InteractDiffusion: Interaction Control in Text-to-Image Diffusion Models

Supplementary Material

6. Implementation Details.

Negative Prompt Negative prompts are commonly used in text-to-image generative works, which give negative guidance to avoids artefacts. We use the following negative prompt for all generation: "longbody, lowres, bad anatomy, bad hands, missing fingers, extra digit, fewer digits, cropped, worst quality, low quality". Tab. 3 shows the ablation experiment on the effect of negative prompt in FGAHOI Swin-Tiny Default setting.

Models	**With** Negative Prompt				**Without** Negative Prompt			
	Quality		HOI Det Score		Quality		HOI Det Score	
	FID	KID	Full	Rare	FID	KID	Full	Rare
StableDiffusion	35.85	0.01297	0.63	0.68	35.47	0.01216	0.78	1.01
GLIGEN	29.35	0.01275	21.73	15.35	28.29	0.01218	21.80	16.21
GLIGEN*	18.82	0.00694	25.23	17.45	19.69	0.00749	25.71	20.09
InteractDiffusion	18.69	0.00676	29.53	23.02	19.23	0.00769	29.37	23.18

Table 3. Comparison between **with** and **without** negative prompt.

Model Complexity. Tab. 4 shows the number of parameters in InteractDiffusion model in comparison with other diffusion-based baselines. The number of trainable parameters of InteractDiffusion is about 210 millions, only 1 millions more than GLIGEN, while introducing new interaction controllability. Note that these parameters counts do not include the text encoder and the VAE, which are same for all methods.

Method	N_{params}	$N_{\text{trainable}}$
StableDiffusion	860M	860M
GLIGEN	1069M	209M
InteractDiffusion	1070M	210M

Table 4. Number of parameters for InteractDiffusion in comparison with other diffusion-based baselines.

Network Architecture. In all experiments, Stable Diffusion V1.4 is used as base model for all methods. We maintain the network architecture except the transformer block in U-Net was adapted to include our Interaction Module.

7. Additional Ablation Studies

7.1. Scheduled Sampling

The scheduled sampling rate ω is a hyper-parameter in Interaction Transformer (Eq. (12)), which could greatly impact the generation as it control the degree of adherence to the interaction conditions. Thus, we ablate this hyper-parameter in interval of 0.1 from 0.0 to 1.0. Fig. 10 and Fig. 11 show the mAP and FID score for different values of scheduled sampling rate ω while Fig. 9 shows qualitative samples for different values of scheduled sampling rate ω.

From Figs. 10 and 11, we find that the interaction controllability improves as ω increases and converges around $\omega = 0.6$ and $\omega = 1.0$ produces best results in term of HOI detection score

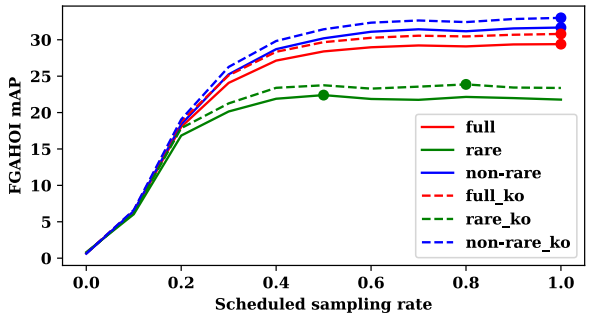

Figure 10. HOI detection score for various ω measured using FGAHOI with Swin-Tiny.

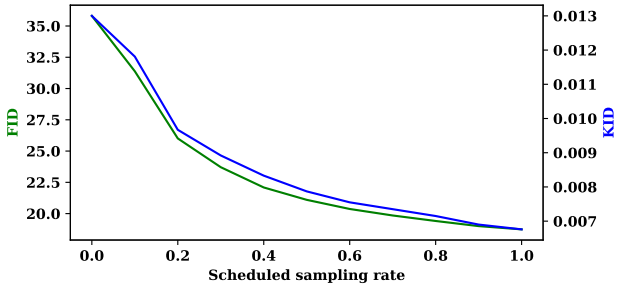

Figure 11. Quality scores for various ω.

for every subset, while FID and KID decreases gradually as ω increases and $\omega = 1.0$ produces least FID and KID distance when compared to original HICO-DET dataset. We recommend $\omega = 0.8$ in most of the cases, as it stride a balance between text caption and interaction condition adherence. In Fig. 9, the interaction correspondence increases gradually as ω increases, which is more obvious especially in range $\omega = 0.1$ to $\omega = 0.3$. When $\omega = 0.0$ is used, the model reduces back to the Stable Diffusion model where the Interaction Transformer is ignored.

