VerbDiff: Text-Only Diffusion Models with Enhanced Interaction Awareness

Supplementary Material

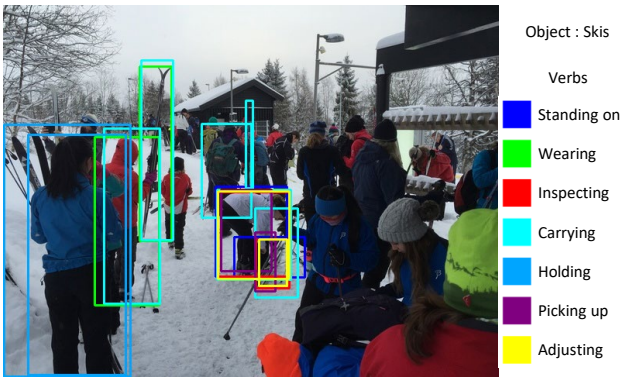

Figure 1. **An image sample from HICO-DET containing multiple human-object interactions.** Each colored box corresponds to a distinct human-object interaction, representing different interaction words.

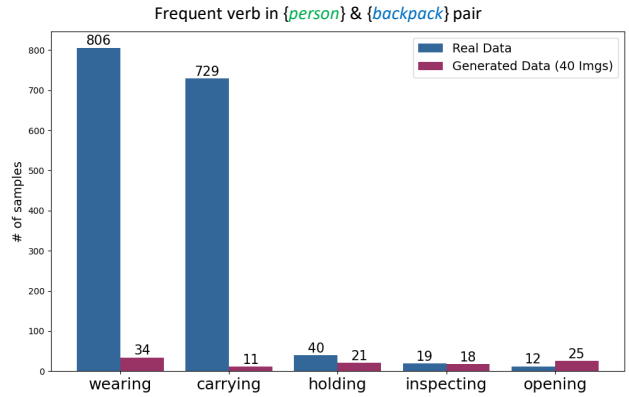

Figure 2. **Frequent verb in real data and the number of samples contain the frequent verb in captions extracted from generated images.** The blue bars indicate the number of samples in the real dataset for each interaction verb associated with "backpack". The purple bars represent the number of captions containing the most frequent verb "wearing," extracted from 40 images generated with ground-truth verbs.

A. Data Analysis

A.1. HICO-DET Dataset

We present an example from HICO-DET containing multiple interactions within a single image. As shown in Fig. 1, the image includes various interaction verbs (*e.g.*, "wearing"), which can confuse the model when trained using the conventional reconstruction loss. To isolate interactions corresponding to the correct interaction verbs, we apply a mask $\mathcal{M}$ during training.

CLIP		S-BERT	
Verb	Score	Verb	Score
riding	1.0	**riding**	1.0
hopping	0.9595	**sitting on**	0.8944
sitting on	0.9585	**straddling**	0.8385
straddling	0.9551	holding	0.8745
walking	0.9512	**walking**	0.8711
carrying	0.9341	carrying	0.8555
pushing	0.9277	jumping	0.8468
jumping	0.9111	**hopping**	0.8436
holding	0.8960	parking	0.8218
parking	0.8950	inspecting	0.7982
inspecting	0.8662	pushing	0.7975
repairing	0.8291	repairing	0.7456
washing	0.8081	washing	0.7253

Table 1. **Cosine similarity comparison between CLIP and Sentence-BERT (S-BERT).** We score the similarity of the verb "riding" with other verbs associated with the object "bicycle" in real data.

A.2. Anchor Interaction Words

Fig. 2 presents two graphs for the human and "backpack" pair example. We generate 40 images per prompt (*e.g.*, "A person {verb} a backpack") and extract captions using InstructBLIP. We then count the captions that include the most frequent verb from the real data (*e.g.*, "wearing"). Across all generated samples, nearly half of the captions contain the frequent verb, which we define as the anchor interaction word for relation disentanglement guidance.

B. Implementation & Evaluation Details

B.1. Implementation Details

We use "black and white image, extra arms, extra legs" as negative prompts when generating biased interaction images during training. Additionally, we include "naked, poor resolution" to negative prompts for image generation in all our experiments. For encoding texts and images, we utilize CLIP ViT-L/14 weights, consistent with the weights used for the text encoder in SD.

