VerbDiff

Text-only Diffusion Models with Enhanced Interaction Awareness

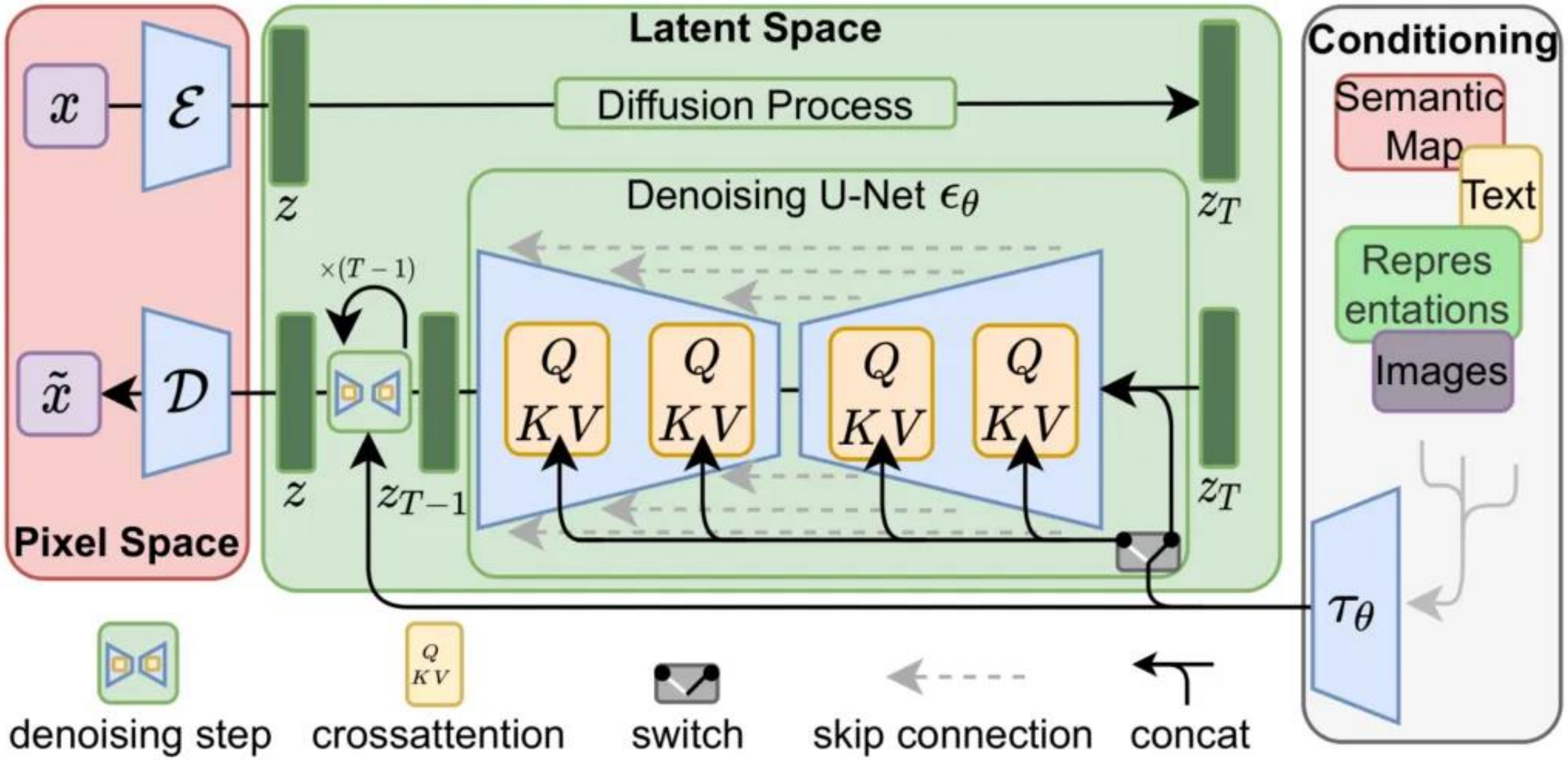
25’ CVPR

Seung Ju Cha, Kwanyoung Lee, Ye-Chan Kim, Hyunwoo Oh, Dong-Jin Kim

HANYANG UNIVERSITY

Contents

Backgrounds



LDM[1]

Stable Diffusion(SD)[1] takes natural prompts as input and can generate photo-realistic images

[1] Rombach, Blattmann, et al. "High-Resolution Image Synthesis with Latent Diffusion Models." CVPR (2022)

However...

SD sometimes struggles to generate accurate interactions between humans and objects[2]

Exiting, Train

Walking, Bicycle

Holding, Backpack

Pouring, Bottle

The generated images do not match with our perceptual interactions

SD cannot reflect the intended interactions as these are detailed semantics when depicting images[2]

[2] Tian Hoe, Jiang, et al. "InteractDiffusion: Interaction Control in Text-to-Image Diffusion Models." CVPR (2024)

Why?

CLIP has a strong bias towards objects or backgrounds[3]

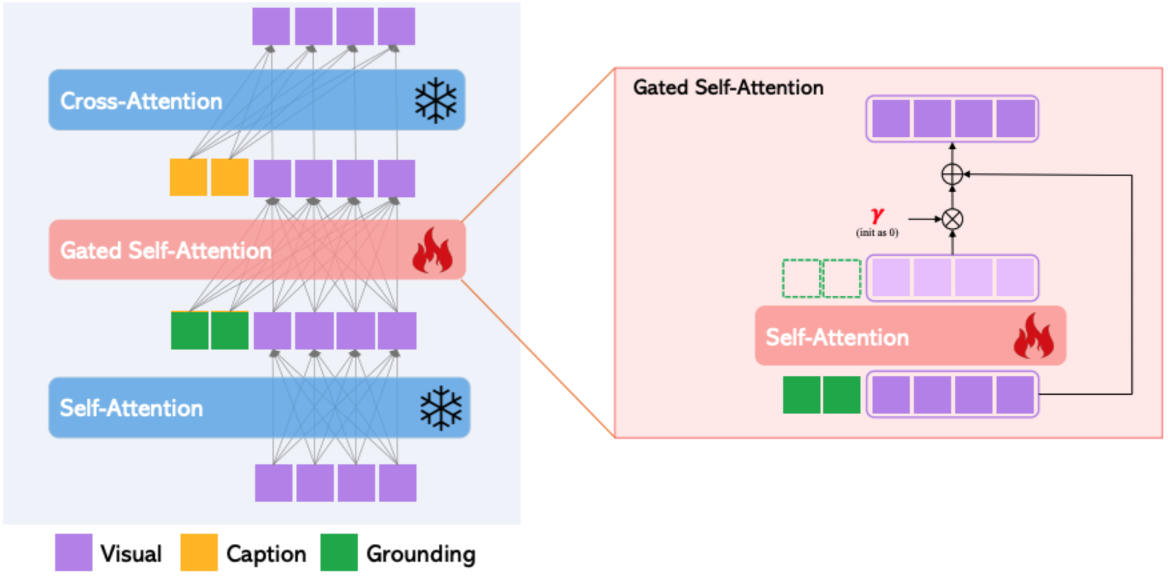
CLIP has **a lack of verb understandability** → affects the diffusion model

[3] Momeni, Liliane, et al. Momeni, Liliane, et al. "Verbs in action: Improving verb understanding in video-language models." ICCV (2023)

