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.

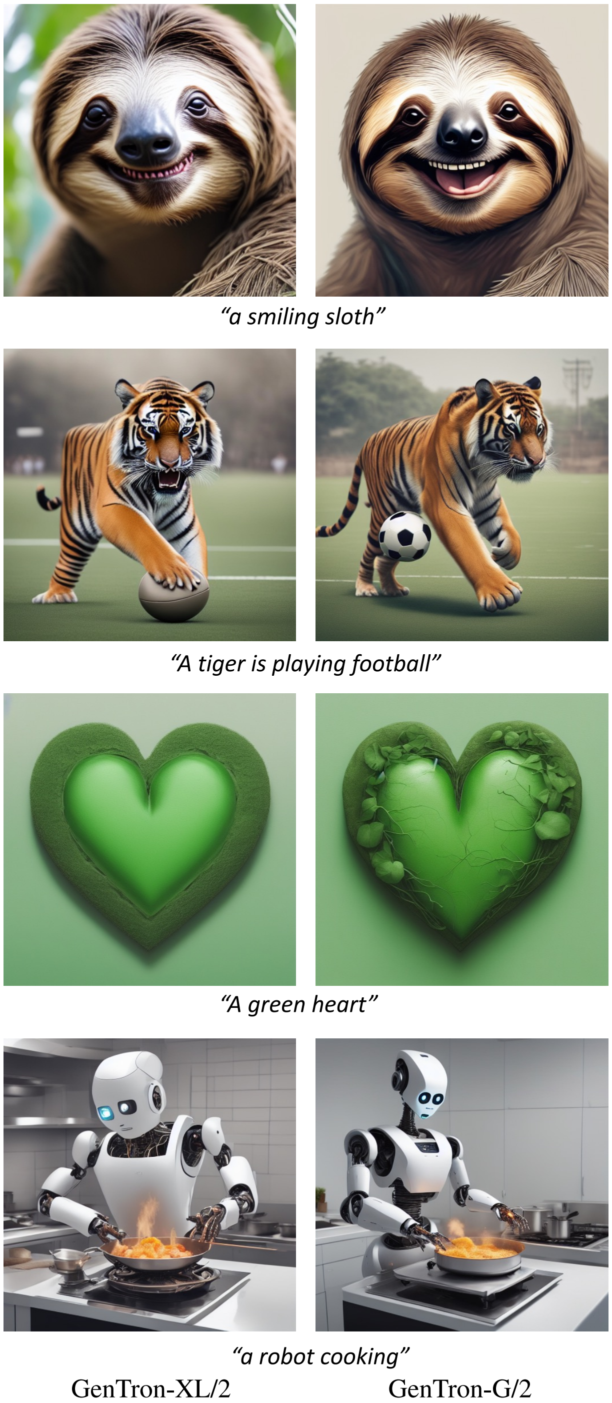
B. More Results

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.



"a smiling sloth"

"A tiger is playing football"

"A green heart"

"a robot cooking"

GenTron-XL/2 GenTron-G/2

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].

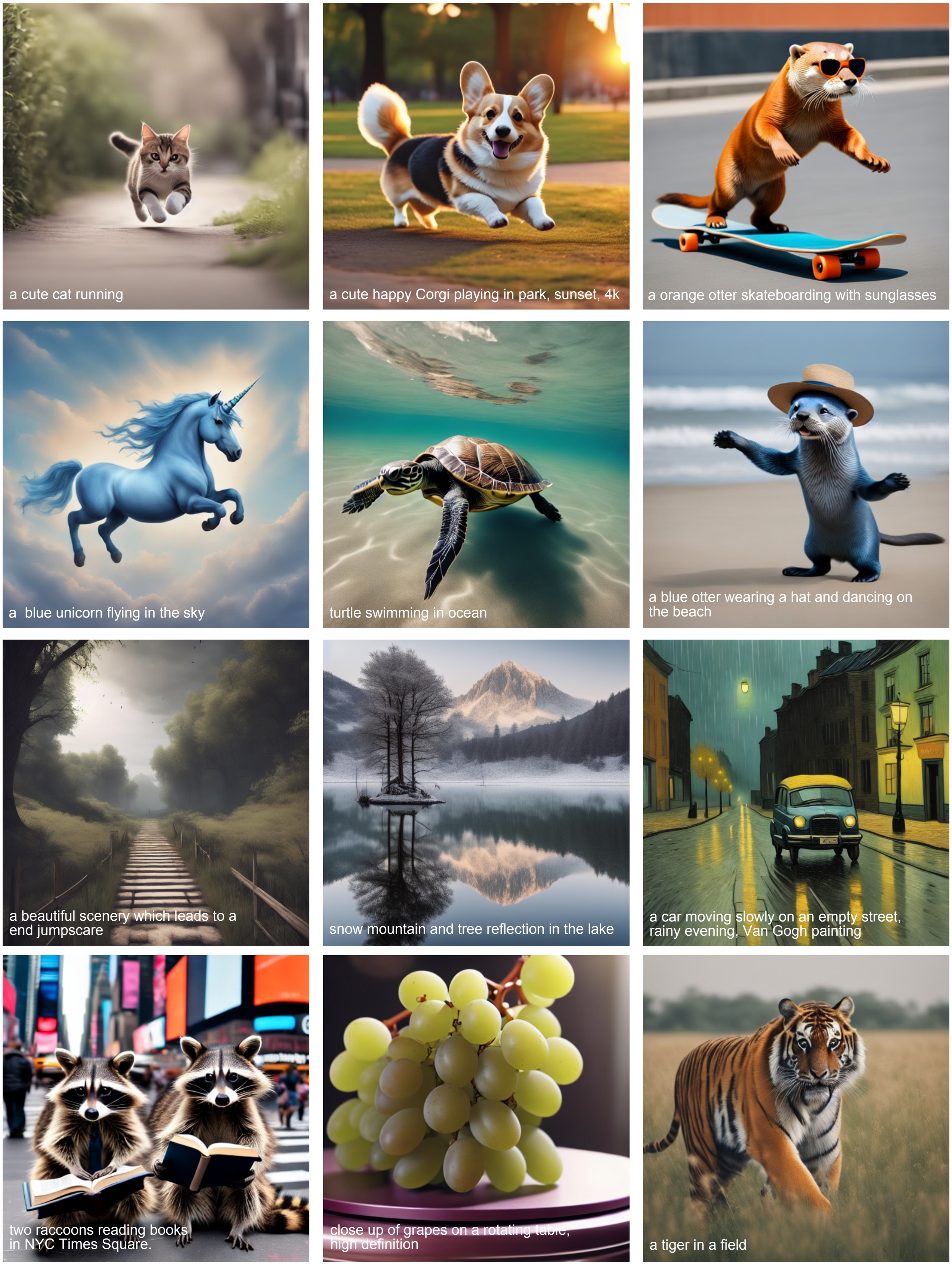

a cute cat running
a cute happy Corgi playing in park, sunset, 4k
a orange otter skateboarding with sunglasses
a blue unicorn flying in the sky
turtle swimming in ocean
a blue otter wearing a hat and dancing on the beach
a beautiful scenery which leads to a end jumpscare
snow mountain and tree reflection in the lake
a car moving slowly on an empty street, rainy evening, Van Gogh painting
two raccoons reading books in NYC Times Square.
close up of grapes on a rotating table, high definition
a tiger in a field

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