B.2. Sentence-BERT Evaluation Metric

We compare the cosine similarity differences between CLIP and Sentence-BERT, focusing on interaction verbs associated with the object "bicycle". Using the text template: "A photo of a person {verb} a bicycle," we calculate cosine similarities, as shown in Tab. 1. As highlighted by the scores for bolded verbs (*e.g.*, "walking"), CLIP evaluates all verbs as highly similar, even when humans perceive them

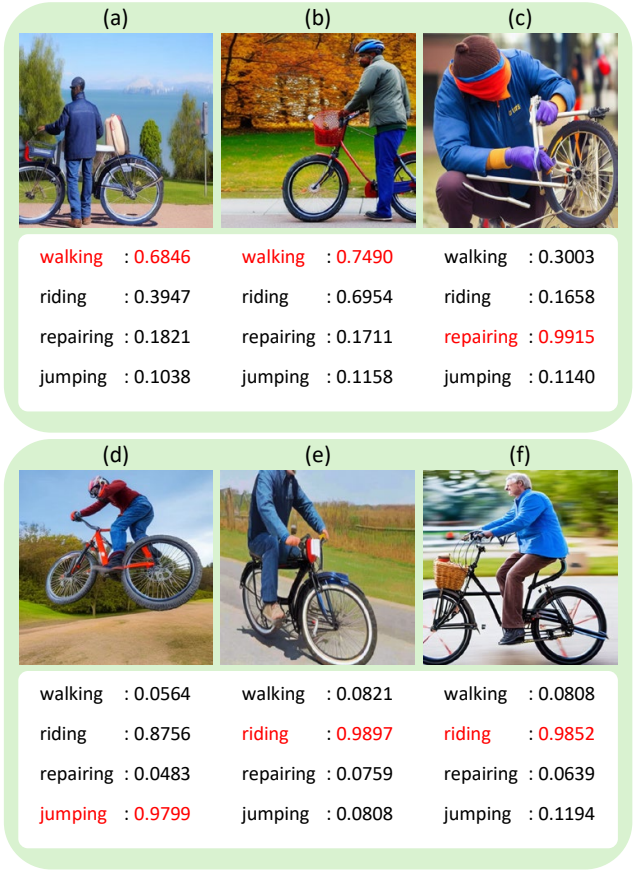

Figure 3. **VQA-score result examples.** We measure the probability that the VQA model answers "yes" for questions based on four verbs associated with the "bicycle" class. The verb with the highest score is highlighted in red.

as distinct. In contrast, S-BERT successfully distinguishes between interaction verbs, better reflecting the interaction differences that humans recognize in real human-object interactions.

B.3. VQA-Score

For assessing the image-to-text (I2T) alignment score and VQA-score, we use CLIP-FlanT5 weights. Additionally, we apply the following question template: "Is this figure showing a {H} {R} a/an {O}? Please answer yes or no". Fig. 3 illustrates VQA-score examples for the "bicycle" class. We use four verbs in the question template and calculate the probability scores for each image. As shown, the VQA-score effectively distinguishes complex prompts involving human-object interactions. In particular, the model differentiates images with ambiguous interactions (b), successfully identifying the human foot on the ground.

$\mathcal{L}_{IDG}$	CLIP	S-BERT	$\mathcal{L}_{rec}$	CLIP	S-BERT
SD	0.725	0.620	SD	0.725	0.620
0.2	0.725	0.638	0.4	0.733	0.634
0.4	0.723	0.638	0.7	0.732	0.636
0.8	**0.729**	**0.640**	1.0	**0.735**	**0.638**
1.2	0.726	0.639	1.5	0.732	0.637
1.6	0.724	0.639	2.0	0.731	0.636

Table 2. **Score variations across the interaction direction guidance (IDG) and reconstruction loss weights.** Scores are evaluated using CLIP and Sentence-BERT (S-BERT) text-to-text similarity metrics. The top row represents the scores from Stable Diffusion (SD).

$\mathcal{L}_{RDG}$	CLIP	S-BERT	m	CLIP	S-BERT
SD	0.725	0.620	SD	0.725	0.620
1	0.735	0.640	0.1	0.733	0.634
5	0.735	0.641	0.2	**0.735**	**0.638**
10	**0.736**	**0.641**	0.4	0.732	0.636
20	0.734	0.640	-	-	-

Table 3. **Score variations based on the weight of relation disentanglement guidance (RDG) and the margin m in $\mathcal{L}_{triple}$ for RDG.** Scores are evaluated using CLIP and Sentence-BERT (S-BERT) text-to-text similarity metrics. The top row represents the scores from Stable Diffusion (SD).