7.2. Model Transferability

In the rapidly evolving field of text-to-image synthesis, personalized Stable Diffusion models have gained popularity for their capacity to generate images with distinct styles and traits. The interaction module's integration allowed for fine-grained interaction control over the generative process without necessitating extensive retraining. In our experiments, we conducted evaluations to assess the impact of the Interaction Module on several personalized Stable Diffusion models, including CuteYukiMix[1], RC-

[1] https://civitai.com/models/28169/cuteyukimixadorable-style

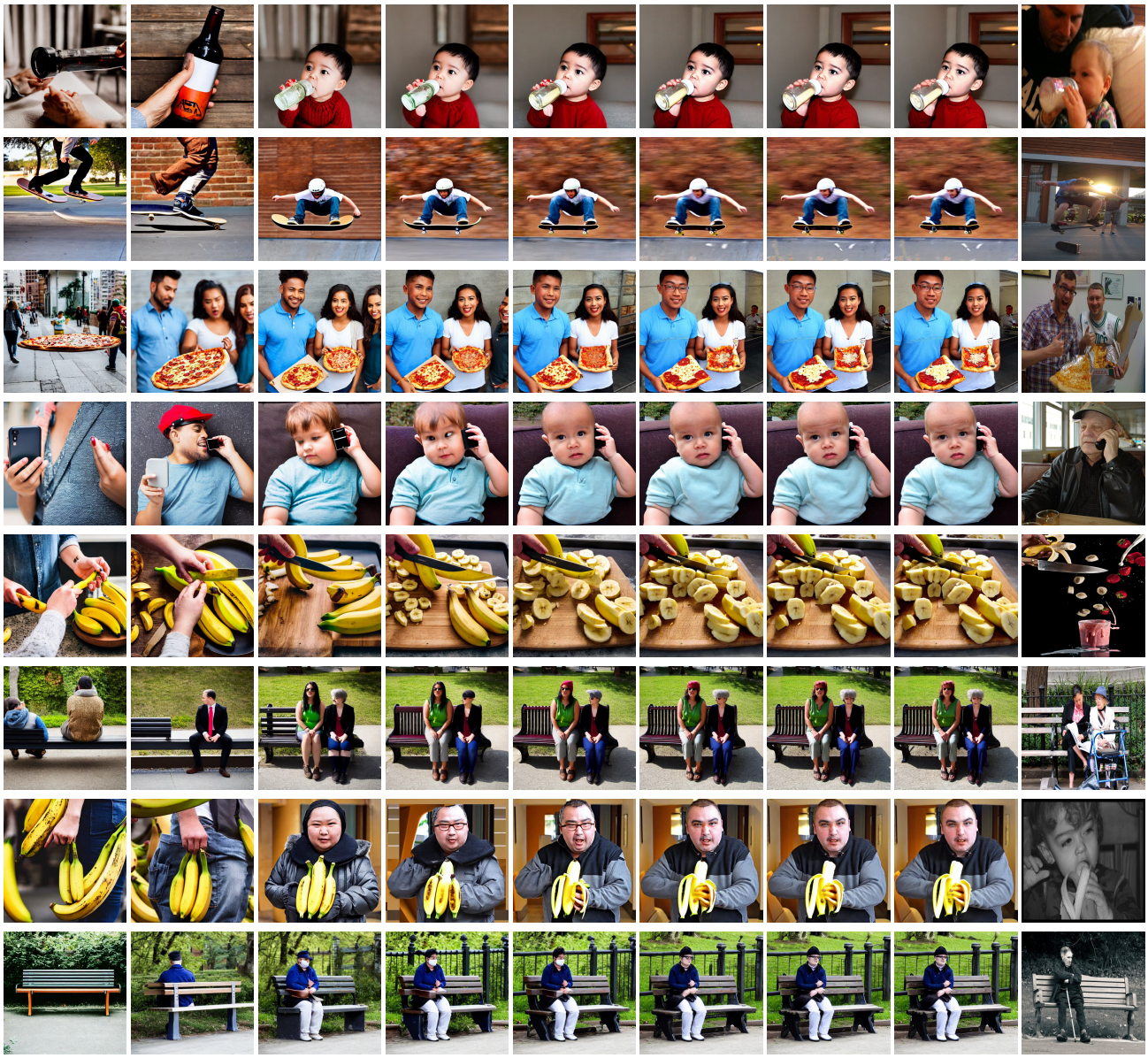

Figure 9. Ablation of scheduled sampling rate ω. It adjust the degree of attentiveness to interaction condition. Zoom in for detail.

NZCartoon3D[2], ToonYou[3], Lyriel[4], DarkSushiMix[5], RealisticVision[6], and ChilloutMix[7]. We observed that our transferable interaction module successfully maintains the unique stylistic attributes of personalized models while offering improved interaction controllability. We demonstrates visualization of InteractDiffusion on various personalized Stable Diffusion models on Fig. 12, further affirming the module's potential to introduce interaction control

[2] https://civitai.com/models/66347/rcnz-cartoon-3d

[3] https://civitai.com/models/30240/toonyou

[4] https://civitai.com/models/22922/lyriel

[5] https://civitai.com/models/24779/dark-sushi-mix-mix

[6] https://civitai.com/models/4201/realistic-vision-v51

[7] https://civitai.com/models/6424/chilloutmix

Figure 12. Visualization of InteractDiffusion on various personalized StableDiffusion models. Zoom in for detail.

without hindering the distinct qualities of these models.

Method	Tiny			Large		
	Full	Unseen	Seen	Full	Unseen	Seen
Zero-shot						
InteractDiffusion (ZS)	28.47(-0.65)	20.75(-3.10)	30.41(-0.03)	30.31(-0.73)	23.06(-2.30)	32.12(-0.34)