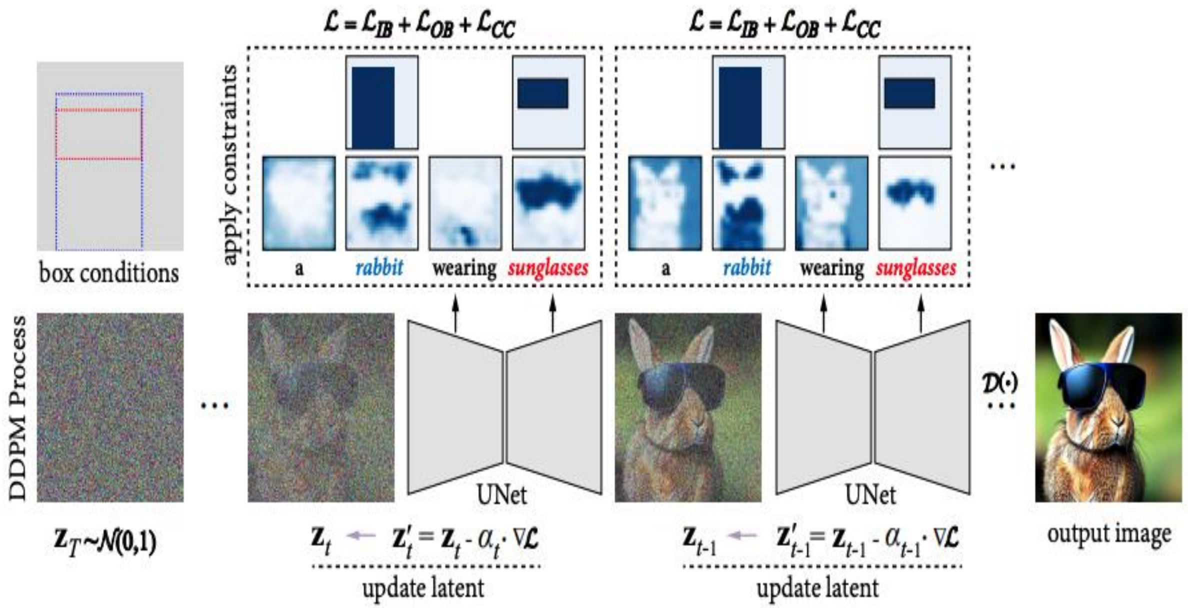
Related Works

To enhance the text understandability

1. Layout (bounding box) based method



GLIGEN[4]

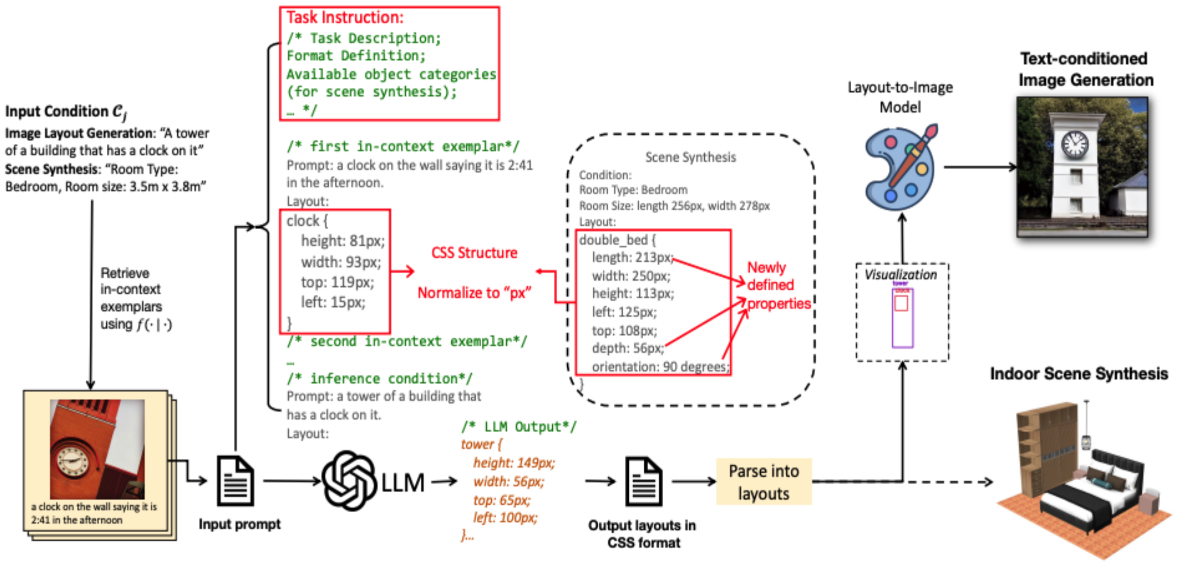


BoxDiff[5]

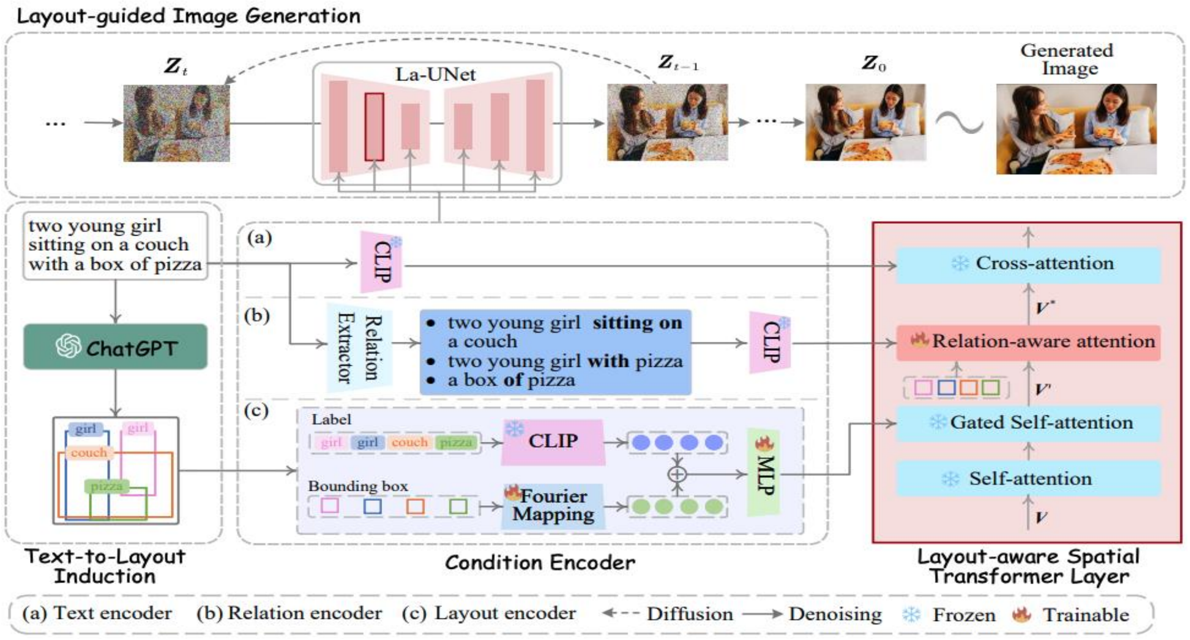
[4] Li, Yuheng, et al. "Gligen: Open-set grounded text-to-image generation." CVPR (2023)
[5] Xie, Jinheng, et al. "Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion." ICCV (2023)

Related Works

2. LLM based method



LayoutGPT[6]



