

ReCorD: Reasoning and Correcting Diffusion for HOI Generation

Supplementary Material

Anonymous Author(s)

This supplementary material provides two templates, *i.e.*, Pose Selection Template and Interaction Template, as described in the main paper. We further offer the details of the conditioned spatial constraints used for the update of the latent, along with the complete algorithm. Additionally, we present how we conduct random subject augmentation to each verb and object pair from HICO-DET [6], VCOCO [7], and the non-spatial category of T2I-CompBench [23] to formulate the input prompt. Lastly, we display more qualitative comparisons and ablation studies of our ReCorD.

1 POSE SELECTION TEMPLATE

In \mathcal{M}_r of our method, we query the VLM-based Pose Selection Agent with previously generated candidates and an input prompt, utilizing the capability of VLMs to process visual information. Below, we provide the template to form the input prompt:

Pose Selection Template:

Given the input images and the prompt, y , which picture contains the most possible pose for the given action? Please answer by number.

2 INTERACTION TEMPLATE

When engaging the Layout Agent, we integrate detected key points \mathcal{P} , the human’s bounding box b_h , and the object’s location b_o into the Interaction Template. Additionally, the image selected by the Pose Selection Agent is provided as supplementary data for the VLM to deliberate on when suggesting the layout. In constructing the Interaction Template, we explicitly design guidelines and procedures for the VLMs to follow. We also enhance the logical flow by directing the VLMs to analyze human postures’ visual attributes using the Chain-of-Thought method [57] and in-context learning examples. A simplified version of the Interaction Template is presented below, with the complete version available in Table S2.

Interaction Template:

Your Role: Expert Human Pose Analyst

Objective: Think step by step, your task is analyzing key points of human pose in square images according to the user’s prompt and manipulating the bounding boxes of objects to the correct locations while maintaining visual accuracy.

[Key Guidelines + Process Steps + In-context Examples]

Your Current Task: Carefully follow the provided guidelines and steps closely to accurately identify the human pose based on . . .

User Prompt: *Input Prompt* y , *Key points*: \mathcal{P} , *Original Human Location*: b_h , *Original Object Location*: b_o .

Reasoning:

3 DETAILED FORMULATION FOR CONDITIONED SPATIAL CONSTRAINTS

The mentioned loss function [59] is crafted to constrain the generative image, ensuring that the cross-attention map of the object token shows sufficiently strong values within the specified bounding box \hat{b}_o . Since such an objective involves three terms: inner-box, outer-box, and corner constraints, we elaborate on the corresponding conditions in the following paragraphs.

Inner-box Constraint The target of \mathcal{L}_{IB} is to regularize the construction so that objects would approach the mask regions. A straightforward solution is to ensure that the objects align with the desired position, *i.e.*, a series of binary masks $\mathbf{M} = \{M_i\}$ and that high responses from cross-attention maps occur only within the mask regions. To accomplish such a purpose, we can compose the objective as

$$\mathcal{L}_{IB} = \sum_{i=1}^N [1 - \frac{1}{K} \sum (\mathcal{A} \cdot M_i)_K], \quad (\text{S1})$$

where $(\cdot)_K$ represents the top- K selection that collects K highest magnitude of resulting representations.

Outer-box Constraint In contrast to the previous constraint, \mathcal{L}_{OB} aims to penalize the model when the attention maps extend beyond the specified area, thereby preventing the object from moving out of the target regions. Accordingly, \mathcal{L}_{OB} is defined as

$$\mathcal{L}_{OB} = \sum_{i=1}^N [\frac{1}{K} \sum (\mathcal{A} \cdot (\mathbb{1} - M_i))_K], \quad (\text{S2})$$

where $\mathbb{1}$ is the matrix containing all elements equal to 1.

