A Additional Results

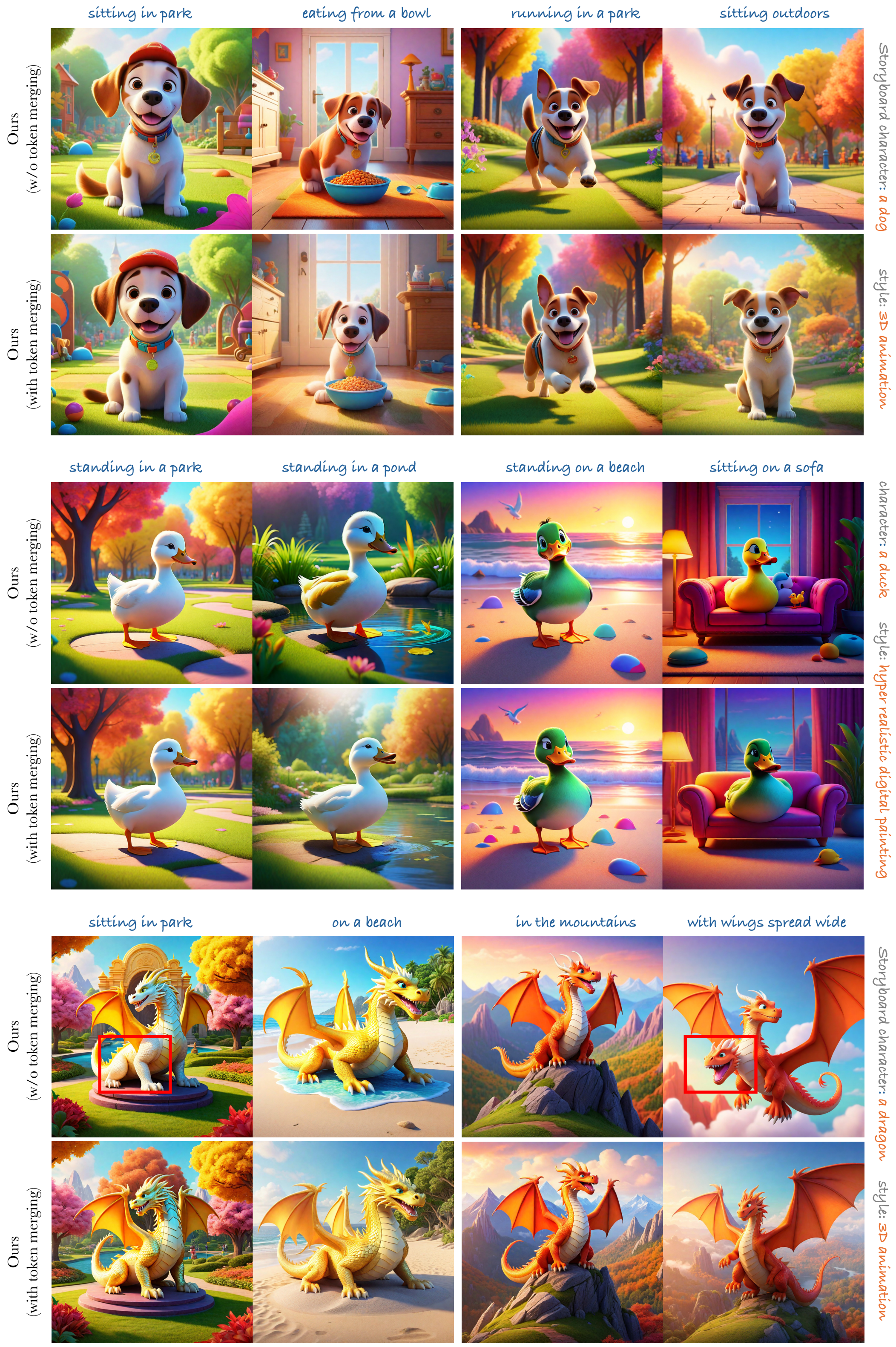


Figure 10: *Additional results with and without token merging for aligning fine-grain features.*