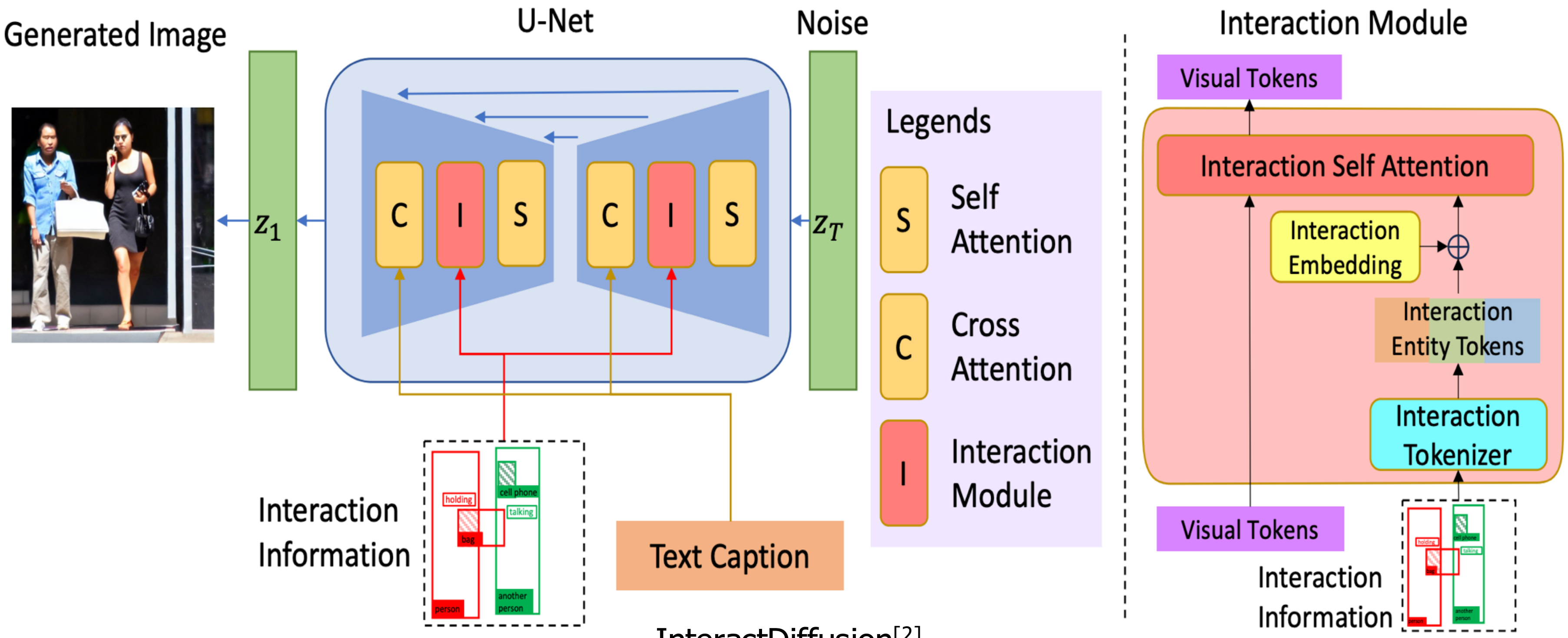
LayoutLLM-T2I[7]

These methods **focus on the spatial location of object relation not the interaction**

[6] Feng, Weixi, et al. "Layoutgpt: Compositional visual planning and generation with large language models." NeruIPS (2024)

[7] Qu, Leigang, et al. "Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation." ACM MM (2023)

Related Works



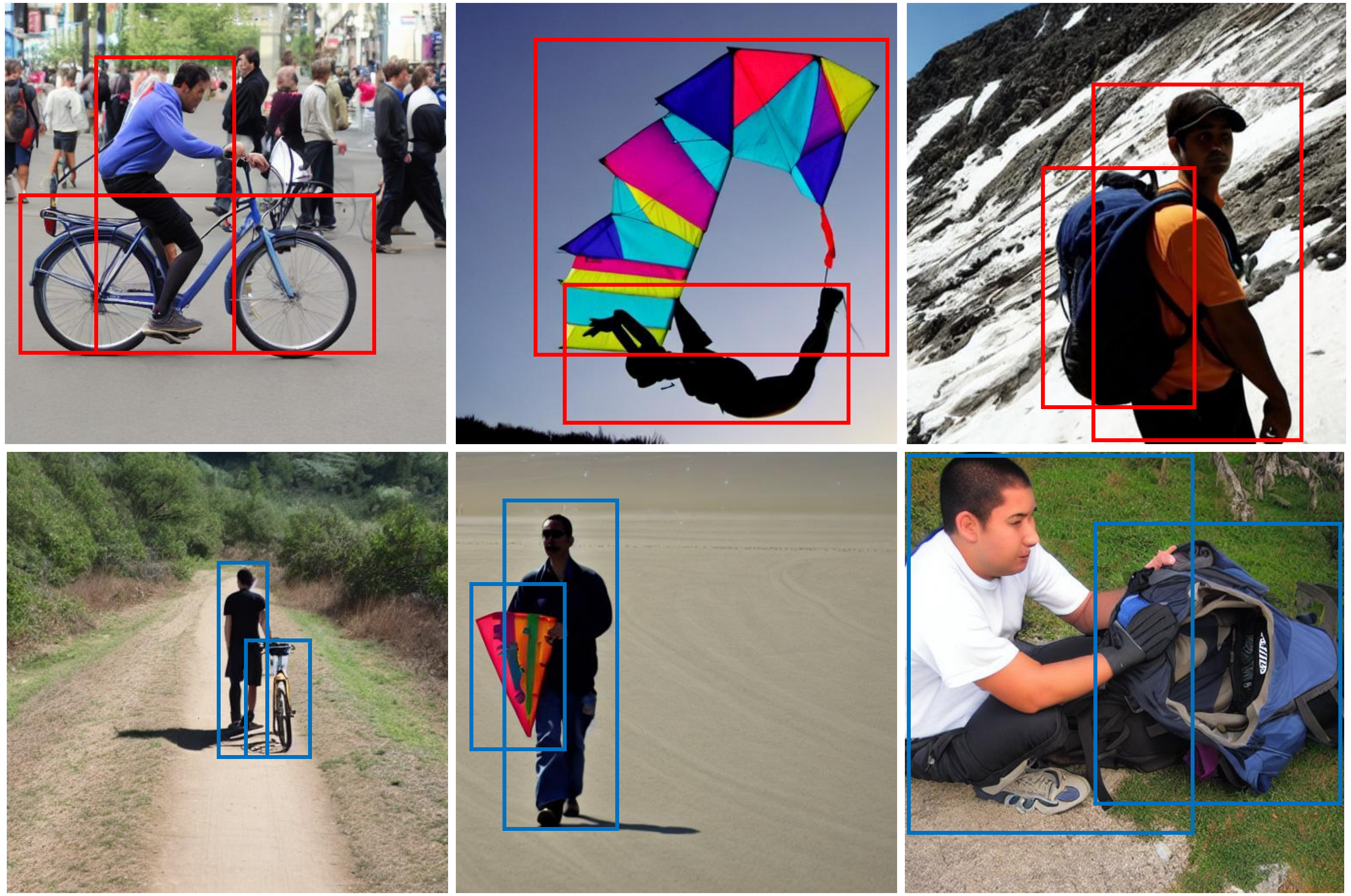
InteractDiffusion[2]

Leverage Bounding Box corresponding to Human, Object

Training Interaction Self-Attention Layer

Research Question

Walking, Bicycle　Carrying, Kite　Holding, Backpack



Success and failure cases of InteractDiffusion[2]

When given bounding boxes that **multiple interaction shares**
e.g., walking/riding bicycle
↓
Cannot distinguish the interaction word semantic difference

Heavily relies on the accurate bounding box
↓
considering the precise boxes for interaction are labor-intensive

Research Question

Problem 1. Still has problem in **distinguishing the semantic differences between interaction words (verbs)**

