

Supplementary Materials: Customizing Text-to-Image Generation with Inverted Interaction

Anonymous Authors

1 IMPLEMENTATION DETAILS

1.1 Benchmark Details

A detailed illustration of the 15 interactive poses introduced in this paper is presented in Fig. 1.

Training: For each of these interactions, we collect 10 diverse exemplar images by searching with a set of keywords on the internet.

Evaluation: We use 200 descriptions to generate 2,000 images, where each description generates 10 images. The used description consists of the learned tokens $[P_1^*]$, $[P_2^*]$ and $[R^*]$ for the 15 interactive poses and diverse subjects to fully examine the generalization of our method. A total of 30 kinds of subject categories are used in the description, such as ‘seal’, ‘sloth’, ‘kangaroo’ and ‘frog’.

1.2 List of the Verb and Preposition

This paper employs two kinds of language priors (verb and preposition) to promote the learning of token embeddings P_1^* , P_2^* and R^* . As for computing the verb prior, a list of verbs corresponding to commonly used action are shown in Tab. 1. As for computing the preposition prior, we collect several prepositions, shown in Tab. 2.

1.3 Training Details

Here, we present more training details of the proposed two-stage inversion framework. In this paper, we mainly focus on the interaction of two subjects, which is more common in daily life.

In the single pose inversion stage, we employ ViTPose[8] to detect the pose of subjects O_1 and O_2 in the exemplar images. Then, we utilize ControlNet conditional on each pose to generate 5 images with different backgrounds and styles, with each image containing only a single subject while maintaining its pose as the original exemplar image. To avoid the model generating the subject with deformed limbs, we perform a closed-loop detector. Specifically, we also use ViTPose to detect the pose for the generated image and filter out the image where the pose of the subject is highly different from the exemplar image. The learnable tokens of pose $[P_1^*]$ and $[P_2^*]$ are separately optimized with 500 iterations with Adamw[5] optimizer by setting $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Each mini-batch consists of 8 samples. The learning rate is set to 2×10^{-4} .

In the interactive pose inversion stage, the original exemplar images are used. We set loss weight $\lambda = 0.01$ for \mathcal{L}_{CA} and optimize the interactive pose token $[R^*]$ with 1000 iterations using Adamw[5] optimizer by setting $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Each mini-batch consists of 8 samples. The learning rate is set to 2×10^{-4} .

We finetune token embedding in the CLIP Text Encoder based on Stable Diffusion V1-5¹.

1.4 Detailed Explanation of Evaluation Metrics

To evaluate pose consistency between interactive poses of generated images and given exemplar images, we propose Pose-S and Pose-KP

Table 1: List of the Verbs for computing the verb prior.

attack	act	box	blow	bounce	brush	carry	catch
chase	cheer	clap	climb	come	cough	crawl	cross
cut	cycle	dine	dive	drink	drop	eat	fall
feed	flick	fly	fold	follow	give	glide	go
grab	growl	guide	hang	hike	hit	hold	hop
hug	hush	juggle	jump	kick	kiss	kneel	knock
lie	lift	look	make	move	nod	open	pat
pick	play	point	pounce	pull	punch	push	put
reach	read	ride	rub	run	salute	shake	shoot
sit	slap	snap	sneeze	sniff	squat	stagger	stand
staple	step	stretch	swing	take	talk	tear	throw
toss	touch	use	walk	warp	wave	wear	whisper
wield	wipe	write	yawn				

Table 2: List of the prepositions for computing the preposition prior.

aboard	about	above	across	after	against	along	among
amongst	astride	at	atop	before	behind	below	beneath
beside	between	beyond	by	down	from	in	including
inside	into	near	of	off	on	onto	outside
over	through	to	toward	under	up	upon	versus
with	within						

metrics based on classifier accuracy. Specifically, we construct a 15-class image classification dataset using exemplar images. We utilize data augmentation strategy including random crop and horizontal flip, then split the training-testing set into a 4:1 ratio. In the Pose-S, we extract pose-related features from the generated images using the encoder of a ViTPose+[9] model trained on a general species dataset and average pool them into a 2048-dimensional vector. We train a linear SVM classifier based on the extracted pose-related features and achieve 91% accuracy(exceeding the 43% accuracy of using PSGFormer[10] as Reversion[2]) on the split test set. In the Pose-KP, we directly detect the poses of subjects in the generated images using ViTPose+. Then, we transform detected poses of each image into a graph where key points and the connections between them are respectively regarded as nodes and edges. We train a two-layer graph convolutional neural network(GCN)[3] on the task of whole-graph classification and achieve 85% accuracy on the split test set. For generated images, we compute the classification accuracy of the generated images as Pose-S and Pose-KP.