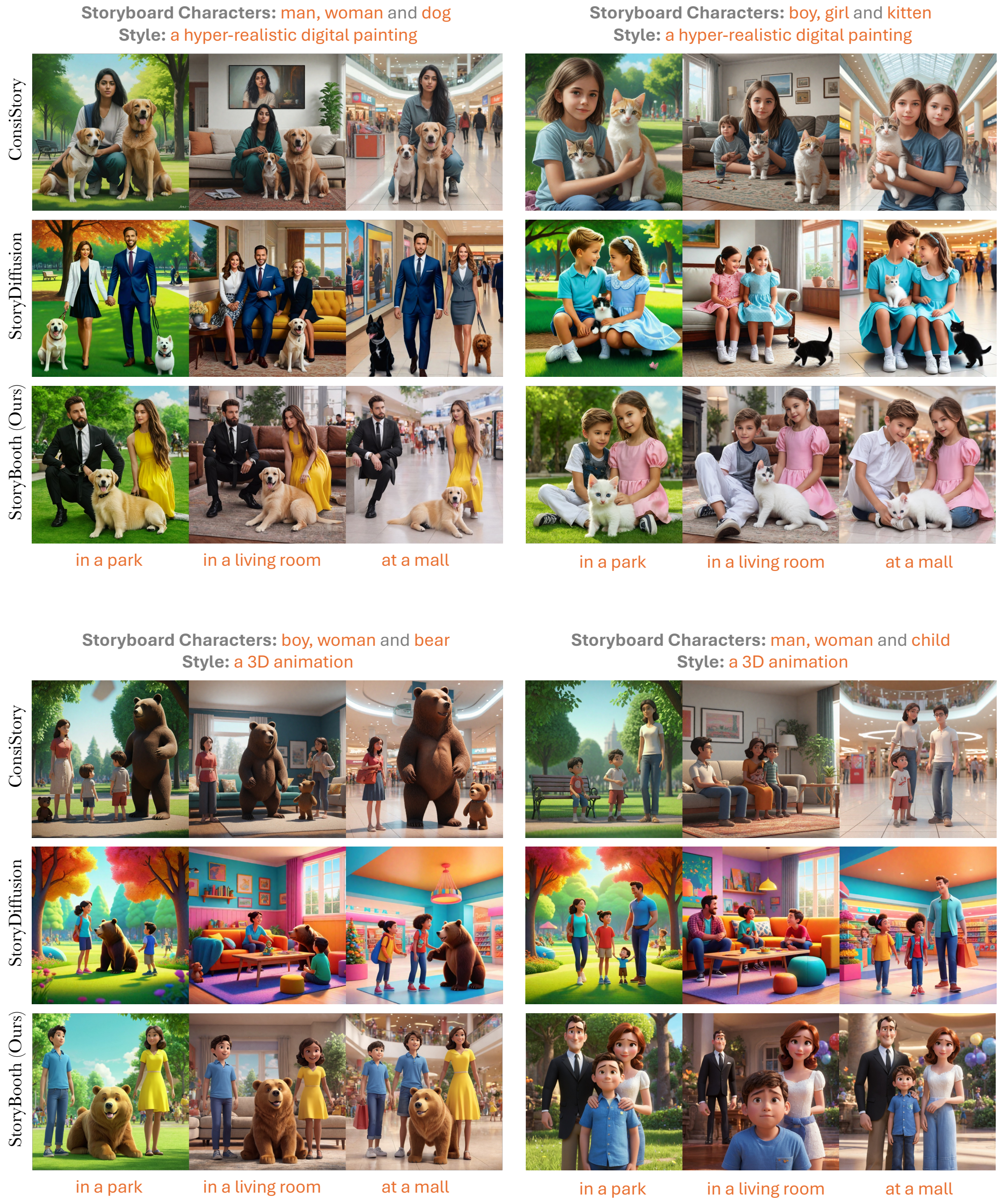


Figure 11: **Additional Results for $N > 2$ characters.**

B Implementation Details

In this section, we provide further details for the implementation of our approach as well as other baselines (Ruiz et al., 2022; Zhou et al., 2024) used while reporting results in the main paper (Sec. 5).

Model Details. Similar to (Zhou et al., 2024; Tewel et al., 2024), we use the official SDXL model Podell et al. (2023) as the underlying pretrained text-to-image generation model while reporting results with all methods (Zhou et al., 2024; Ruiz et al., 2022; Gal et al., 2022; Ye et al., 2023) (including *storybooth* as proposed in the main paper). Results with BLIP-Diffusion (Li et al., 2024a) and Storygen (Liu et al., 2024) are reported directly using pretrained models obtained from paperauthors. Similar to (Tewel et al., 2024), Dreambooth (Ruiz et al., 2022) training is done using low-rank (LoRA) finetuning (Hu et al., 2021) with rank of 4.

All results are reported at 1024×1024 resolution while using 50 inference steps during the reverse diffusion process. Unless otherwise specified, a fixed classifier-free guidance scale (Ho & Salimans, 2022) $\alpha_{cfg} = 5.0$ is used for all experiments. A LLaMA3-8B model AI@Meta (2024) is used as the underlying large language model for generating storyboard layouts from the storyboard prompt $\mathcal{P}$.

Region-Based Storyboard Generation. A key idea behind our approach is to use region-based planning and generation Yang et al. (2024a) in order to *apriori* localize the placement of different characters across the storyboard. This helps us accurately apply cross-frame self-attention bounding (refer Sec. 4) in order to allow tokens for each subject to pay attention to *only* tokens from the same subject. However, this requires the generated images to align with the initially predicted character layouts across the storyboard. While several region-based solutions for the same are feasible (Dahary et al., 2024; Li et al., 2023), for simplicity we use the region-based cross-attention masking from (Yang et al., 2024a) for ensuring adherence to the input prompt. Equal distribution of weights (Yang et al., 2024a) are used for all character level τ_i^k and overall-frame level storyboard prompts $\mathcal{T}_i$.

Bounded Cross-frame Self-Attention. We apply bounded self-attention (refer Sec. 4) both within the same frame as well as across different frames in order to reduce inter-character leakage. The self-attention bounding is applied predominantly on the *up-blocks* of the SDXL UNet (Podell et al., 2023) and is used between timesteps $t \in [1000, 200]$. More importantly, we observe that the naive use of self-attention bounding alone can reduce output image quality. We therefore utilize a dropout $\beta_d = 0.5$ which allows reduces intercharacter leakage while still preserving output image quaility.

Token Merging. In order to better align the fine-grain features of different subjects we use token merging (refer Sec. 4.3) in order to place a hard-constraint on the appearance of the character features across different frames. We use a positive $\alpha = 0.4$ for token merging from timesteps $t \in [950, 600]$. Furthermore, we also use *early negative token unmerging* (refer Sec. 4.4) with $\alpha = -0.5$ from timesteps $t \in [1000, 950]$ in order to encourage pose-variance (refer Fig. 6).

C Evaluation and User Study

Datasets. Given the training-free nature of our proposed approach, we do not rely on the use of any public datasets. For evaluation purposes, we utilize the storyboard-prompt dataset from (Tewel et al., 2024) which consists of 100 storyboard generation prompts with a single main subject across diverse settings. For multi-subject evaluation, we construct an analogoes multi-subject prompt dataset using LLaMA3-8B (AI@Meta, 2024), where given a set of potential subjects (*e.g.*, *cat, dog, hedgehog, owl, bear, duck etc*.) and the possible scene locations (*e.g.*, *mountains, park, beach etc*.), we prompt the language model $\mathcal{M}$ to generate a dataset of storyboard prompts $\mathcal{D} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots \mathcal{P}_M\}$, (where $M = 100$) placing two randomly selected subjects in different settings.