Fully Seen						
GLIGEN*	25.23	17.77	27.10	26.45	19.23	28.25
InteractDiffusion	29.12	23.85	30.44	31.04	25.36	32.46
Reference						
HICO-DET	29.81	22.69	32.59	37.11	32.59	38.24

Table 5. Zero-shot performance of InteractDiffusion compared to default fully-seen setting. Comparison were made in relatively to Fully-Seen setting.

Figure 13. Visualization of InteractDiffusion and others demonstrating the generation of *different objects* for the same action.

Figure 14. Visualization of InteractDiffusion demonstrating the generation of *different actions* for the same object. Zoom in for detail.

7.3. Zero-shot experiments

Following the setting in zero-shot HOI detection work [32], we choose 120 HOI classes from total 600 classes in HICO-DET as unseen subset which does not involve in training, while the remaining 480 classes are in seen subset, which will be used in training. We use the same split as in [32]. We train the InteractDiffusion for similar number of iterations as the default setting to ensure fairness.

Tab. 5 shows the zero-shot performance of InteractDiffusion. In seen subset, no significant performance drop is observed, while for unseen setting, we observe mAP drop of only 3.10 and 2.30 for FGAHOI with Swin-Tiny and Swin-Large backbones, respectively. This shows that our InteractDiffusion only suffer a minor drop in its zero-shot performance, demonstrate its capability in generate unseen interaction combinations.

8. More Qualitative Results

In Fig. 13, we visualize how our InteractDiffusion renders different objects with the same action; while Fig. 14 shows how our InteractDiffusion renders different actions with the same object. This shows that our model can generate various combinations of interactions that maintain the coherence and naturalness of interactions between people and objects.

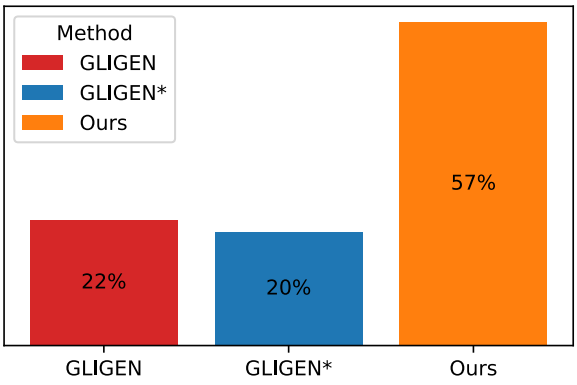

Figure 15. Overview of user preferences.