To better understand the proposed metrics and their complementarity, we visualize the generated images for different prediction results, as Fig. 2. We present detected poses of generated images and the predicted result by Pose-S and Pose-KP. As the first and fourth columns of Fig. 2, the Pose-S and Pose-KP can effectively assess the interactive poses between subjects. However, pose-S, which relies on the visual feature of the generated image, is confusing when assessing the image where the appearance of the subject’s limbs is vague. For example, in the second column of Fig. 2, one feet of the left wolf is not visual salience and the Pose-S misclassifies the kicking as handshaking. However, the Pose-KP correctly evaluates the results thanks to the robustness of the pose detector. Furthermore, as shown in the third column of Fig. 2, the metric Pose-S, based on

¹<https://huggingface.co/runwayml/stable-diffusion-v1-5>

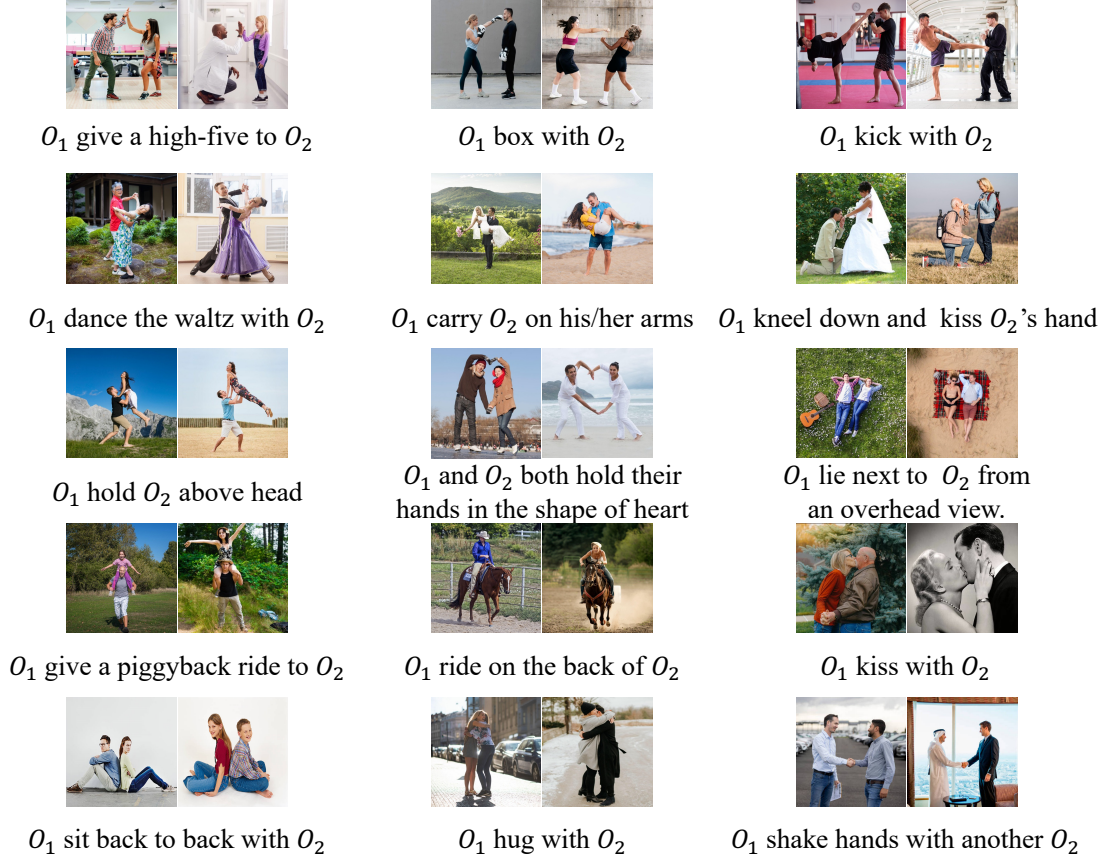


Figure 1: Illustration of the introduced 15 interactive poses.

image feature, provides accurate judgments when the pose detector is influenced by the detection results.

1.5 Implementation Details of the t-SNE Visualization

To visualize the distribution of the learned token embedding in the text embedding space, we randomly select 5 interactive poses with the learned poses embedding P_1^* , P_2^* and interaction embedding R^* . We represent about 50 embeddings of the commonly used words for each adjective, preposition, adverb, noun, proper noun and verb. The learned embeddings P_1^* and P_2^* are closer to the verb embedding cluster, while the embedding R^* are closer to the preposition embedding cluster. Compared to that, the token embeddings without using the language prior to the optimization process tend to be clustered in the inaccurate embedding space, such as ‘Adjectives’, ‘noun’ and ‘Conjunction’, which depends on the specific type of interactive pose and training exemplar images.