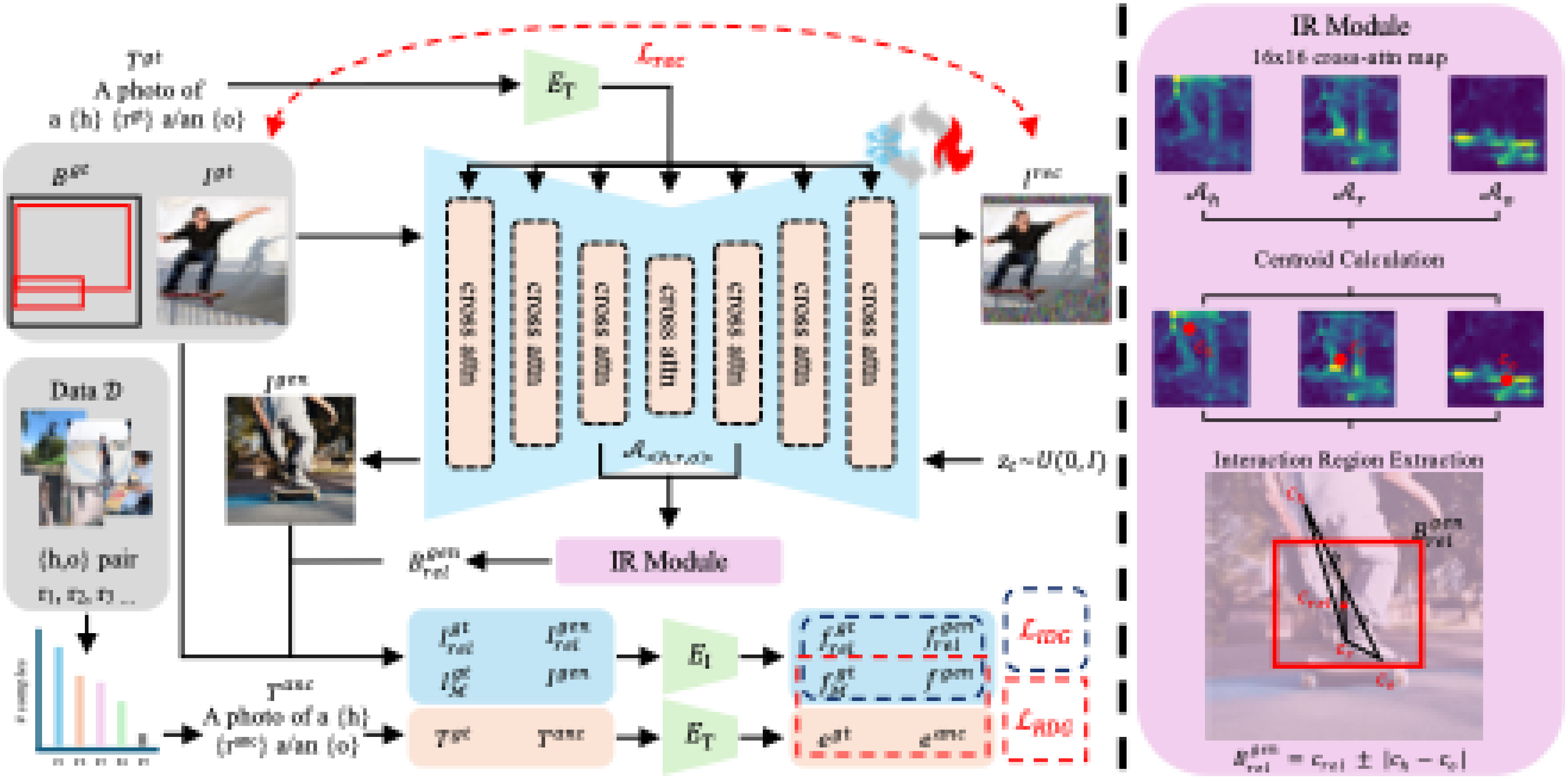
Problem 2. The generated images **heavily rely on precise bounding boxes which is labor-intensive** to provide

Can we enhance the understandability of interaction words in SD in a more generalized way?

Propose VerbDiff with two additional guidances

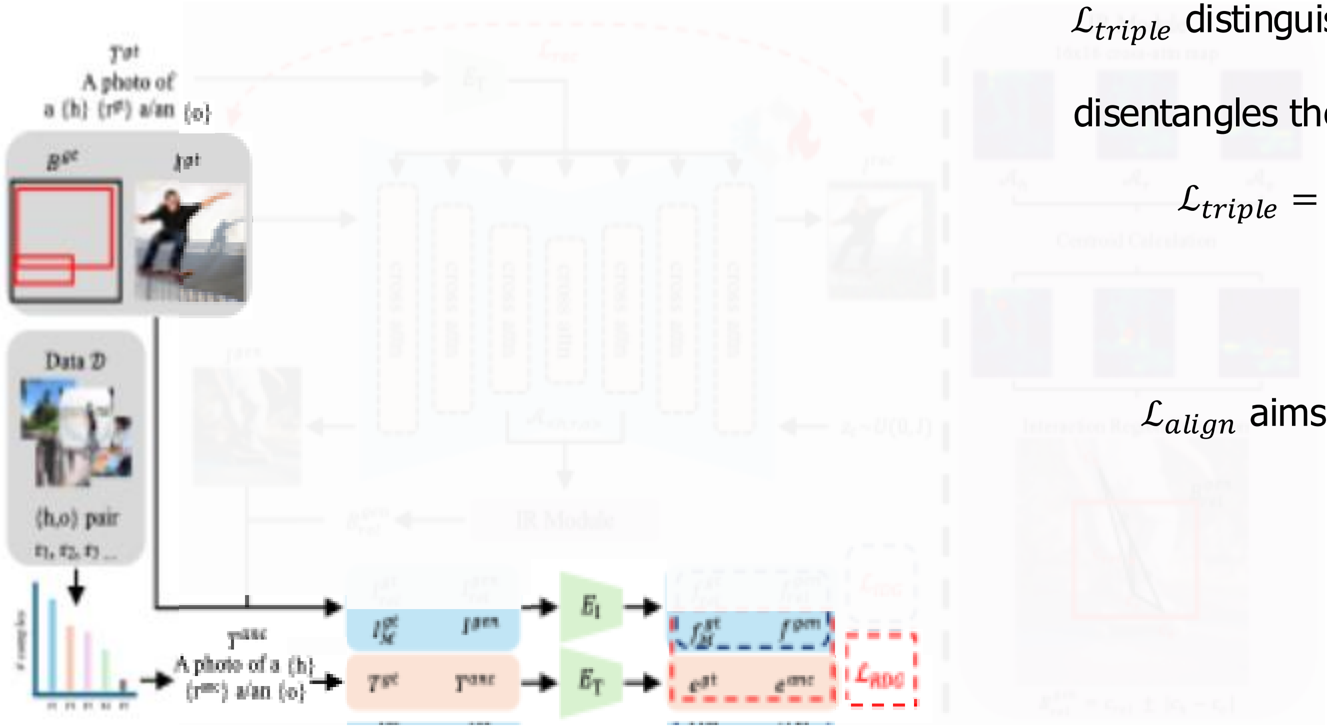
1. **R**elation **D**isentanglement **G**uidance based on **frequency-based anchor text**
2. **I**nteraction **D**irection **G**uidance with **IR modules** that capture detailed interaction regions