Model	SOV-STG-S (Acc ↑)			
	Def.		KO.	
	Full	Rare	Full	Rare
SD	16.09	4.59	18.22	4.85
w/o α	17.83	5.48	19.12	5.65
w/ α	**22.59**	**7.62**	**24.79**	**7.83**

Table 4. **HOI accuracy comparison with and without adaptive interaction modification α.** Scores are evaluated using SOV-STG-S weights. The top row represents the results from Stable Diffusion (SD).

C. Additional Ablations

C.1. Model Weight Comparison

We compare the weight hyperparameters for relation disentanglement guidance (RDG), interaction direction guidance (IDG), reconstruction loss, and the margin value m in $\mathcal{L}_{triple}$ of RDG in Tab. 3 and Tab. 2, respectively. Additionally, we include the second-best model (SD) in the top row of each table for reference. The overall scores differ from the main evaluation because we evaluate in different settings to observe score differences more clearly.

Tab. 4 shows the HOI accuracy differences when using adaptive interaction modification α. To highlight the effectiveness of α in balancing modification across interaction words, we compare it under the same settings as the main evaluation table. The results demonstrate that adaptive interaction modification successfully balances the extent of interaction adjustments, preventing overfitting interaction

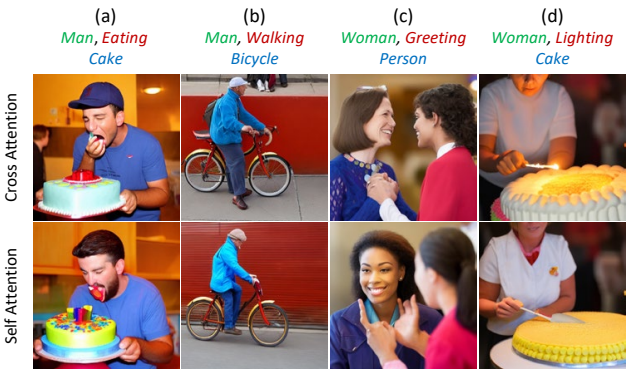

Figure 4. **Comparison between self-attention and cross-attention layer tuning.** While the cross-attention-tuned model generates images with accurate interactions, the self-attention-tuned model fails to capture precise interactions.

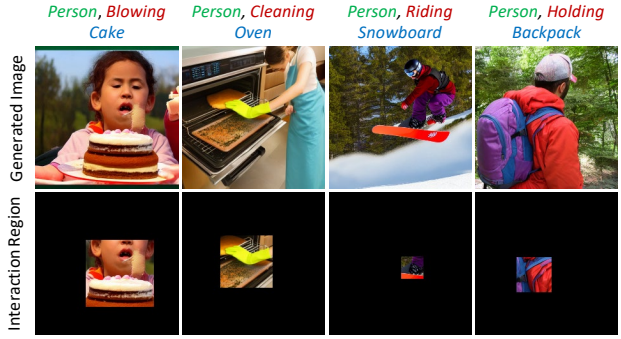

Figure 5. **Interaction region in generated images.**

words with many samples in the real dataset.

C.2. Self-Attention & Cross-Attention

We show the interaction differences depending on which layer is tuned in Fig. 4. As shown, the self-attention-tuned model fails to understand the semantics of interaction verbs and generates images with inaccurate interactions. However, the cross-attention tuned model (VerbDiff) accurately depicts the intended interactions. This demonstrates that the cross-attention layer better reflects nuanced interactions, enhancing the interaction word understanding in SD.

C.3. Interaction Region

We present the generated images and the extracted interaction regions utilized during training in Fig. 5. As can be seen, the IR module extract quite accurate interaction region between humans and objects.

D. Additional Qualitative Results

Fig. 7 presents additional results comparing VerbDiff with SD, GLIGEN, InteractDiffusion, Our model generates high-quality images with accurate interactions that closely resemble the ground truth and those produced by DALL-E 3. For example, in Fig. 7 (e), it successfully captures a man lighting a candle, while GLIGEN fails to generate the candle, and InteractDiffusion does not depict the action correctly.

Figure 6. **Failure Cases with previous methods.**

Additionally, Fig. 8 shows diverse images from complex prompts with multiple interactions. Even when provided with precise bounding boxes extracted from our generated images, GLIGEN and InteractDiffusion fail to produce the intended interactions or objects accurately. In contrast, our model captures interactions effectively without additional conditions, achieving quality comparable to DALL-E 3 and enhancing interaction understanding within SD.