Evaluation Metrics. We also evaluate the performance of the proposed approach quantitatively, by evaluating the text-to-image alignment and output character consistency across different storyboard frames (refer Sec. 5 of main paper). In particular, we use the recently proposed VQAScore (Lin et al., 2024) for evaluating text-to-image alignment, as it has been observed to show significantly higher-correlation with the human preferences over traditionally used CLIP-Score (Hessel et al., 2021) especially when scaling to multiple characters (Lin et al., 2024). Also, similar to Tewel et al. (2024), Dreamsim (Fu et al., 2023) cosine similarity is used for evaluating character consistency, as it is observed to show higher correlation with human-evaluation for image-to-image similarity as opposed to traditionally used CLIP-I (Radford et al., 2021) and DINO (Oquab et al., 2023) scores.

Human-user Study. In addition to quantitative evaluation (refer Table 1), we also perform an anonymous human user study wherein the T2I alignment (Lin et al., 2024) and character-consistency (Fu et al., 2023) are evaluated by actual human users (refer Tab. 2 for results). In particular, given a storyboard prompt $\mathcal{P}$ and image-level prompts $\{\mathcal{T}_1, \mathcal{T}_2, \ldots \mathcal{T}_N\}$, the participants are shown a pair of storyboard outputs comparing our method with prior works. The user study consists of two separate tasks 1) Evaluating T2I Alignment and 2) Evaluating character consistency.

For evaluating text-to-image alignment, the participants are shown a pair of output images and the input prompt, and asked to select the image with the best alignment between the output image and input text prompt. Similarly for evaluating character-consistency, given a set of desired storyboard characters, the participants are shown a pair of storyboard outputs and asked to select the one with the better consistency for all storyboard characters. Empirically we found that unlike single-character consistency, judging cross-frame consistency for multiple characters is significantly harder even for human annotators when using $N > 2$ frames. We therefore use $N = 2$ frames when evaluating multi-subject consistency using human evaluation in order to get better quality annotations. Additionally, in order to remove data noise, we use a repeated comparison (control seed) for each user. Responses of users who answer differently to this repeated seed are discarded while reporting the final results. Please refer Fig. 18 for a screenshot of the quantitative human user-study setup for both tasks.

D Discussion and Limitations

While the proposed approach helps improve both text-to-image alignment as well as character-consistency when scaling to multiple characters, it still has some limitations. *First,* the proposed cross-frame self-attention bounding approach relies on the use of cross-attention masking drive region-based generation Yang et al. (2024a). Thus, weaknesses of underlying region-based generation approach can sometimes become our weaknesses. Recall that region-based storyboard generation helps *apriori* localize the placements of different characters and is used for reducing inter-character leakage (refer Sec. 4). While the use of self-attention bounding further helps improve the layout consistency in the storyboard frames (refer Fig. 3 from main paper), it may still struggle in scenarios where the underlying cross-attention driven region-based generation shows poor performance. In future, the use of more advanced or off-the-shelf pretrained region-based generation models (Li et al., 2023; Dahary et al., 2024) can help consistensy with the predicted storyboard layouts.

Second, we note that while negative token unmerging helps increase pose-variance, the output results may sometimes exhibit similar poses in the lack of defining action phrases in the input prompt (*e.g*. running, sitting, standing *etc*.). Nevertheless, we note that explicitly prompting the large-language model $\mathcal{M}$ to describe the character activity in each frame can help alleviate this problem.

Finally, as noted in Tab. 1 of the main paper, while the proposed approach helps achieve better T2I alignment and character-consistency over prior works, the prompt-alignment performance decreases when scaling to multiple characters. In particular, we observe a decline in T2I alignment score for our approach from 0.78 to 0.63 when scaling to multiple characters (prior training-free *state-of-the-art storydiffusion* Zhou et al. (2024) declines to 0.407). This, leaves much room for improvement of consistent multi-character storyboard generation and storytelling. Combining the proposed training-free approach (Sec. 4) with selective fine-tuning the cross-frame self-attention (Guo et al., 2024) using multi-character data presents as interesting direction for future work. However, the same is out of scope of this paper, and we leave it here as direction for future research.

