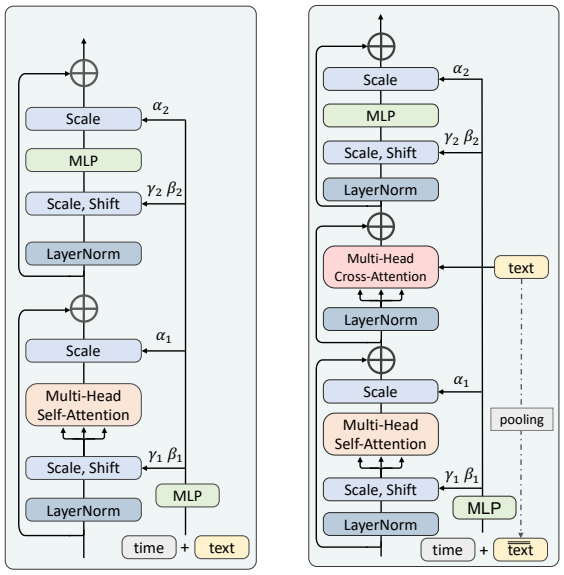
We first introduce the preliminaries in Section 3.1, and then present the details of GenTron for text-to-image generation in Section 3.2, which includes text encoder models, embedding integration methods, and scaling up strategy of GenTron. Lastly, in Section 3.3, we extend GenTron's application to video generation, building on top of the T2I foundations laid in previous sections.

3.1. Preliminaries

Diffusion models. Diffusion models [27] have emerged as a family of generative models that generate data by performing a series of transformations on random noise. They are characterized by a forward and a backward process. Given an instance from the data distribution $x_0 \sim p(x_0)$, random Gaussian noise is iteratively added to the instance in the forward noising process to create a Markov Chain of random latent variable $x_1, x_2, ..., x_T$ following:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_i}x_{t-1}, \beta_t\mathbf{I}), \qquad (1)$$

where $\beta_1, ...\beta_T$ are hyperparameters corresponding to the noise schedule. After a large enough number of diffusion steps, x_T can be viewed as a standard Gaussian noise. A denoising network ϵ_θ is further trained to learn the backward process, *i.e.*, how to remove the noise from a noisy input [27]. For inference, an instance can be sampled starting from a random Gaussian noise $x_T \sim \mathcal{N}(0; \mathbf{I})$ and denoised



(a) adaLN-Zero (b) Cross attention

Figure 2. **Text embedding integration architecture.** We directly adapt adaLN from DiT [45], substituting the one-hot *class* embedding with *text* embedding. For cross attention, different from the approach in [45], we maintain the use of adaLN to model the combination of *time* embedding and the aggregated *text* embedding.

step-by-step following the Markov Chain, *i.e.*, by sequentially sampling x_{t-1} to x_0 with $p_\theta(x_{t-1}|x_t)$:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\Big) + \sigma_t\mathbf{z}, \qquad (2)$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$, $\alpha_t = 1-\beta_t$ and σ_t is the noise scale. In practical application, the diffusion sampling process can be further accelerated using different sampling techniques [40, 61].

Latent diffusion model architectures. Latent diffusion models (LDMs) [55] reduce the high computational cost by conducting the diffusion process in the latent space. First, a pre-trained autoencoder [21, 36] is utilized to compress the raw image from pixel to latent space, then the diffusion models, which are commonly implemented with a U-Net [56] backbone, work on the latent space. Peebles *et al*. proposed DiT [45] to leverage the transformer architecture as an alternative to the traditional U-Net backbone for *class*-conditioned image generation, adopting the adaptive layernorm (adaLN [46]) for class conditioning mechanism, as shown in Figure 2a.

3.2. Text-to-Image GenTron

Our GenTron is built upon the DiT-XL/2 [45], which converts the latent of shape 32×32×4 to a sequence of non-overlapping tokens with a 2×2 patchify layer [20]. Then,

these tokens are sent into a series of transformer blocks. Finally, a standard linear decoder is applied to convert these image tokens into latent space.

While DiT has shown that transformer-based models yield promising results in *class*-conditioned scenarios, it did not explore the realm of T2I generation. This field poses a considerable challenge, given its less constrained conditioning format. Moreover, even the largest DiT model, DiT-XL/2, with its 675 million parameters, is significantly overshadowed by current U-Nets [15, 47], which boast over 3 billion parameters. To address these limitations, our research conducts a thorough investigation of transformer-based T2I diffusion models, focusing specifically on text conditioning approaches and assessing the scalability of the transformer architecture by expanding GenTron to more than 3 billion parameters.