Methods



T^{gt}
A photo of a {h} {r^{gt}} a/an {o}
E_T
$\mathcal{L}_{rec}$
B^{gt}
I^{gt}
I^{rec}
cross attn
$\mathcal{A}_{\{h,r,o\}}$
$z_t \sim U(0, I)$
Data $\mathcal{D}$
{h,o} pair
r_1, r_2, r_3 ...
I^{gen}
B^{gen}_{rel}
IR Module
T^{aug}
A photo of a {h} {r^{aug}} a/an {o}
I^{gt}_{rel}
I^{gen}_{rel}
I^{gt}
I^{gen}
E_I
f^{gt}_{rel}
f^{gen}_{rel}
f^{gt}
f^{gen}
$\mathcal{L}_{RDG}$
T^{gt}
T^{aug}
e^{gt}
e^{aug}
$\mathcal{L}_{RDG}$
IR Module
16x16 cross-attn map
$\mathcal{A}_h$
$\mathcal{A}_r$
$\mathcal{A}_o$
Centroid Calculation
Interaction Region Extraction
B^{gen}_{rel}
$B^{gen}_{rel} = c_{rel} \pm |c_h - c_o|$

Relation Disentanglement Guidance



$\mathcal{L}_{triple}$ distinguish the semantic difference between interaction words

disentangles the input words from the frequency-based anchor word

$$\mathcal{L}_{triple} = \max(0, m + sim(f^{gen}, e^{gt}) - sim(f^{gen}, e^{anc}))$$

$\mathcal{L}_{align}$ aims to control the accurate interaction in image level

$$\mathcal{L}_{align} = 1 - \frac{f_{\mathcal{M}}^{gt} \cdot f^{gen}}{|f_{\mathcal{M}}^{gt}||f^{gen}|}$$

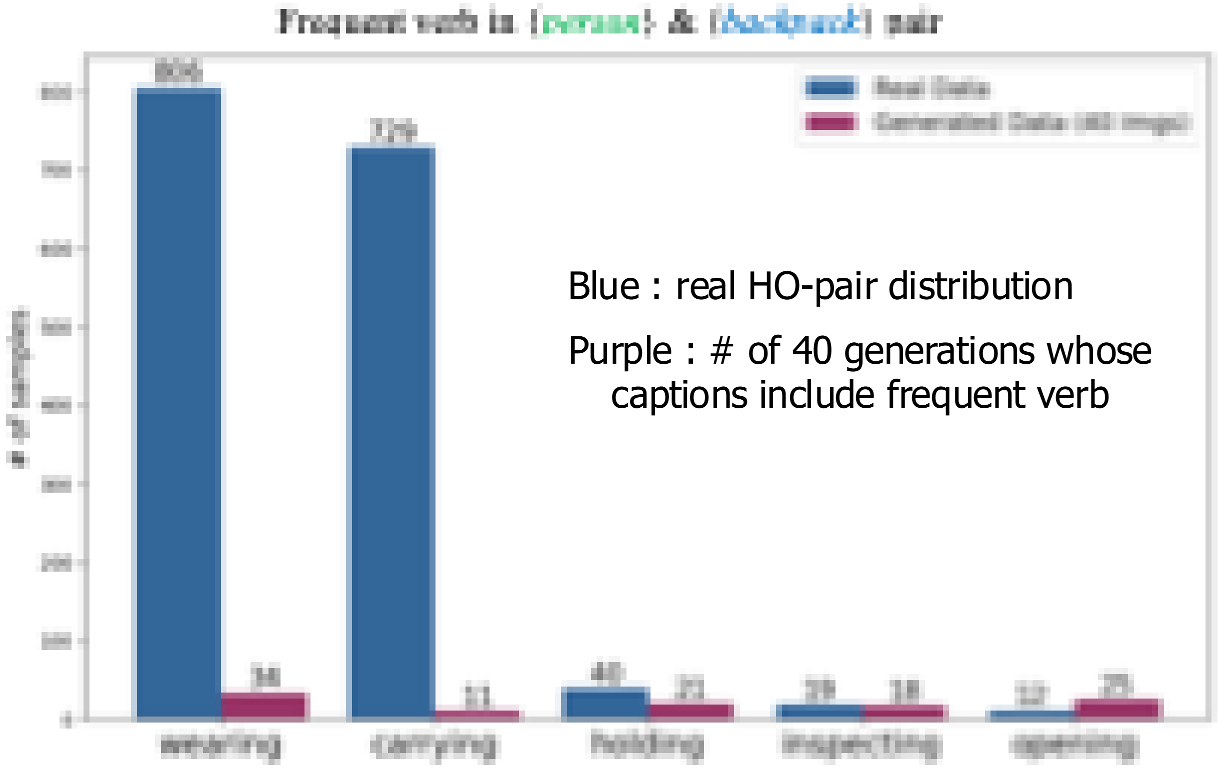
Frequency-based anchor interaction word

"a man **holding** a backpack"

SD

Captioner
(InstructBLIP)

"man **wearing** a backpack on his back"



Blue : real HO-pair distribution

Purple : # of 40 generations whose captions include frequent verb

SD tends to generate the images that matches with the most frequent verb!

Define the anchor interaction word as the most frequent verb along each h,o pair

$$r^{anc} = \arg\max_{r \in R_o} \mathcal{C}\,(r|o)$$

Adaptive Interaction Modification Number

Each h,o pair has different kinds of verbs and different number of images

The more the sample, the more the model effected

e.g., carrying backpack >> opening backpack

To **balance the modification extent**

$$\alpha(k) = \frac{1 - \beta^{n_k}}{1 - \beta}$$

$$\beta = \frac{N-1}{N}$$

n_k number of samples in class k
N: total number of samples in dataset

$$\mathcal{L}_{RDG} = \alpha \cdot (\mathcal{L}_{tiple} + \mathcal{L}_{align})$$

α : adaptive interaction modification number

Interaction Direction Guidance

w/ Relation Disentanglement Guidance (RDG)

The images **misalign with human expectations**, especially in the **local region**

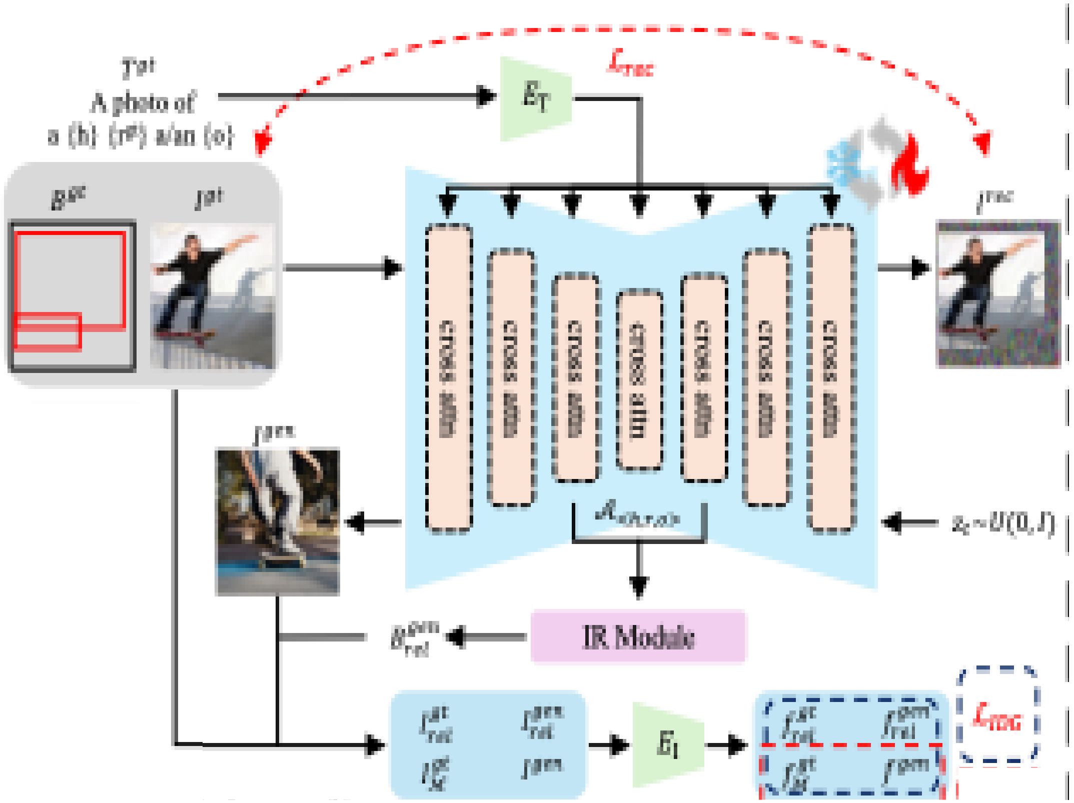
RDG focus on aligning global interaction feature