2 QUANTITATIVE ABLATION STUDY

In order to further analyze the effectiveness of the proposed components, we present a quantitative result of ablation studies as shown in Tab. 3. As the second row of Tab. 3, we do not use the language prior for two-stage inversion, *i.e.*, *w/o Prior*, the results of these

metrics suffer from severe decrease due to failing to decouple interactive poses from exemplar images. As shown in the third row of Tab. 3, we skip the single subject pose inversion stage and directly invert the interactive pose token $[R^*]$. Specifically, for a fair comparison, we extend token $[R]$ to three pseudo-words $[P_1][R][P_2]$ using verb prior for the tokens $[P_1]$ and $[P_2]$, and preposition prior for the token $[R]$. The results indicate that, despite using the same exploration space and prior constraints as the proposed method, a lack of explicit modelling of the pose of a single subject results in diminished accuracy of interactive poses. The effectiveness of the proposed cross-attention loss \mathcal{L}_{CA} is shown in the fourth row of Tab. 3. We observe that the incorrect interaction details appear in the generated images when the learning of token embedding R is not guided by the attention to the key interaction regions, thus leading to performance degradation in pose accuracy.

3 MORE QUALITATIVE RESULTS

3.1 Single Pose Inversion Results

In Fig. 3, we demonstrate generated images using the trained embeddings P_1^* and P_2^* in the single subject pose inversion stage. The results demonstrate the pseudo-words $[P_1^*]$ and $[P_2^*]$ can decompose the pose of a single subject from an interactive pose using generated exemplar images.

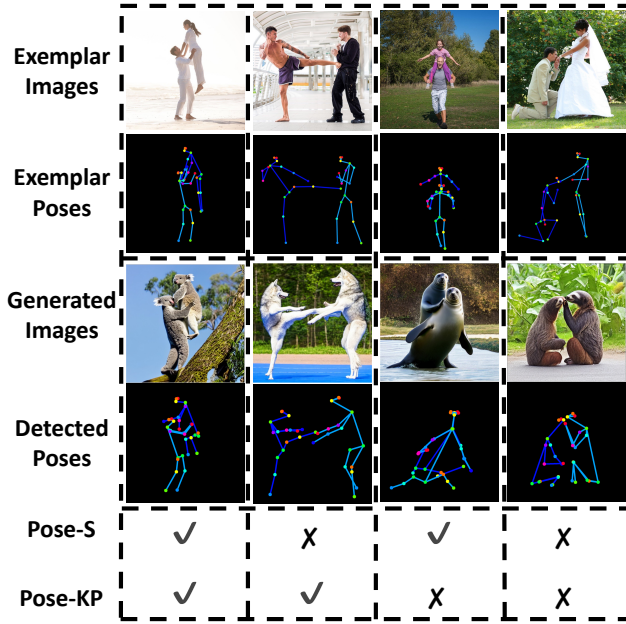


Figure 2: Examples of evaluating interactive pose using Pose-S and Pose-KP.

Table 3: Quantitative ablation study using proposed metrics(%).

Method	CLIP-T↑	CLIP-S↑	Pose-S↑	Pose-KP↑
SD [7]	23.51	21.87	7.05	6.36
w/o Prior	23.45	22.06	16.42	13.85
w/o Stage I	24.82	22.68	20.63	16.90
w/o L_{CA}	25.36	22.79	35.29	33.57
Ours	25.74	22.93	40.81	38.64

3.2 Interactive Pose Inversion Results

We present more generated images of the proposed method as Fig. 4 and Fig. 5. Furthermore, our method also has the power to create interactive poses between animated characters as shown in Fig. 6.

3.3 Interaction between Subjects of Different Species

As mentioned in [1, 4, 6], Stable Diffusion[7] is challenging to generate multiple subjects with different species due to the attribute binding issue. Precisely, given a text including multiple subjects with different colors, textures and appearances, SD can not properly align the attribute information with the specific subject in the generated image. Therefore, we integrate our method with Divide-and-Bind[4] to avoid texture entanglement or appearance leakage between subjects, as shown in Fig. 7. Furthermore, we do not use Divide-and-Bind in other experiments for fair comparison with other methods.