3.2.1 From *Class* to *Text* Condition

T2I diffusion models rely on textual inputs to steer the process of image generation. The mechanism of text conditioning involves two critical components: firstly, the selection of a text encoder, which is responsible for converting raw text into text embeddings, and secondly, the method of integrating these embeddings into the diffusion process. For a complete understanding, we have included in the appendix a detailed presentation of the decisions made in existing works concerning these two components.

Text encoder model. Current advancements in T2I diffusion techniques employ a variety of language models, each with its unique strengths and limitations. To thoroughly assess which model best complements transformer-based diffusion methods, we have integrated several models into GenTron. This includes the text towers from multi-modal models, CLIP [51], as well as a pure large language model, Flan-T5 [13]. Our approach explores the effectiveness of these language models by integrating each model independently with GenTron to evaluate their individual performance and combinations of them to assess the potential properties they may offer when used together.

Embedding integration. In our study, we focused on two methods of embedding integration: adaptive layernorm and cross-attention. (1) **Adaptive layernorm (adaLN).** As shown in Figure 2a, this method integrates conditioning embeddings as normalization parameters on the feature channel. Widely used in conditional generative modeling, such as in StyleGAN [33], adaLN serves as the standard approach in DiT [45] for managing class conditions. (2) **Cross-attention.** As illustrated in Figure 2b, the image feature acts as the `query`, with textual embedding serving as `key` and `value`. This setup allows for direct interaction between the image feature and textual embedding through an attention mechanism [63]. Besides, different from the cross-attention discussed in [45], which processes the class embedding and time embedding together by firstly concatenating them, we maintain the use of adaLN in conjunction with the cross-attention to separately model the *time* embedding. The underlying rationale for this design is our belief that the time embedding, which is consistent across all spatial positions, benefits from the global modulation capabilities of adaLN. Moreover, we also add the pooled text embeddings to the time embedding followings [2, 29, 44, 47].

Model	Depth	Width	MLP Width	#Param.
GenTron-XL/2	28	1152	4608	930.0M
GenTron-G/2	48	1664	6656	3083.8M

Table 1. **Configuration details of GenTron models**.

3.2.2 Scaling Up GenTron

To explore the impact of substantially scaling up the model size, we have developed an advanced version of GenTron, which we refer to as GenTron-G/2. This model was constructed in accordance with the scaling principles outlined in [69]. We focused on expanding three critical aspects: the number of transformer blocks (*depth*), the dimensionality of patch embeddings (*width*), and the hidden dimension of the MLP (*MLP-width*). The specifications and configurations of the GenTron models are detailed in Table 1. Significantly, the GenTron-G/2 model boasts over 3 billion parameters. To our knowledge, this represents the largest transformer-based diffusion architecture developed to date.

3.3. Text-to-Video GenTron

In this subsection, we elaborate on the process of adapting GenTron from a T2I framework to a T2V framework. Sec. 3.3.1 will detail the modifications made to the model's architecture, enabling GenTron to process video data. Furthermore, Sec. 3.3.2 will discuss the challenges encountered in the domain of video generation and the innovative solutions we have proposed to address them.

3.3.1 GenTron-T2V Architecture

Transformer block with temporal self-attention. It is typically a common practice to train video diffusion models from image diffusion models by adding new temporal modeling modules [5, 23, 30, 72]. These usually consist of 3D convolutional layers and temporal transformer blocks that focus on calculating attention along the temporal dimension. In contrast to the traditional approach [5], which involves adding both temporal convolution layers and temporal transformer blocks to the T2I U-Net, our method integrates only lightweight temporal

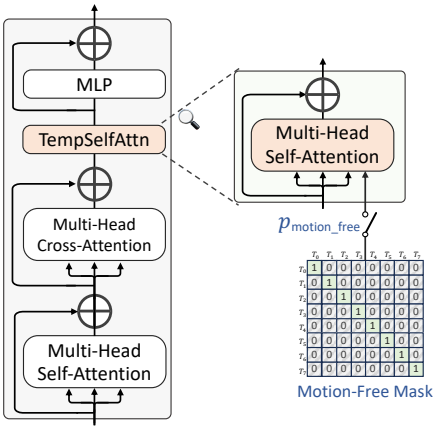


Figure 3. **GenTron Transformer block with** TempSelfAttn **and Motion-Free Mask.** The temporal self-attention layer is inserted between the cross-attention and the MLPs. The motion-free mask, which is an identity matrix, will be utilized in the TempSelfAttn with a probability of $p_{\text{motion\_free}}$. We omit details like text conditioning, LN here for simplicity, which could be found in Figure 2.