To focus on the finer, localized interaction region

Apply Interaction Direction Guidance(IDG)

Leveraging the interaction region extracted from Interaction Region module

Interaction Direction Guidance



Modify the CLIP direction loss into the interaction region loss

$$L_{IDG} = 1 - \frac{(f_{\mathcal{M}}^{gt} - f^{gen}) \cdot (f_{rel}^{bias})}{|f_{\mathcal{M}}^{gt} - f^{gen}||f_{rel}^{bias}|}$$

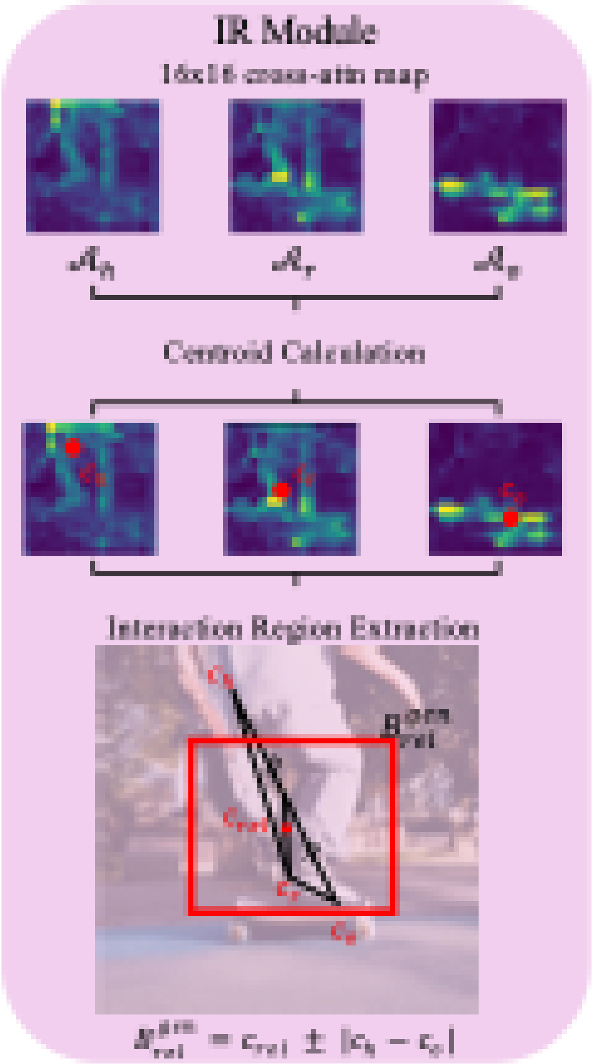
$$f_{rel}^{bias} = f_{rel}^{gt} - f_{rel}^{gen}$$

Can extract interaction region from ground-truth image

How to extract interaction regions from generated images?

Propose Interaction Region module (IR module)

Interaction Region Module



Cross-attention maps in SD reflect the existence of each token in prompts[9]

Extract 16x16 resolution map corresponding to each human, relation, object token

Calculate the centroid $c_{h,r,o}$ of each aggregated $\mathcal{A}_{h,r,o}$

$$c = \frac{1}{\sum_{h,w} \mathcal{A}} \begin{bmatrix} \sum_{h,w} w \cdot \mathcal{A} \\ \sum_{h,w} h \cdot \mathcal{A} \end{bmatrix}$$

Define c_{rel} to be the centroid of the triangle defined by three centroids

Extract interaction region B_{rel}^{gen}

$$B_{rel}^{gen} = c_{rel} \pm |c_h - c_o|_2^2$$

Training only the cross-attention layers in SD

$$\mathcal{L}_{rec} = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),t} [||(\epsilon \odot \mathcal{M} - \epsilon_\theta(z_t, t, T) \odot \mathcal{M})||]$$

$$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}_{rec} + \lambda_2 \cdot \mathcal{L}_{RDG} + \lambda_3 \cdot \mathcal{L}_{IDG}$$

[9] Hertz, Mokady, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626, 2022

Evaluations

Metric

1.CLIP & Sentence-BERT[10] similarity

2.HOI Classification Accuracy

3.VQA-Score[11]

1. Extract caption from generated image (InstrcutBLIP[12]) and calculate the embedding cosine similarity

2. HOI classification accuracy (K.O. : assume the object is correct, only compare the verb, Def. : both object and verb must be accurate)

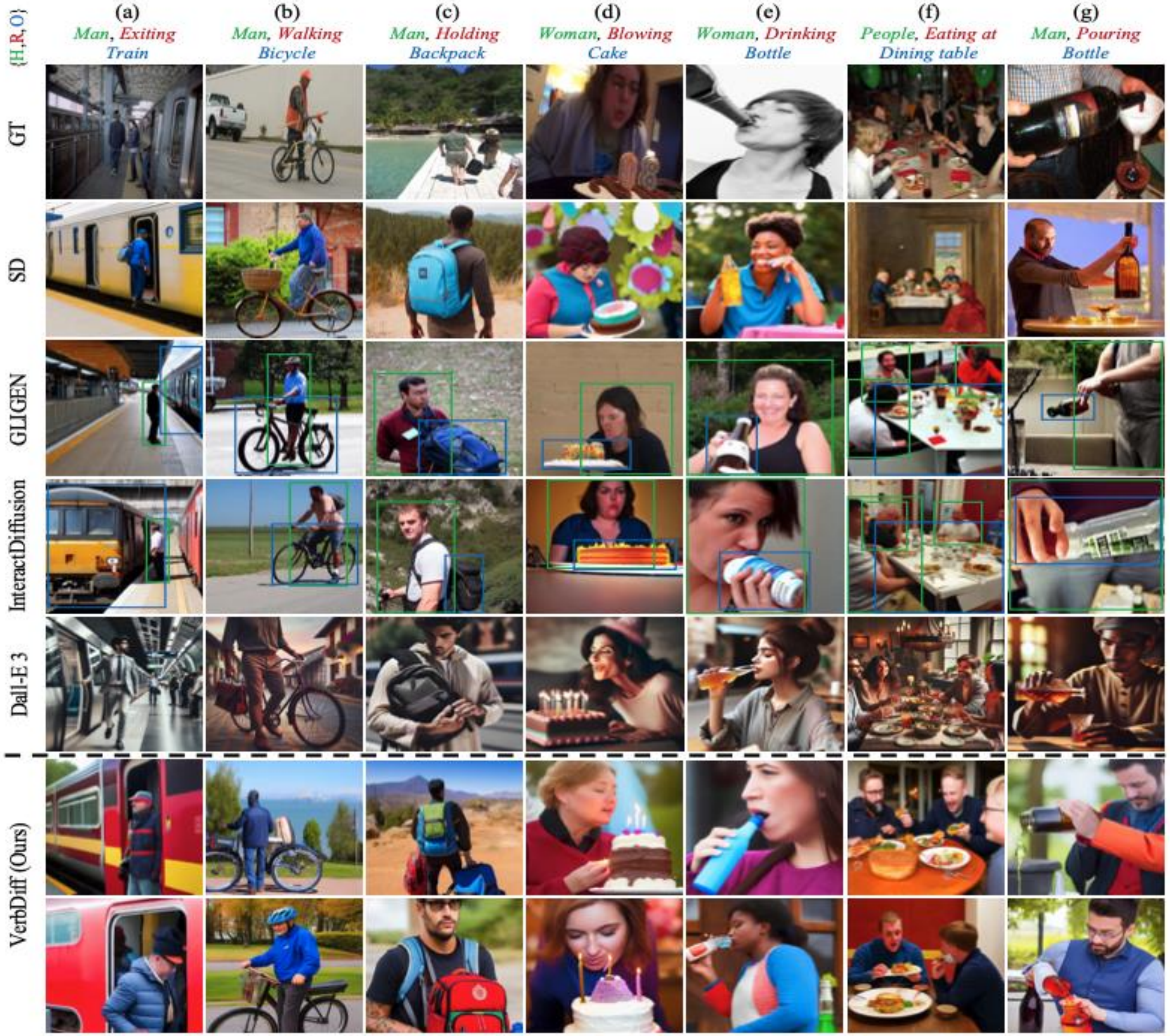
3. The average of probability that the model answer yes to a given question