REFERENCES

- [1] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–10.
- [2] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin C.K. Chan, and Ziwei Liu. 2023. ReVersion: Diffusion-Based Relation Inversion from Images. *arXiv preprint arXiv:2303.13495* (2023).
- [3] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [4] Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. 2023. Divide & bind your attention for improved generative semantic nursing. *arXiv preprint arXiv:2307.10864* (2023).
- [5] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [6] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. 2024. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems* 36 (2024).
- [7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- [8] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. In *Advances in Neural Information Processing Systems*.
- [9] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. ViTPose+: Vision Transformer Foundation Model for Generic Body Pose Estimation. *arXiv preprint arXiv:2212.04246* (2022).
- [10] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. 2022. Panoptic scene graph generation. In *European Conference on Computer Vision*. Springer, 178–196.

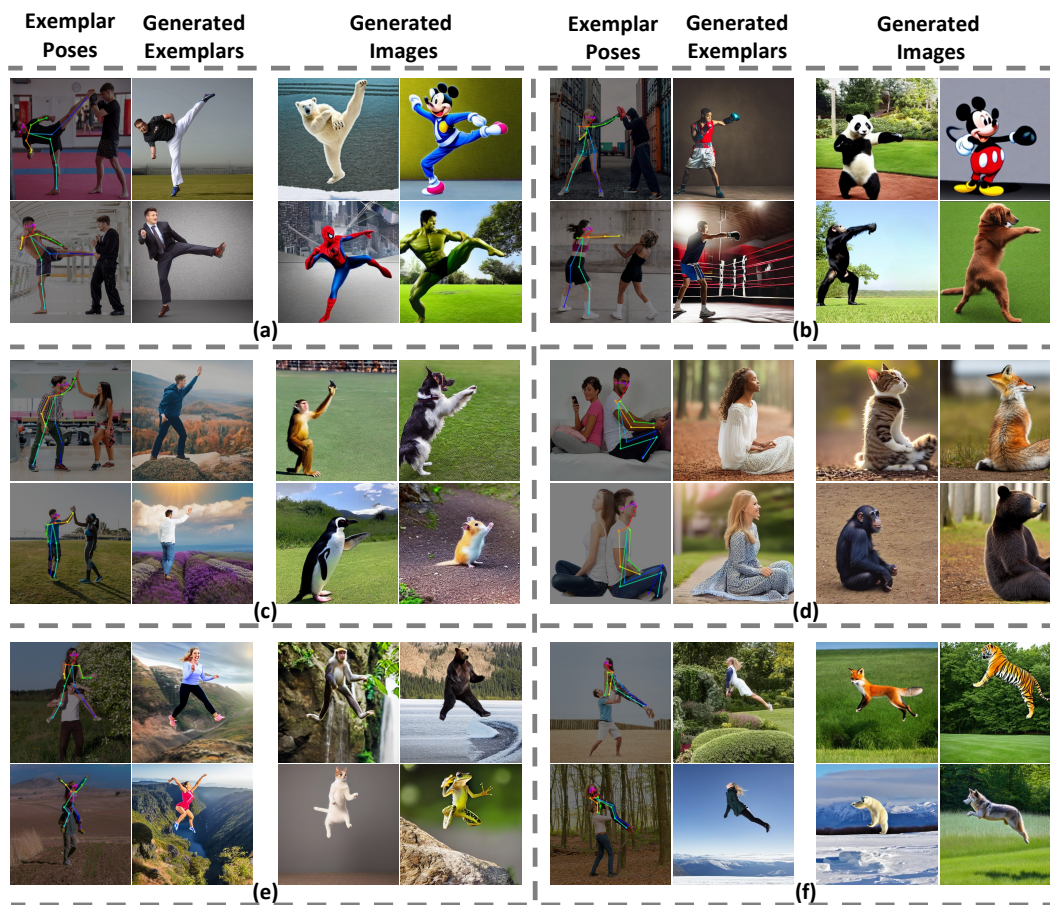


Figure 3: Qualitative results of single subject pose inversion.