self-attention (TempSelfAttn) layers into each transformer block. As depicted in Figure 3, the TempSelfAttn layer is placed right after the cross-attention layer and before the MLP layer. Additionally, we modify the output of the cross-attention layer by reshaping it before it enters the TempSelfAttn layer and then reshape it back to its original format once it has passed through. This process can be formally represented as:

$$\mathbf{x} = \texttt{rearrange}(\mathbf{x}, \texttt{(b t) n d} \rightarrow \texttt{(b n) t d}) \quad (3)$$

$$\mathbf{x} = \mathbf{x} + \text{TempSelfAttn}(\text{LN}(\mathbf{x})) \quad (4)$$

$$\mathbf{x} = \texttt{rearrange}(\mathbf{x}, \texttt{(b n) t d} \rightarrow \texttt{(b t) n d}) \quad (5)$$

where `b`, `t`, `n`, `d` represent the batch size, number of frames, number of patches per frame, and channel dimension, respectively. `rearrange` is a notation from [54]. We discovered that a simple TempSelfAttn layer suffices to capture motion, a finding that aligns with observations in a recent study [65]. In addition, only using TempSelfAttn makes it convenient to *turn on* and *turn off* the temporal modeling, which would be discussed in Sec. 3.3.2.

Initialization. We use the pre-trained T2I model as a basis for initializing the shared layers between T2I and T2V models. In addition, for the newly added TempSelfAttn layers, we initialize the weights and biases of the output project layers to zero. This ensures that at the beginning of the T2V fine-tuning stage, these layers produce a zero output, effectively functioning as an identity mapping in conjunction with the shortcut connection.

3.3.2 Motion-Free Guidance

Challenges encountered. We observed a notable phenomenon in the current T2V diffusion models [5, 30] where the per-frame visual quality significantly lags behind that of T2I models [15, 47, 55, 67]. Furthermore, our analysis revealed a remarkable degradation in visual quality in the T2V models post-fine-tuning, especially when compared to their original T2I counterparts. We note that these problems generally exist in current T2V diffusion models, not limited to our transformer-based T2V.

Problem analysis and insights. We presume that the observed lag in the visual quality of T2V primarily stems from two factors: the nature of video data and the fine-tuning approach. Firstly, publicly available video datasets often fall short in both quality and quantity compared to image datasets. For instance, [58] has more than 2B English image-text pairs, whereas the current widely used video dataset, WebVid-10M [1] contains only 10.7M video-text pairs. Additionally, many video frames are compromised by motion blur and watermarks, further reducing their visual quality. This limited availability hampers the development of robust and versatile video diffusion models. Secondly, the focus on optimizing temporal aspects during video fine-tuning can inadvertently compromise the spatial visual quality, resulting in a decline in the overall quality of the generated videos.

Solution I: joint image-video training. From the data aspect, we adopt the joint image-video training strategy [9, 16, 30, 65] to mitigate the video data shortages. Furthermore, joint training helps to alleviate the problem of domain discrepancy between video and image datasets by integrating both data types for training.

Solution II: motion-free guidance. We treat the temporal motion within a video clip as a special *conditioning* signal, which can be analogized to the textual conditioning in T2I/T2V diffusion models. Based on this analogy, we propose a novel approach, *motion-free guidance* (MFG), inspired by classifier-free guidance [6, 26], to modulate the weight of motion information in the generated video.

In a particular training iteration, our approach mirrors the concept used in classifier-free guidance, where conditioned text is replaced with an empty string. The difference is that we employ an identity matrix to *nullify* the temporal attention with a probability of $p_{\text{motion\_free}}$. This identity matrix, which is depicted in Figure 3 (Motion-Free Mask), is structured such that its diagonal is populated with ones, while all other positions are zeroes. This configuration confines the temporal self-attention to work within a single