[10] Reimers, N. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." *arXiv preprint arXiv:1908.10084* (2019).

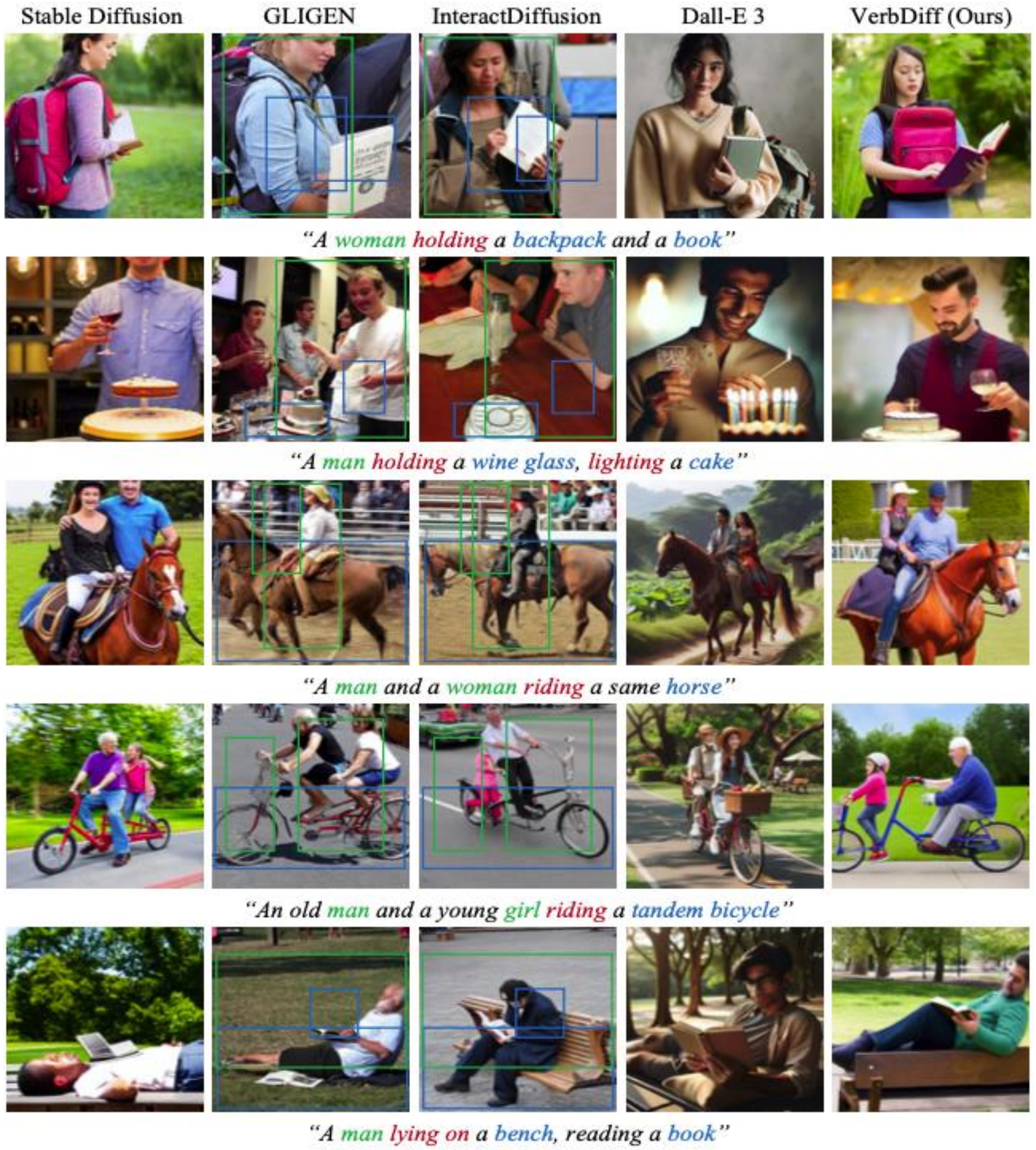
[11] Lin, Zhiqiu, et al. "Evaluating text-to-visual generation with image-to-text generation." *European Conference on Computer Vision*. Springer, Cham, 2025.

[12] Lin, Zhiqiu, et al. "Evaluating text-to-visual generation with image-to-text generation." *European Conference on Computer Vision*. Springer, Cham, 2025.

Single Interaction



Multi-Interactions



Experiments

Models	CLIP		S-BERT	VQA-Score	
	T2T	T2I	T2T	I2T	VQA
SD [25]	0.725	**0.251**	0.620	0.769	0.765
GLIGEN [15]	0.683	0.238	0.554	0.679	0.674
InteractDiffusion [10]	0.703	0.232	0.575	0.728	0.734
VerbDiff (Ours)	**0.733**	0.242	**0.633**	**0.771**	**0.766**

Table 1. **Similarity comparison between VerbDiff and other models.** We evaluate scores on CLIP, S-BERT [24] and a large vision-language alignment benchmark VQA-Score [17].

Models	SOV-STG-S (Acc ↑)				SOV-STG-Swin-L (Acc ↑)			
	Def.		KO.		Def.		KO.	
	Full	Rare	Full	Rare	Full	Rare	Full	Rare
SD [25]	16.09	4.59	18.22	4.85	20.08	8.07	21.69	8.66
GLIGEN [15]	15.88	4.85	17.91	5.24	17.83	7.00	19.35	7.57
InteractDiffusion [10]	19.67	7.00	21.31	7.69	23.53	10.27	24.86	11.18
VerbDiff (Ours)	**22.59**	**7.62**	**24.79**	**7.83**	**27.05**	**12.60**	**28.43**	**13.18**

Table 2. **HOI accuracy comparison between VerbDiff and previous methods.** Def. and KO. refer to Default and Known Object.