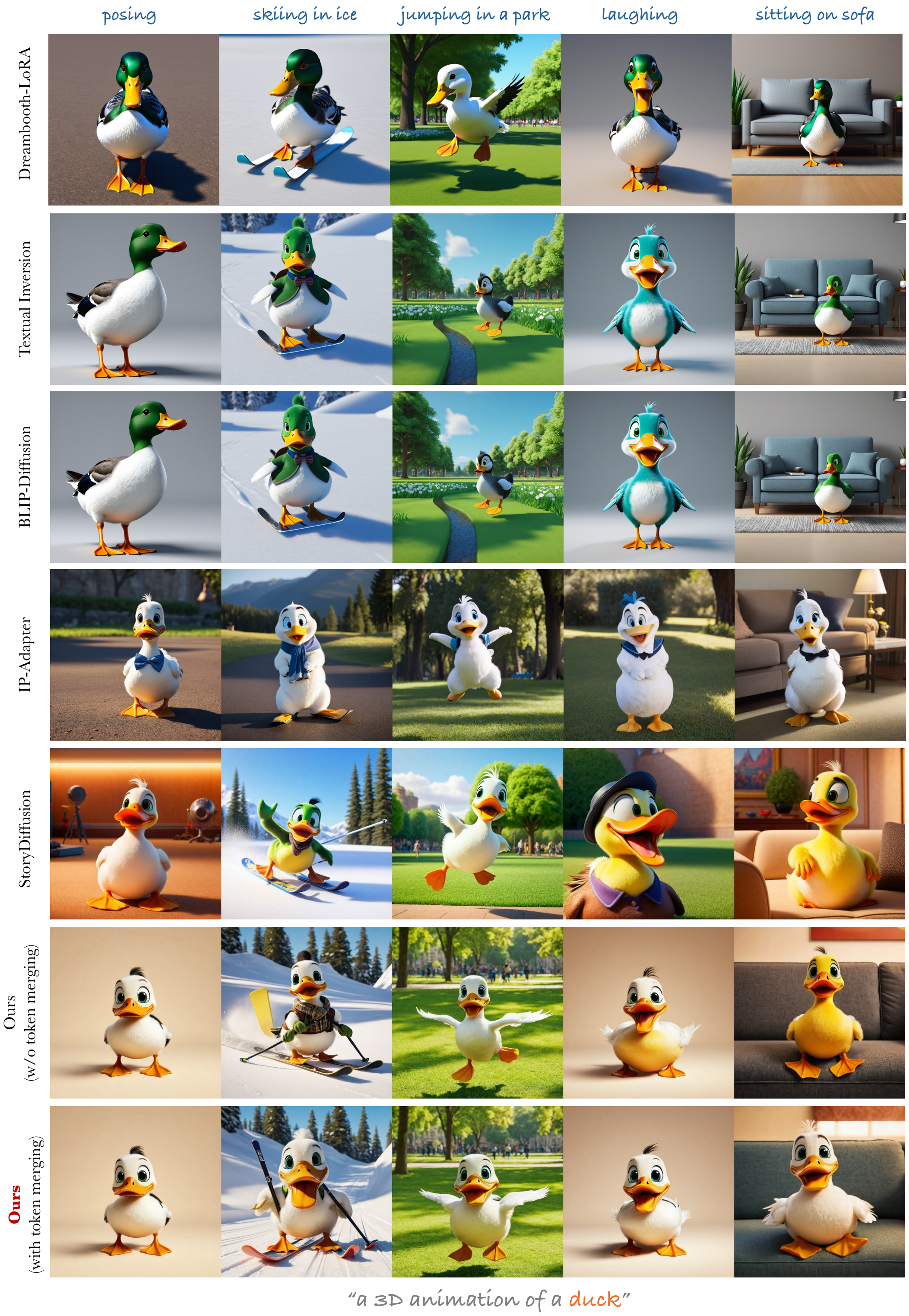


Figure 12: *Additional Comparisons:* comparing our approach with prior works (refer Sec. 5)

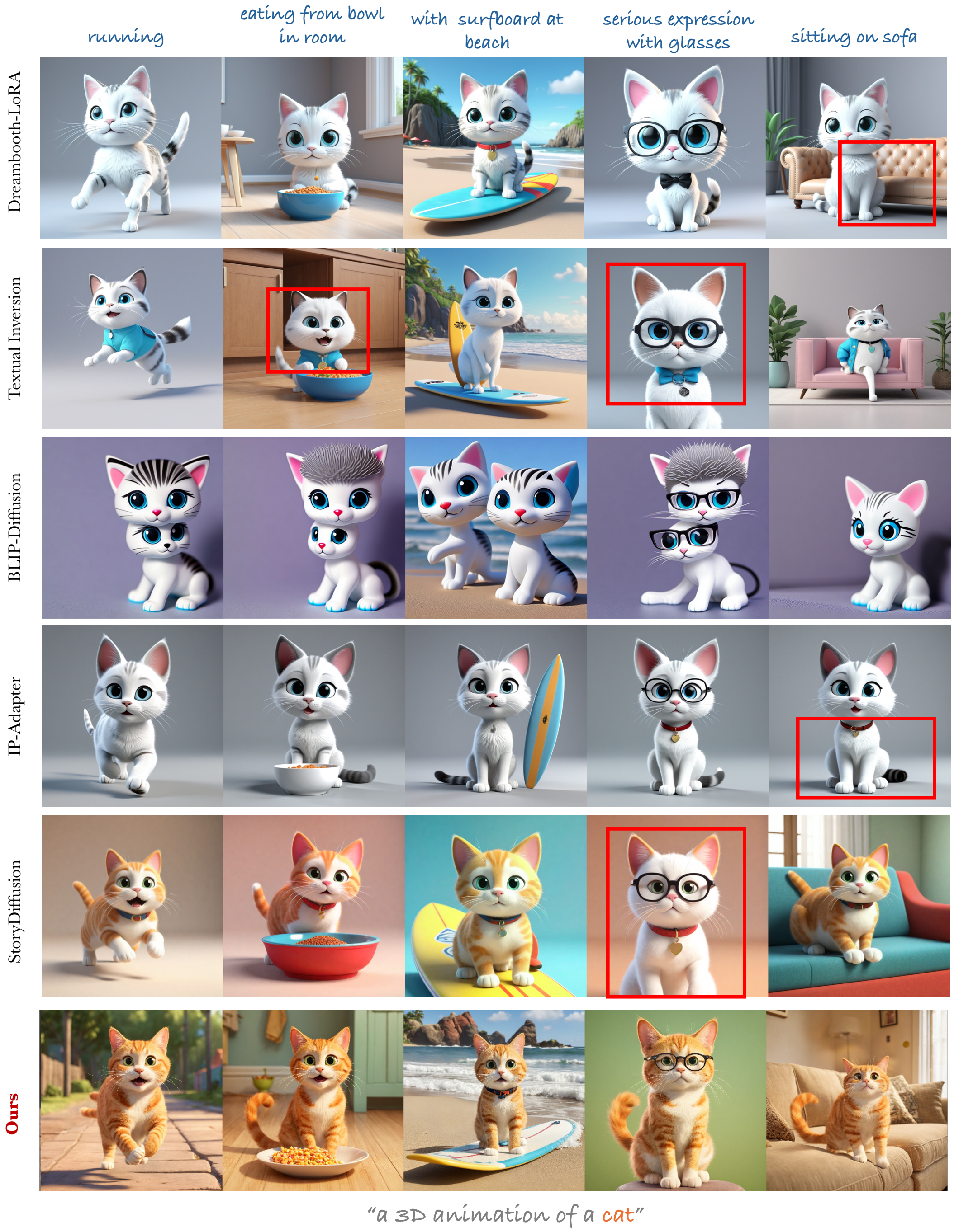


Figure 13: *Additional Comparisons:* comparing our approach with prior works (refer Sec. 5)