Conditioning			Scale	Attribute Binding			Object Relationship		Complex	**Mean**
Text Encoder	Type	Integration		Color	Shape	Texture	Spatial	Non-spatial		
CLIP-L [51]	MM	adaLN-zero	XL/2	36.94	42.06	50.73	9.41	30.38	36.41	34.32
CLIP-L [51]	MM	cross-attn	XL/2	73.91	51.81	68.76	19.26	31.80	41.52	47.84
T5-XXL [13]	LLM	cross-attn	XL/2	74.90	55.40	70.05	20.52	31.68	41.01	48.93
CLIP-T5XXL	MM + LLM	cross-attn	XL/2	75.65	55.74	69.48	20.67	31.79	41.44	49.13
CLIP-T5XXL	MM + LLM	cross-attn	G/2	**76.74**	**57.00**	**71.50**	**20.98**	**32.02**	**41.67**	**49.99**

Table 2. **Conditioning and model scale in GenTron.** We compare GenTron model variants with different design choices on T2I-CompBench [31]. The text encoders are from the language tower of the multi-modal (MM) model, the large language model (LLM), or a combination of them. The GenTron-G/2 with CLIP-T5XXL performs best. Detailed discussions can be found in Sec. 4.2.

frame. Furthermore, as introduced in Sec. 3.3.1, temporal self-attention is the sole operator for temporal modeling. Thus, using a motion-free attention mask suffices to disable temporal modeling in the video diffusion process.

During inference, we have text and motion conditionings. Inspired by [6], we can modify the score estimate as:

$$\begin{aligned}\tilde{\epsilon}_\theta = &\ \epsilon_\theta(x_t, \varnothing, \varnothing) \\ &+ \lambda_T \cdot (\epsilon_\theta(x_t, c_T, c_M) - \epsilon_\theta(x_t, \varnothing, c_M)) \\ &+ \lambda_M \cdot (\epsilon_\theta(x_t, \varnothing, c_M) - \epsilon_\theta(x_t, \varnothing, \varnothing))\end{aligned} \tag{6}$$

where c_T and c_M represent the text conditioning and motion conditioning. λ_T and λ_M are the guidance scale of standard text and that of motion, controlling how strongly the generated samples correspond with the text condition and the motion strength, respectively. We empirically found that fixing $\lambda_T = 7.5$ and adjusting $\lambda_M \in [1.0, 1.3]$ for each example tend to achieve the best result. This finding is similar to [6], although our study utilizes a narrower range for λ_M.

Putting solutions together. We can integrate solution I and II together in the following way: when the motion is omitted at a training step, we load an image-text pair and repeat the image $T-1$ times to create a *pseudo* video. Conversely, if motion is included, we instead load a video clip and extract it into T frames.

4. Experiments

4.1. Implementation Details

Training scheme For all GenTron model variations, we employ the AdamW [39] optimizer, maintaining a constant learning rate of 1×10^{-4}. We train our T2I GenTron models in a multi-stage procedure [47, 55] with an internal dataset, including a low-resolution (256×256) training with a batch size of 2048 and 500K optimization steps, as well as high-resolution (512×512) with a batch size of 784 and 300K steps. For the GenTron-G/2 model, we further integrate Fully Sharded Data Parallel (FSDP) [70] and activation checkpointing (AC), strategies specifically adopted to optimize GPU memory usage. In our video experiments, we train videos on a video dataset that comprises approximately 34M videos. To optimize storage usage and enhance data loading efficiency, the videos are pre-processed to a resolution with a short side of 512 pixels and a frame rate of 24 FPS. We process batches of 128 video clips. Each clip comprises 8 frames, captured at a sampling rate of 4 FPS.

Model	Attribute Binding			Obj. Relation		Comp.	**Mean**
	Color	Shape	Texture	Spat.	Non-spat.		
LDM v1.4	37.65	35.76	41.56	12.46	30.79	30.80	31.50
LDM v2	50.65	42.21	49.22	13.42	30.96	33.86	36.72
Composable v2	40.63	32.99	36.45	8.00	29.80	28.98	29.47
Structured v2	49.90	42.18	49.00	13.86	31.11	33.55	36.60
Attn-Exct v2	64.00	45.17	59.63	14.55	31.09	34.01	41.41
GORS	66.03	47.85	62.87	18.15	31.93	33.28	43.35
DALL·E 2	57.50	54.64	63.74	12.83	30.43	36.96	42.68
LDM XL	63.69	54.08	56.37	20.32	31.10	40.91	44.41
PixArt-α	68.86	55.82	70.44	20.82	31.79	41.17	48.15
GenTron	**76.74**	**57.00**	**71.50**	**20.98**	**32.02**	**41.67**	**49.99**

Table 3. **Comparison of alignment evaluation on T2I-CompBench [31].** Results show our advanced model, GenTron-CLIPT5XXL-G/2, achieves superior performance across multiple compositional metrics compared to previous methods.