9. User Preference Study

To evaluate user preferences, we conducted a user study with 86 respondents. Each of the 5 prompts was accompanied by a set of 12 images for evaluation. Among these images, 4 were generated by each method: GLIGEN, fine-tuned GLIGEN (GLIGEN*), and our proposed method. Users were asked to select the 4 images that

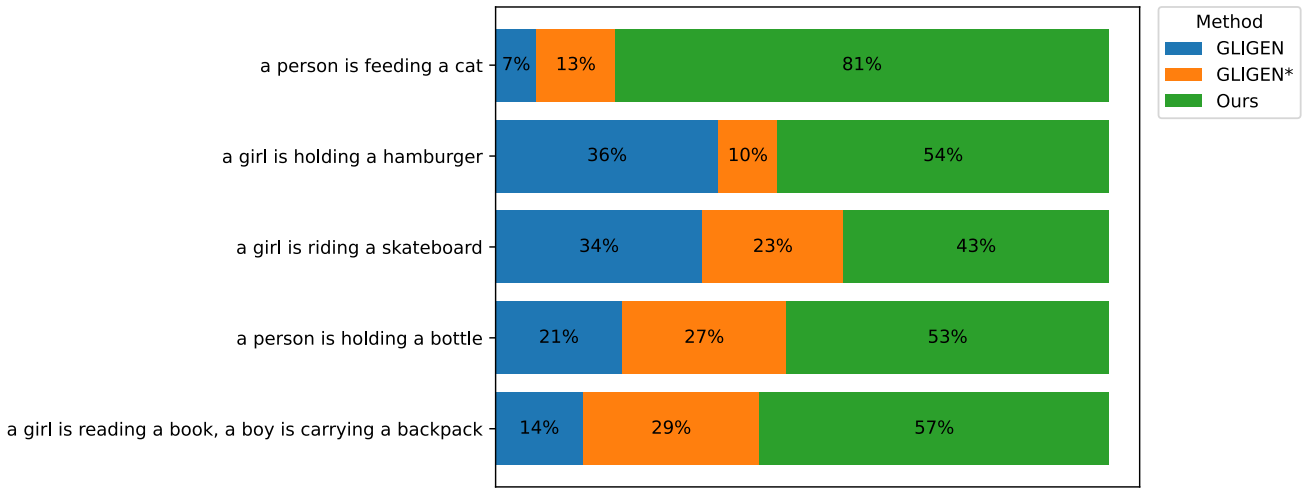

Figure 16. User preferences for each text prompt.

best corresponded to the given text prompt out of the 12 randomly-shuffled choices provided.

Fig. 15 show the overall user preferences among GLIGEN, GLIGEN* and our InteractDiffusion. Among the total responses, the majority of users (57.50%) preferred the images generated by our method (Ours), while GLIGEN and GLIGEN* received 22.33% and 20.17% of the votes, respectively. The high percentage of preferences for our method suggests that it performed comparatively better in satisfying user preferences in the evaluated scenarios. This could be attributed to various factors, such as the effectiveness of our interaction control mechanism, the quality of generated images, or other user-centric considerations.

On the other hand, Fig. 16 offers insights into user preferences categorized by each text prompt, providing a more detailed analysis of how preferences vary across different prompts. For all text prompt, our method received the highest preference. While the prompt "*a girl is riding a skateboard*" demonstrates a relatively balanced distribution of user preferences across the three methods, the prompt "*a person is feeding a cat*" exhibits a more pronounced bias towards our method. This disparity suggests that certain interaction scenarios may be more effectively conveyed by our approach compared to others.

10. Limitations

Despite significant improvements in various metrics, the generated interaction still show some difference from realistic, especially in finer detail. This could be discovered on the mAP of larger detector (*i.e*. FGAHOI(Swin-Large)), which pays attention to the finer detail in detecting HOI. Besides, we discovered that existing large pretrained models(CLIP[25],StableDiffusion[27]) are object-focused in pre-training stage, thus lack of understanding of interaction, which hinders the performance of InteractDiffusion in controlling the interaction. We expect that a more diversely trained large model that includes the both object and interaction could boost the interaction controllability of InteractDiffusion.