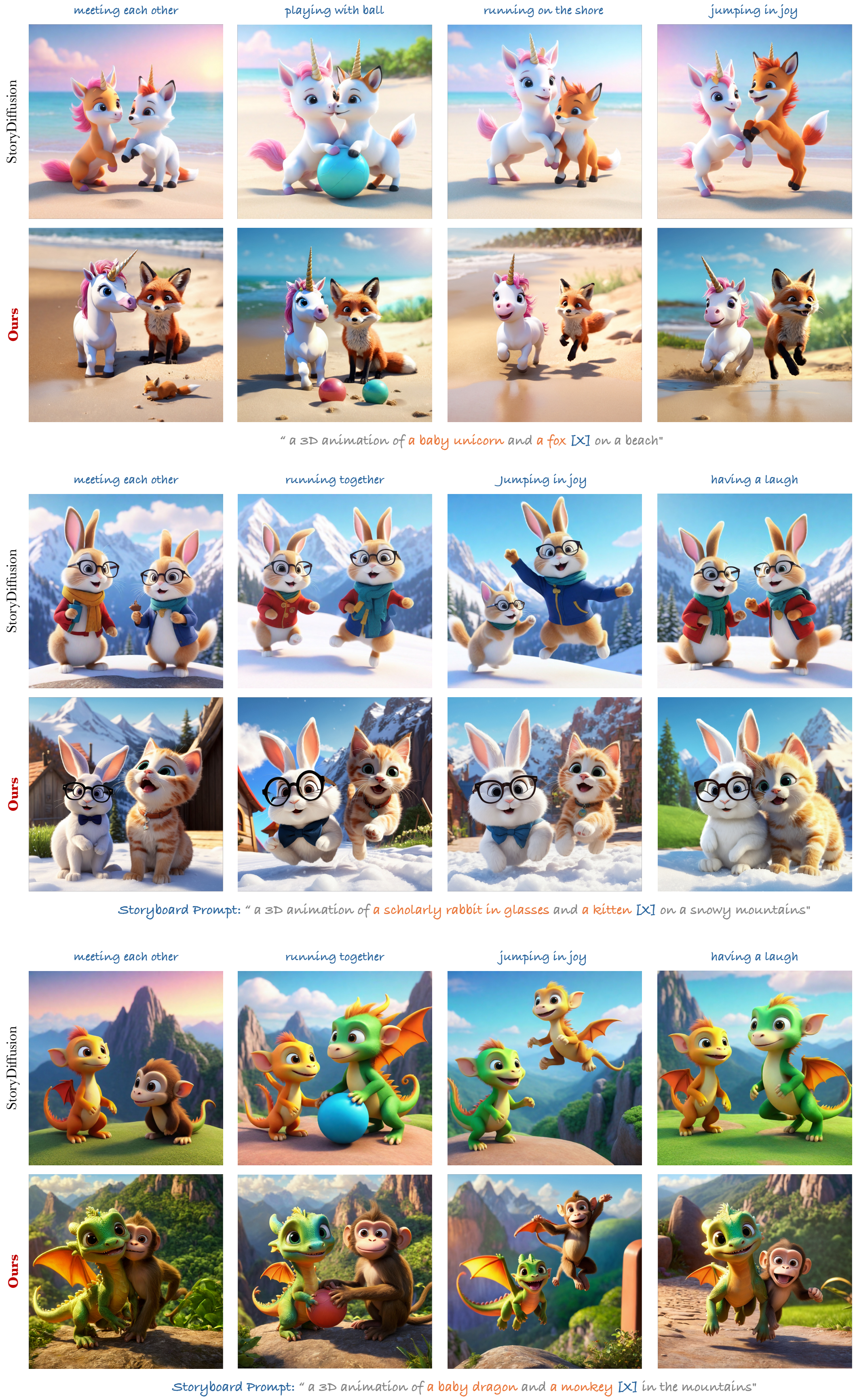
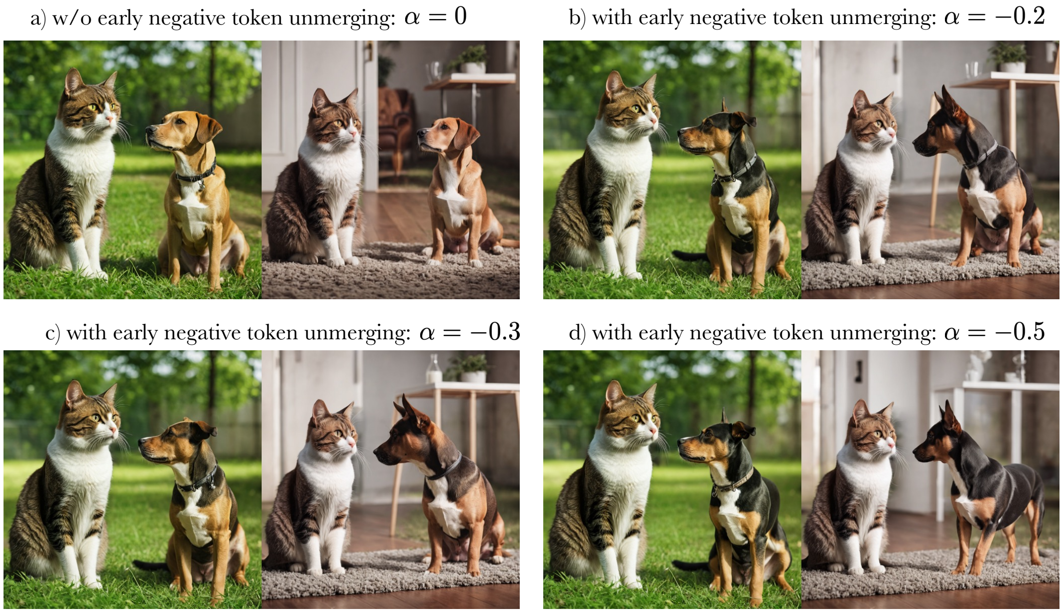