Evaluation metrics. We mainly adopt the recent T2I-CompBench [31] to compare GenTron model variants, following [4, 10]. Specifically, we compare the attribute binding aspects, which include *color*, *shape*, and *texture*. We also compare the spatial and non-spatial object relationships. Moreover, user studies are conducted to compare visual quality and text alignment.

4.2. Main Results of GenTron-T2I

In this subsection, we discuss our experimental results, focusing on how various conditioning factors and model sizes impact GenTron's performance. Table 2 presents the quantitative findings of our study. Additionally, we offer a comparative visualization, illustrating the effects of each conditioning factor we explored. A comparison to prior art is also provided.

(a) adaLN-Zero (b) Cross attention

Figure 4. **adaLN-Zero *vs*. cross attention.** The prompt is *"A panda standing on a surfboard in the ocean in sunset."* Cross attention exhibits a distinct advantage in the *text*-conditioned scenario.

(a) GenTron-XL/2 (b) GenTron-G/2

Figure 5. **Effect of model scale.** The prompt is "a cat reading a newspaper". The larger model GenTron-G/2 excels in rendering finer details and rationalization in the layout of the cat and newspaper. More comparisons can be found in the appendix.

Method	#Param.	FID-30K↓	CLIP-Score↑	T2I-CompBen↑
Imagen	3.0B	**7.27**	0.27	-
Parti-750M	0.8B	10.71	-	-
Parti-3B	3.0B	8.10	-	-
GigaGAN	1.0B	9.09	0.322	-
MUSE-3B	3.0B	7.88	0.320	-
SD v1.4	0.9B	12.94	0.325	31.50
SDXL	2.6B	17.82	0.329	44.41
GenTron-XL/2	0.9B	14.21	0.326	49.13
GenTron-G/2	3.1B	14.53	**0.335**	**49.99**

Table 4. **Comparison to T2I models.**

Cross attention *vs*. adaLN-Zero. Recent findings [45] conclude that the adaLN design yields superior results in terms of the FID, outperforming both cross-attention and in-context conditioning in efficiency for *class*-based scenarios. However, our observations reveal a limitation of adaLN in handling free-form text conditioning. This shortcoming is evident in Figure 4, where adaLN's attempt to generate a panda image falls short, with cross-attention demonstrating a clear advantage. This is further verified quantitatively in the first two rows of Table 2, where cross-attention uniformly excels over adaLN in all evaluated metrics.

This outcome is reasonable considering the nature of *class* conditioning, which typically involves a limited set of fixed signals (*e.g.*, the 1000 one-hot class embeddings for ImageNet [18]). In such contexts, the adaLN approach, operating at a spatially global level, adjusts the image features uniformly across all positions through the normalization layer, making it adequate for the static signals. In contrast, cross-attention treats spatial positions with more granularity. It differentiates between various spatial locations by dynamically modulating image features based on the cross-attention map between the text embedding and the image features. This spatial-sensitive processing is essential for free-from *text* conditioning, where the conditioning signals are infinitely diverse and demand detailed representation in line with the specific content of textual descriptions.

Comparative analysis of text encoders. In Table 2 (rows two to four), we conduct a quantitative evaluation of various text encoders on T2I-CompBench, ensuring a fair comparison by maintaining a consistent XL/2 size across models. Results reveal that GenTron-T5XXL outperforms GenTron-CLIP-L across all three attribute binding and spatial relationship metrics, while it demonstrates comparable performance in the remaining two metrics. This suggests that T5 embeddings are superior in terms of compositional ability. These observations are in line with [2], which utilizes both CLIP and T5 embeddings for training but tests them individually or in combination during inference. Unlike eDiff, our approach maintains the same settings for both training and inference. Notably, GenTron demonstrates enhanced performance when combining CLIP-L and T5XXL embeddings, indicating the model's ability to leverage the distinct advantages of each text embedding type.

Scaling GenTron up. In Figure 5, we showcase examples from the PartiPrompts benchmark [68] to illustrate the qualitative enhancements achieved by scaling our model up from approximately 900 million to 3 billion parameters. Both models operate under the same CLIP-T5XXL condition. The larger GenTron-G/2 model excels in rendering finer details and more accurate representations, particularly in rationalizing the layout of objects like cats and newspapers. This results in image compositions that are both more coherent and more realistic. In comparison, the smaller GenTron-XL/2 model, while producing recognizable images with similar color schemes, falls short in terms of precision and visual appeal.