E. Limitations

VerbDiff generates high-quality images with accurate interactions by improving the interaction word comprehension in SD. While our model achieves better scores across various evaluation metrics, there are cases where several humans perform the same interaction within a single training image. As shown in Fig. 6, VerbDiff occasionally miscounts the number of humans and inherits the limitations of the Stable Diffusion (SD) framework.

Additionally, it still produces some deviations from real interactions when handling complex prompts involving multiple human-object interactions. This limitation may arise from the model focusing on distinguishing interaction words without considering similarities between interactions involving different objects (*e.g.*, "walking a bicycle" and "walking a motorcycle"). We anticipate that incorporating more generalized representations of interactions between humans and objects could further enhance interaction comprehension in text-to-image models.

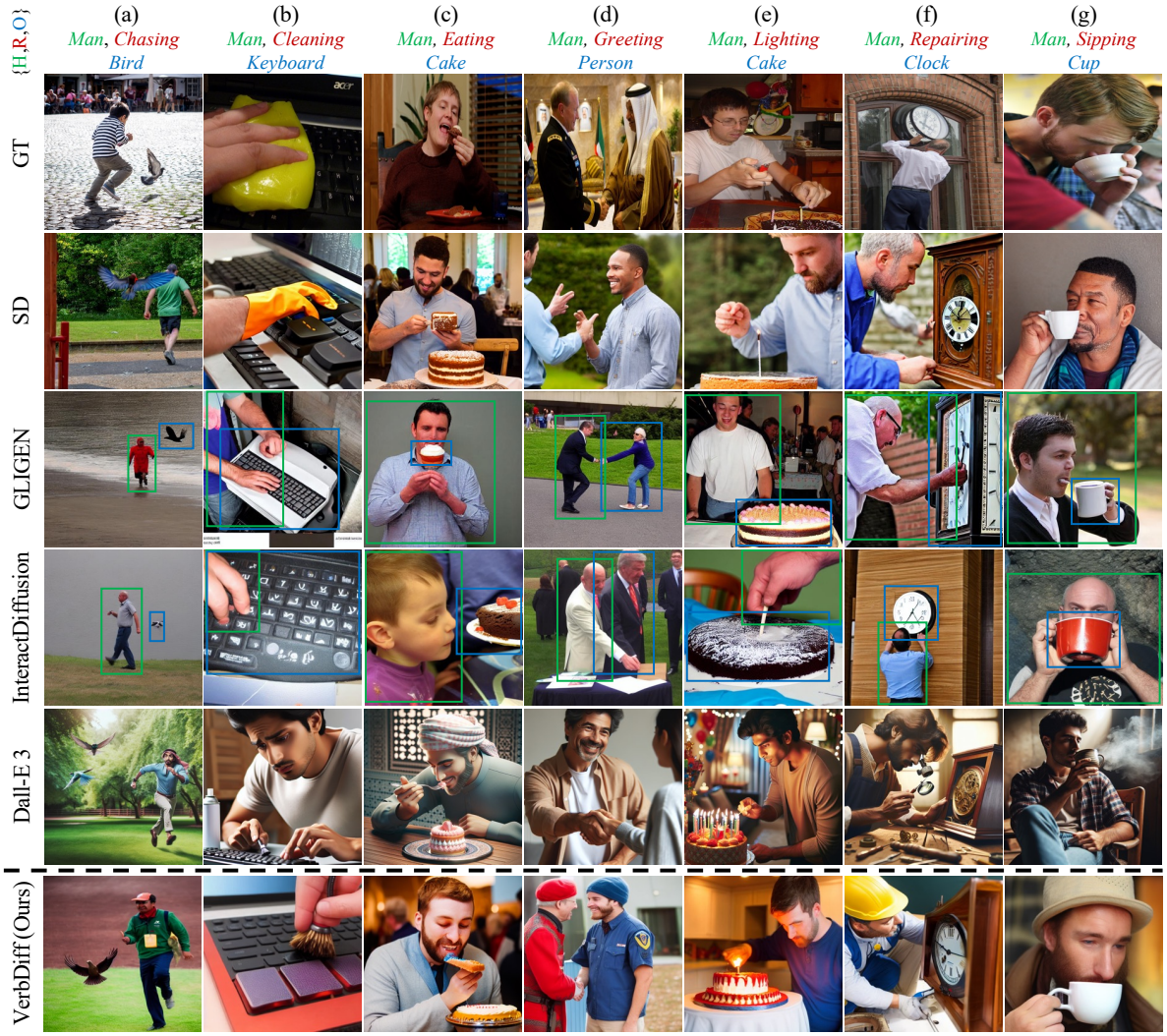

Figure 7. **Additional single interaction qualitative results.** The colored boxes mean the input bounding boxes.