meeting each other
playing with ball
running on the shore
jumping in joy
StoryDiffusion
Ours
" a 3D animation of a baby unicorn and a fox [X] on a beach"
meeting each other
running together
Jumping in joy
having a laugh
StoryDiffusion
Ours
Storyboard Prompt: " a 3D animation of a scholarly rabbit in glasses and a kitten [X] on a snowy mountains"
meeting each other
running together
jumping in joy
having a laugh
StoryDiffusion
Ours
Storyboard Prompt: " a 3D animation of a baby dragon and a monkey [X] in the mountains"

Figure 14: *Additional results for multi-character storyboard generation.*



a) w/o early negative token unmerging: $\alpha = 0$

b) with early negative token unmerging: $\alpha = -0.2$

c) with early negative token unmerging: $\alpha = -0.3$

d) with early negative token unmerging: $\alpha = -0.5$

"a photo of a cat and dog [in a park, in a living room] "

Figure 15: *Visualizing pose-variation with negative-token merging coefficient*. In order to visualize the effect of $\alpha < 0$ for early negative token merging we generate a storyboard with a *a cat and dog* in discrete settings (*park, living room*) without specifying the action (*e.g*., sitting, standing *etc*.). We then increase the negative-token merging coefficient $\alpha < 0$. We observe that without negative token unmerging the pose of both cat and dog is very similar. As the value of α is gradually varied the pose variance between the subjects increase with the *dog* appearing to gradually turn to a standing position, while still maintaining consistency for both storyboard characters (*cat and dog*).

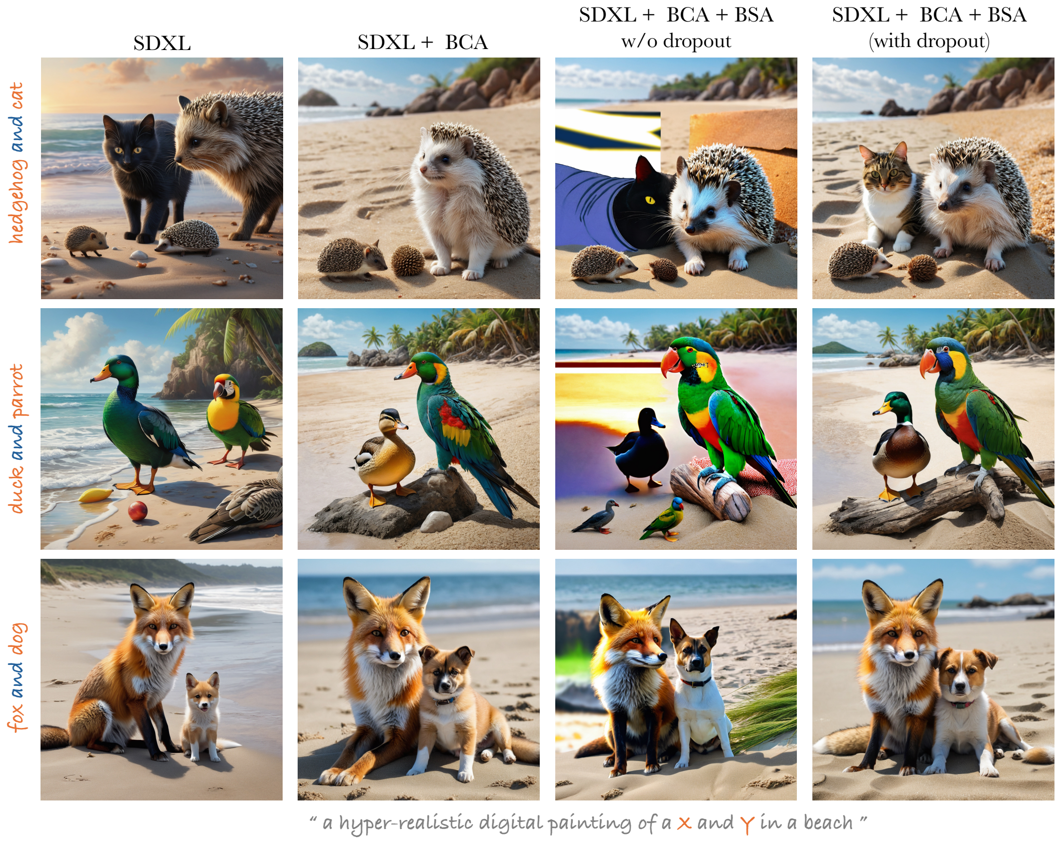
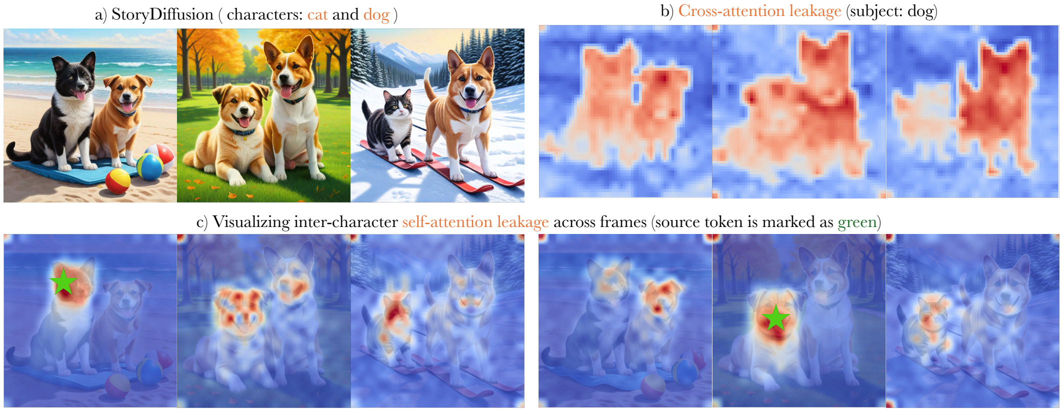