Furthermore, the superiority of GenTron-G/2 is quantitatively affirmed through its performance on T2I-CompBench, as detailed in Table 2. The increase in model size correlates with significant improvements across all evaluative criteria for object composition, including attributes and relationships. Additional comparative examples are provided in the appendix for further illustration.

Figure 6. **GenTron-T2V examples.** Prompts are "A giant tortoise is making its way across the beach" and "A dog swimming".

.

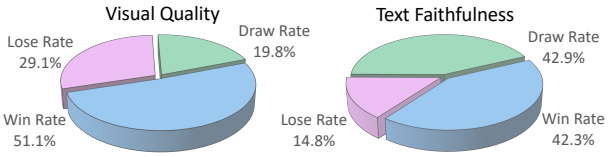


Figure 7. **Visualization of the human preference of our method *vs*. Latent Diffusion XL.** Our method received a significantly higher number of votes as the winner in comparisons of visual quality and text faithfulness, with a total of 3000 answers.

Comparison to prior work. In Table 3, we showcase the alignment evaluation results from T2I-CompBench. Our method demonstrates outstanding performance in all areas, including attribute binding, object relationships, and complex compositions. This indicates a heightened proficiency in compositional generation, with a notable strength in color binding. In this aspect, our approach surpasses the previous state-of-the-art (SoTA) work [10] by over 7%.

In Table 4, we further thoroughly compare the zero-shot FID-30K, CLIP-Score, and T2I-CompBench metrics with previous T2I models. In contrast to U-Net based SDv1.4, GenTron uses about *four times* ***less*** *data* (550M *vs*. 2B), while achieving better CLIP-scores and T2I-CompBench results. Although GenTron does not achieve a superior FID score compared to SDv1.4, it's important to note that recent studies, including Pick-a-Pic [37] and SDXL [47], have highlighted that FID scores often misrepresent human preferences in generative models, sometimes even **negatively** correlating with visual aesthetics.

User study Figure 7 shows the human preference of our method versus SDXL. We used standard prompts in PartiPrompt2 [68] to generate 100 images using both methods and ask people for their preference blindly after shuffling. We received a total of three thousand responses on the comparisons of visual quality and text faithfulness, with our method emerging as the clear winner by a clear margin.

4.3. GenTron-T2V Results

In Figure 6, We showcase several samples generated by GenTron-T2V, which are not only visually striking but also temporally coherent. This highlights GenTron-T2V's effectiveness in creating videos that are both aesthetically pleasing and consistent over time.

Figure 8. **Effect of motion-free guidance.** GenTron-T2V with motion-free guidance has a clear visual appearance improvement. The prompt is "A lion standing on a surfboard in the ocean in sunset".

Effect of motion-free guidance. In Figure 8, we present a comparison between our GenTron variants, with motion-free guidance (MFG) and without. For this comparison, critical factors such as the pre-trained T2I model, training data, and the number of training iterations were kept constant to ensure a fair evaluation. The results clearly indicate that GenTron-T2V, when integrated with MFG, shows a marked tendency to focus on the central object mentioned in the prompt, often rendering it in greater detail. Specifically, the object typically occupies a more prominent, central position in the generated video, thereby dominating the visual focus across video frames.

5. Conclusion

In this work, we provide a thorough exploration of transformer-based diffusion models for text-conditioned image and video generation. Our findings shed light on the properties of various conditioning approaches and offer compelling evidence of quality improvement when scaling up the model. A notable contribution of our work is the development of GenTron for video generation, where we introduce motion-free guidance. This innovative approach has demonstrably enhanced the visual quality of generated videos. We hope that our research will contribute to bridging the existing gap in applying transformers to diffusion models and their broader use in other domains.