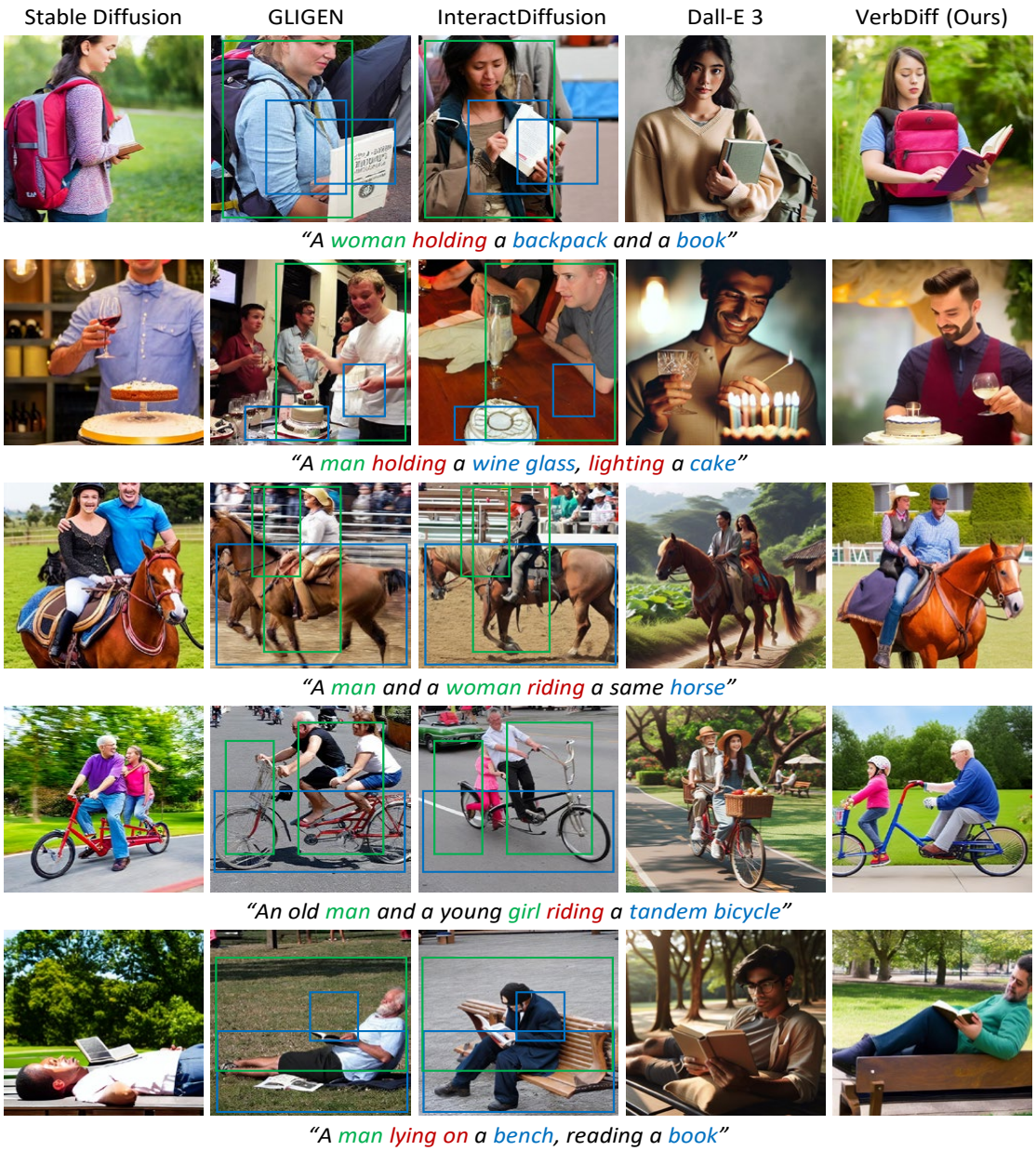

Figure 8. **Additional multiple interactions qualitative results.** The colored boxes represent the corresponding human and object bounding boxes. We extract the box from our generated images and apply it to the GLIGEN and InteractDiffusion.