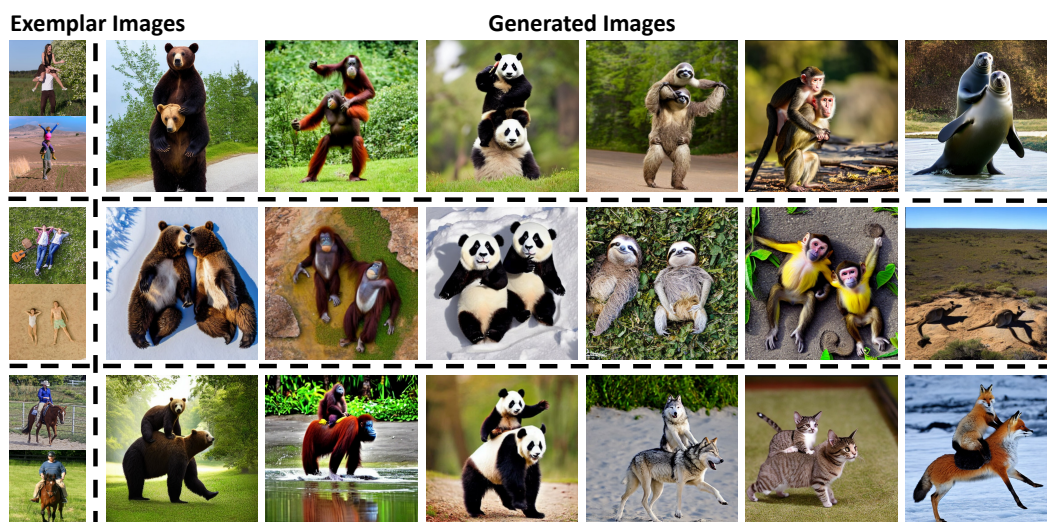


Figure 4: More qualitative results for customized interaction generation.



Figure 5: More qualitative results for customized interaction generation.

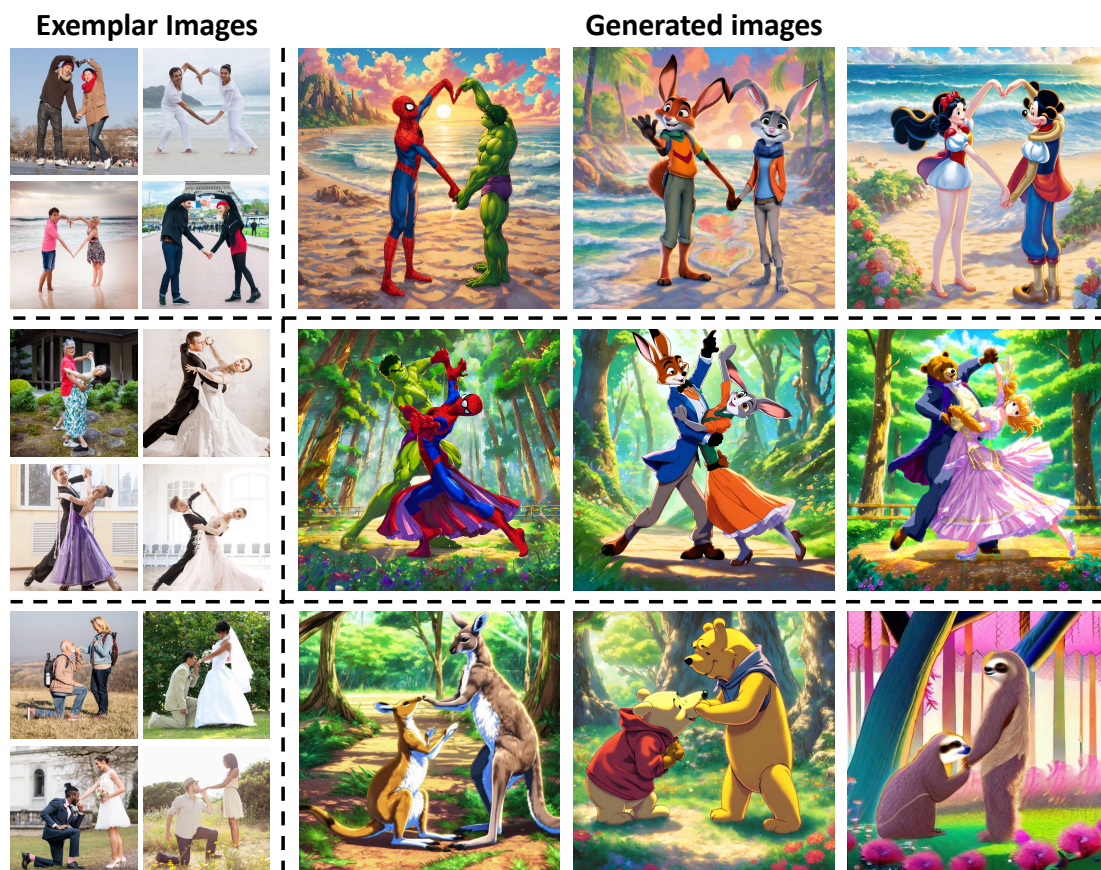


Figure 6: Generation with customized interactive pose between animated characters.



Figure 7: Generation with customized interactive pose between subjects of different species.