Acknowledgement. This paper is partially supported by the National Key R&D Program of China No.2022ZD0161000 and the General Research Fund of Hong Kong No.17200622.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 5

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 4, 7, 12

[3] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, 2023. 3

[4] 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, 2023. 6, 12

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2, 3, 4, 5

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 5, 6

[7] 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. *OpenAI*, 2024. 2

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[9] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2, 3, 5

[10] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 3, 6, 8, 12

[11] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2

[13] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2, 4, 6, 12

[14] Yuren Cong, Mengmeng Xu, christian simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. FLATTEN: optical FLow-guided ATTENtion for consistent text-to-video editing. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[15] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 2, 4, 5, 12

[16] Yatin Dandi, Aniket Das, Soumye Singhal, Vinay Namboodiri, and Piyush Rai. Jointly trained image and video generation using residual vectors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3028–3042, 2020. 2, 5

[17] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023. 2

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 2

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 3

[21] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Pro-*

ceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12873–12883, 2021. 3

[22] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. *arXiv preprint arXiv:2303.14389*, 2023. 3

[23] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 4

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[25] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 3

[26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2, 5

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[28] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[29] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022. 4

[30] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 2, 4, 5

[31] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2, 6

[32] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 12

[33] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 4

[34] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 2

[35] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 2

[36] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 3

[37] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 8

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[40] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 3

[41] Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. Vdt: An empirical study on video diffusion with transformers. *arXiv preprint arXiv:2305.13311*, 2023. 3

[42] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023. 2

[43] 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. 2, 12

[44] Alexander Quinn 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. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 4, 12

[45] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 2, 3, 4, 7, 12

[46] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2, 3

[47] 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. 2, 4, 5, 6, 8, 12

[48] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023. 2

[49] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018. 2

[50] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4, 6, 12

[52] 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. 12

[53] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021. 2

[54] Alex Rogozhnikov. Einops: Clear and reliable tensor manipulations with einstein-like notation. In *International Conference on Learning Representations*, 2021. 5

[55] 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*, pages 10684–10695, 2022. 2, 3, 5, 6

[56] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2, 3

[57] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2, 12

[58] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 5

[59] Alex Shonenkov, Misha Konstantinov, Daria Bakshandaeva, Christoph Schuhmann, Ksenia Ivanova, and Nadiia Klokova. Deepfloyd if: A novel state-of-the-art open-source text-to-image model. https://github.com/deep-floyd/IF, 2023. 12

[60] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[61] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[62] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[64] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 3

[65] 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. 2, 5

[66] Yilin Wang, Zeyuan Chen, Liangjun Zhong, Zheng Ding, Zhizhou Sha, and Zhuowen Tu. Dolfin: Diffusion layout transformers without autoencoder. *arXiv preprint arXiv:2310.16305*, 2023. 3

[67] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. RAPHAEL: Text-to-image generation via large mixture of diffusion paths. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 5

[68] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. Featured Certification. 7, 8, 12

[69] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12104–12113, 2022. 2, 4

[70] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023. 6

[71] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023. 3

[72] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 3, 4

Method	Text Encoder	Integration
single text encoder		
GLIDE [43]	CLIP-B [51]	cross-attention
SDv1.4 / SDv1.5	CLIP-L [51]	cross-attention
SDv2.0 / SDv2.1	OpenCLIP-H [32]	cross-attention
DALL-E 2 [52]	CLIP [51]	cross-attention
DALL-E 3 [4]	T5-XXL [13]	cross-attention
Imagen [57]	T5-XXL [13]	cross-attention
DeepFloyd IF [59]	T5-XXL	cross-attention
DiT [45]	N/A	adaLN
PixArt-α [10]	T5-XXL [13]	cross-attention
multiole text encoders		
eDiff-I [2]	CLIP-L & T5-XXL	cross-attention
SDXL [47]	CLIP-L & OpenCLIP-bigG	cross-attention
Emu [15]	CLIP-L & T5-XXL	cross-attention

Table 5. Summary of conditioning strategies used by existing image diffusion models.

A. Summary of Conditioning Mechanism

In Table 5, we present a summary of the conditioning approaches utilized in existing text-to-image diffusion models. Pioneering studies, such as [44, 52], have leveraged CLIP's language model to guide text-based image generation. Furthermore, Saharia *et al*. [57] found that large, generic language models, pretrained solely on text, are adept at encoding text for image generation purposes. Additionally, more recently, there has been an emerging trend towards combining different language models to achieve more comprehensive guidance [2, 15, 47]. In this work, we use the interleaved cross-attention method for scenarios involving multiple text encoders, while reserving plain cross-attention for cases with a single text encoder. The interleaved cross-attention technique is a specialized adaptation of standard cross-attention, specifically engineered to facilitate the integration of two distinct types of textual embeddings. This method upholds the fundamental structure of traditional cross-attention, yet distinctively alternates between different text embeddings in a sequential order. For example, in one transformer block, our approach might employ CLIP embeddings, and then in the subsequent block, it would switch to using Flan-T5 embeddings.