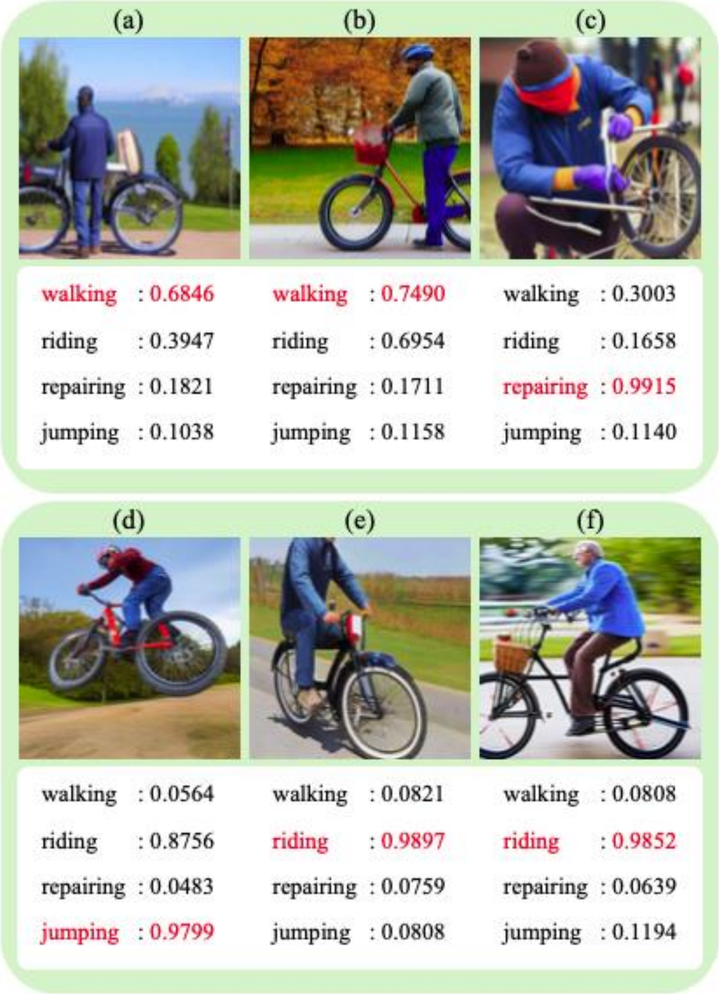
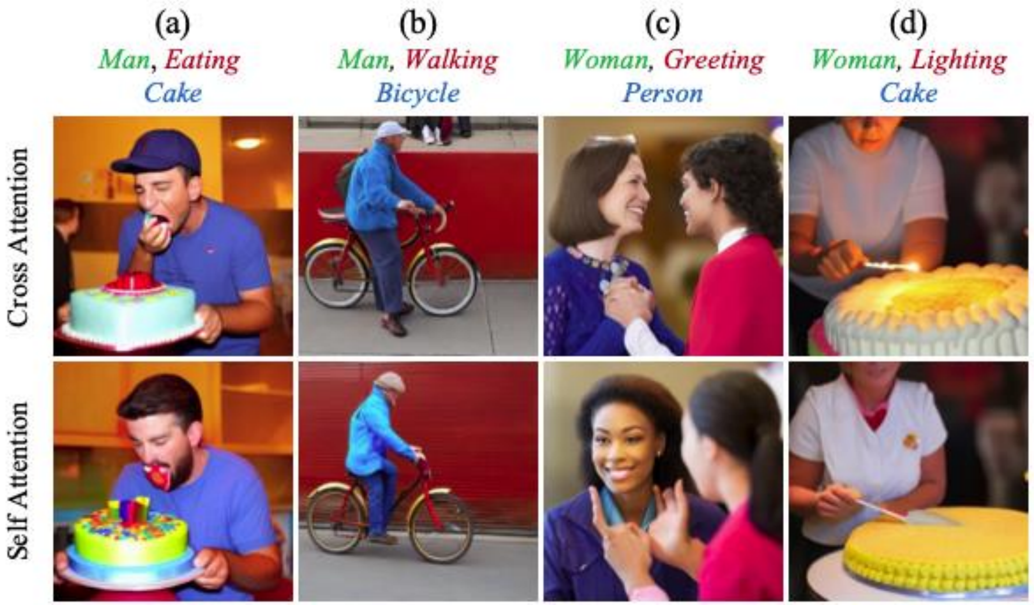


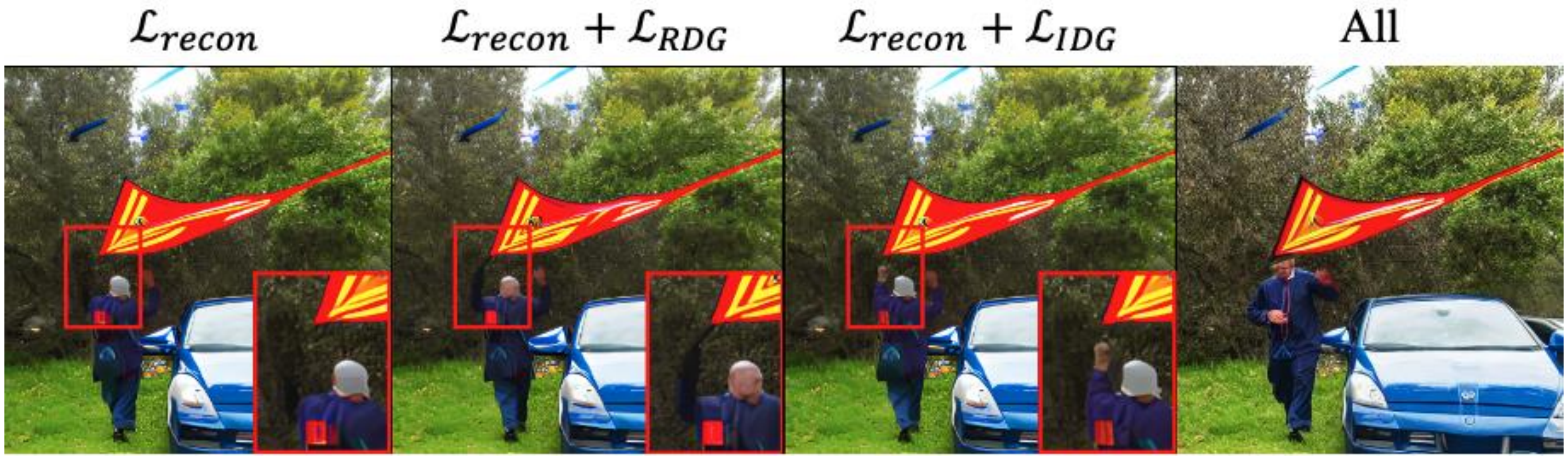
Figure 3. **VQA-score result examples.** We measure the probability that the VQA model answers "yes" for questions based on four verbs associated with the "bicycle" class. The verb with the highest score is highlighted in red.

Ablations

VerbDiff	$\mathcal{L}_{rec}$	$\mathcal{L}_{triple}$	$\mathcal{L}_{align}$	$\mathcal{L}_{IDG}$	CLIP T2T	S-BERT T2T	HOI Acc Def.	HOI Acc KO.
(a)	✓				0.691	0.582	19.38	20.89
(b)	✓	✓	✓		0.700	0.589	20.32	21. 87
(c)	✓			✓	0.699	0.588	20.21	21.67
(d)	✓	✓		✓	0.710	0.610	23.39	24.51
All (Ours)	✓	✓	✓	✓	**0.733**	**0.633**	**27.05**	**28.43**

Table 3. **Ablation of the VerbDiff guidance settings.** We score the similarity and HOI accuracy on the Full setting. We use the SOV-STG-Swin-L model. Combining all the proposed loss functions shows the best performance.





CVPR Nashville JUNE 11-15, 2025

Thank you.