SDXL
SDXL + BCA
SDXL + BCA + BSA w/o dropout
SDXL + BCA + BSA (with dropout)
hedgehog and cat
duck and parrot
fox and dog
" a hyper-realistic digital painting of a X and Y in a beach "

Figure 16: *Role of dropout in preserving image quality after self-attention bounding.* We observe that bounded cross-attention (BCA) computation (Yang et al., 2024a) itself can often be insufficient for removing inter-character leakage within the same frame. The *bounded self-attention* (BSA) as proposed in the paper (Sec. 4) can help reduce inter-character leakage but its naive application can result in reduction of image quality (Col-3). Therefore, despite its simplicity dropout pays a critical role in the proposed *bounded self-attention* (BSA) approach, and helps maintain output image quality while still reducing inter-character leakage. ***Note:*** Please note that *bounded self-attention* (BSA) as proposed in the paper (Sec. 4) also helps cross-frame inter-character leakage, however, reduction in cross-frame leakage is often not feasible unless inter-character leakage within the same frame is minimized first. Thus, the proposed approach helps address both problems in a joint formulation.

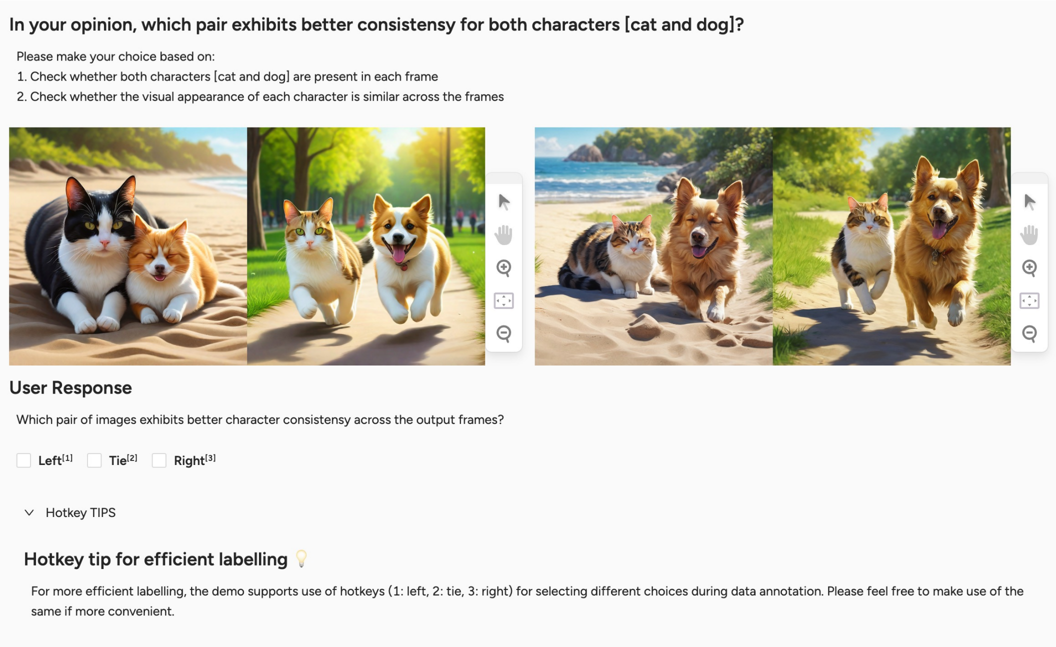


a) StoryDiffusion (characters: cat and dog)

b) Cross-attention leakage (subject: dog)

c) Visualizing inter-character self-attention leakage across frames (source token is marked as green)

Figure 17: *Visualizing cross-frame self-attention leakage.* We observe that sharing self-attention across different frames (Zhou et al., 2024) exacerbates inter-character leakage with tokens of one subject (*dog*) paying attention to tokens of the other subject (*cat*) *and vice-versa*, across different frames. Moreover, addressing this leakage using cross-attention masking (Tewel et al., 2024) is also not feasible due to the correspondingly increased cross-attention leakage (b) between output characters.



In your opinion, which pair exhibits better consistensy for both characters [cat and dog]?

Please make your choice based on:
1. Check whether both characters [cat and dog] are present in each frame
2. Check whether the visual appearance of each character is similar across the frames

User Response

Which pair of images exhibits better character consistensy across the output frames?