B. More Results

B.1. Additional Model Scaling-up Examples

We present additional qualitative results of model scaling up in Figure 9. All prompts are from the PartPrompt [68].

C. Additional GenTron-T2I Examples

We present more GenTron-T2I example in Figure 10.

Figure 9. **More examples of model scaling-up effects.** Both models use the CLIP-T5XXL conditioning strategy. Captions are from PartiPrompt [68].

D. Additional GenTron-T2V Examples

Additional GenTron-T2V results are available on our website[1].

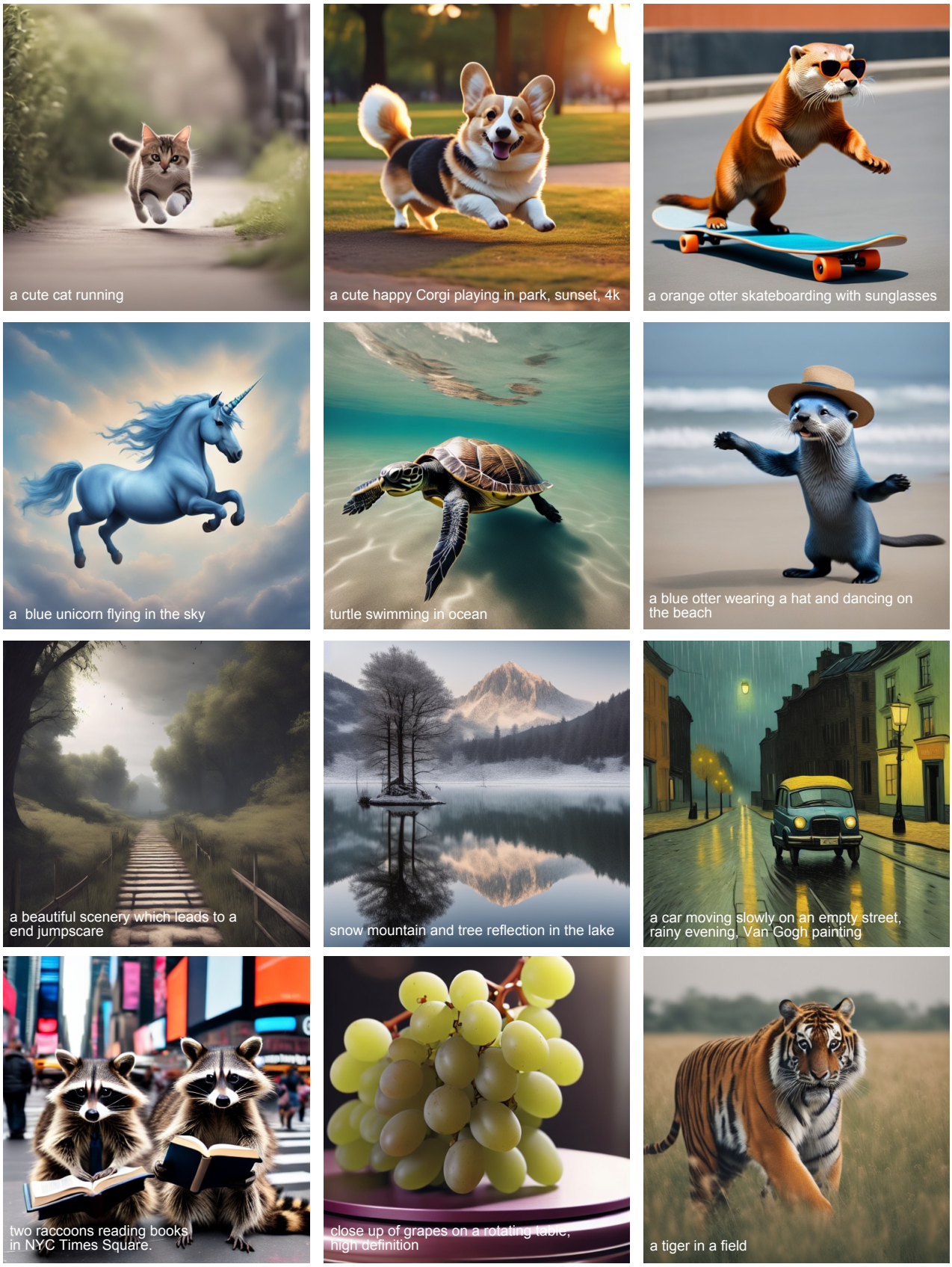


Figure 10. **GenTron-T2I examples.**