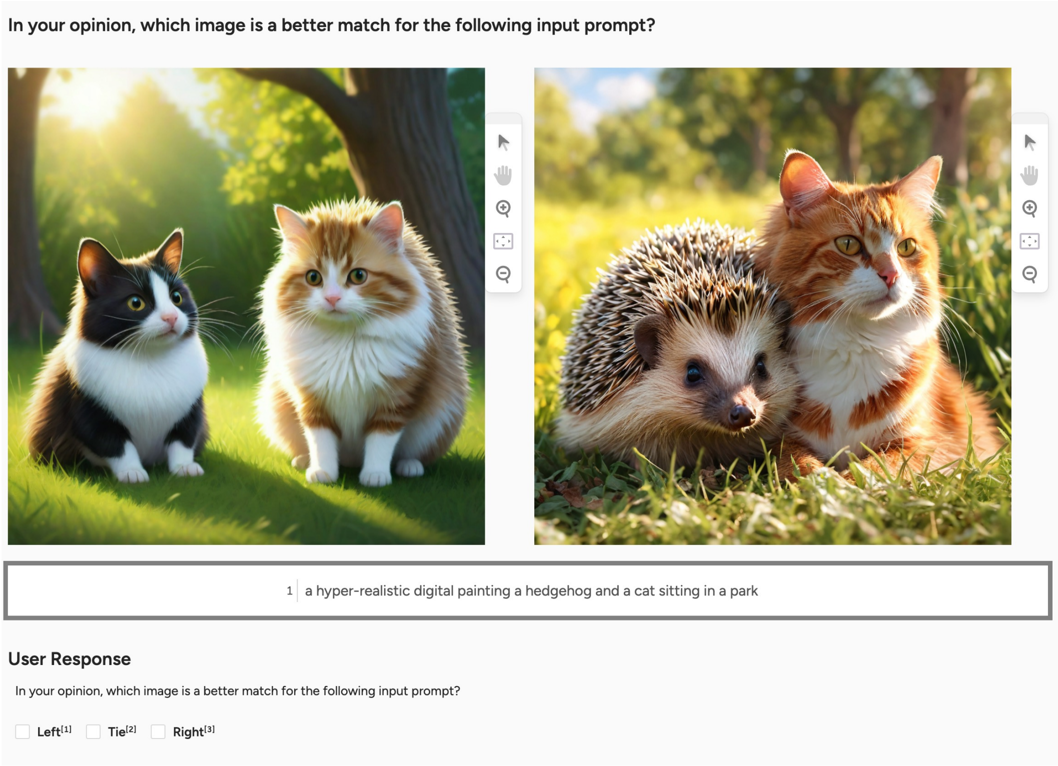
□ Left[1] □ Tie[2] □ Right[3]

Hotkey TIPS

Hotkey tip for efficient labelling 💡

For more efficient labelling, the demo supports use of hotkeys (1: left, 2: tie, 3: right) for selecting different choices during data annotation. Please feel free to make use of the same if more convenient.

(a) *Setup for user-study evaluating multi-character consistency (refer Sec. 5)*



In your opinion, which image is a better match for the following input prompt?

1 a hyper-realistic digital painting a hedgehog and a cat sitting in a park

User Response

In your opinion, which image is a better match for the following input prompt?

□ Left[1] □ Tie[2] □ Right[3]

(b) *Setup for user-study evaluating multi-character consistency (refer Sec. 5).*

Figure 18: *Setup for user-study comparing our method with prior works (refer Sec. 5)*

References

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md. 16

Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. The chosen one: Consistent characters in text-to-image diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024. 1, 3

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 3

Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be yourself: Bounded attention for multi-subject text-to-image generation. *arXiv preprint arXiv:2403.16990*, 2024. 1, 5, 16, 17

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

Zhangyin Feng, Yuchen Ren, Xinmiao Yu, Xiaocheng Feng, Duyu Tang, Shuming Shi, and Bing Qin. Improved visual story generation with adaptive context modeling. *arXiv preprint arXiv:2305.16811*, 2023. 1, 3

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 9, 16

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1, 3, 7, 9, 10, 15

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. 17

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 16

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 16

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 16

Hyeonho Jeong, Gihyun Kwon, and Jong Chul Ye. Zero-shot generation of coherent storybook from plain text story using diffusion models. *arXiv preprint arXiv:2302.03900*, 2023. 1, 3

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024a. 1, 3, 7, 9, 10, 15

Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7486–7495, 2024b. 3, 7

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22511–22521, 2023. 3, 16, 17

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. 9, 16

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm-open-ended visual storytelling via latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6190–6200, 2024. 1, 2, 3, 7, 9, 10, 15

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 16

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 1

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. 7, 9, 15, 16

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*, pp. 8748–8763. PMLR, 2021. 16

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 1, 2, 3, 7, 9, 10, 15, 16

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4): 1–18, 2024. 1, 2, 3, 7, 9, 15, 16, 23

Khai N Truong, Gillian R Hayes, and Gregory D Abowd. Storyboarding: an empirical determination of best practices and effective guidelines. In *Proceedings of the 6th conference on Designing Interactive systems*, pp. 12–21, 2006. 1

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 3

Bichen Wu, Ching-Yao Chuang, Xiaoyan Wang, Yichen Jia, Kapil Krishnakumar, Tong Xiao, Feng Liang, Licheng Yu, and Peter Vajda. Fairy: Fast parallelized instruction-guided video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8261–8270, 2024. 3

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and CUI Bin. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*, 2024a. 2, 3, 4, 5, 16, 17, 22

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*, 2024b. 1, 2, 3, 7, 9

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 7, 8, 9, 10, 15

Zhengqing Yuan, Ruoxi Chen, Zhaoxu Li, Haolong Jia, Lifang He, Chi Wang, and Lichao Sun. Mora: Enabling generalist video generation via a multi-agent framework. *arXiv preprint arXiv:2403.13248*, 2024. 3

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 4

Canyu Zhao, Mingyu Liu, Wen Wang, Jianlong Yuan, Hao Chen, Bo Zhang, and Chunhua Shen. Moviedreamer: Hierarchical generation for coherent long visual sequence, 2024. URL https://arxiv.org/abs/2407.16655. 1, 3

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024. 1, 2, 3, 4, 7, 9, 10, 15, 17, 23