Corner Constraint As \mathcal{L}_{IB} and \mathcal{L}_{OB} regularize the position and hold limited spatial conditions, we introduce \mathcal{L}_{CC} specified for corner restrictions. Ideally, we uniformly retrieve L samples from the margin between M_i and \mathcal{A} in the embedding space around the corner coordinates. The objective that corresponds to the axis is optimized as

$$\mathcal{L}_x = \sum_{i=1}^N [\frac{1}{L} \sum \text{U}(\|\xi_x(M_i) - \xi_x(\mathcal{A})\|, L, x_{min}, x_{max})] \quad (\text{S3})$$

$$\mathcal{L}_y = \sum_{i=1}^N [\frac{1}{L} \sum \text{U}(\|\xi_y(M_i) - \xi_y(\mathcal{A})\|, L, y_{min}, y_{max})], \quad (\text{S4})$$

where $\xi_x(\cdot) \in \mathbb{R}^W$ and $\xi_y(\cdot) \in \mathbb{R}^H$ project the mask M_i and cross-attention map \mathcal{A} using the max operator along x-axis and y-axis, respectively. We aim to bring $\xi(M_i)$ close to $\xi(\mathcal{A})$ by this objective. Eventually, the corner constraint can be represented as

$$\mathcal{L}_{CC} = \mathcal{L}_x + \mathcal{L}_y. \quad (\text{S5})$$

By coupling these constraints within our correction mechanism, we can preserve poses and avoid attention overlapping. This superiority gives ReCorD adjustment ability and integrates the reasoning capabilities of VLM to render interactions effectively.

4 INTERACTION-CORRECTING ALGORITHM

Given an optimal pose and layout proposed by \mathcal{M}_r , \mathcal{M}_c combines these to produce accurate HOI images. Our complete algorithm in \mathcal{M}_c can be derived as follows:

Algorithm 1: Interaction-Correcting Algorithm.

```

1 Input: A full prompt  $y$  with  $N$  tokens, an intransitive
   prompt  $\tilde{y}$ , an object token index  $m$ , a seed  $s$  selected and an
   object bounding box  $\tilde{b}_o$  proposed by  $\mathcal{M}_r$ .
2 Output: Interaction-corrected image latent  $z_0$ .
3  $z_T \sim \mathcal{N}(0, I)$  a Gaussian noise sampled with seed  $s$ 
4 for  $t \in [T_2, \dots, 0]$  do
5    $A \leftarrow \text{LDM}(z_t, y, t, s)$ 
6    $\tilde{A} \leftarrow \text{LDM}(z_t, \tilde{y}, t, s)$ 
7    $\mathcal{A}_{\text{cross}}, \mathcal{A}_{\text{self}} \leftarrow$  manipulating  $A, \tilde{A}$  by eq. (2)
8    $\tilde{A}_m = \mathbb{1} - A_m$ 
9   for  $n \in [1, \dots, N]$  do
10    if  $n \neq m$  then
11       $\hat{A}_n \leftarrow \tilde{A}_m \odot A_n$ 
12    end
13     $\hat{z}_t \leftarrow z_t - \alpha_t \cdot \nabla \mathcal{L}$ 
14     $z_{t-1} \leftarrow \text{LDM}(\hat{z}_t, y, t, s)$ 
15 end
16 Return  $z_0$ 

```

5 DETAILS OF AUGMENTATION

HICO-DET and VCOCO For HICO-DET and VCOCO, we derive HOI triplets and diversify the subjects to include men, women, and individuals of various ages, providing a total of 15 variations. Ultimately, this process yields 7,650 HOI prompts for HICO-DET and 2,550 for VCOCO. The complete list of subjects is as follows:

Subject Augmentation of HICO-DET & VCOCO:

“man, woman, boy, girl, old man, old woman, teenager, child, young man, young woman, adult, kid, elderly person, middle-aged person, toddler”

T2I-CompBench Non-Spatial Relationship Category For T2I-CompBench, we initially remove any prompts that do not involve HOIs, such as those including dogs, cats, *etc.*, to align with our experimental framework. We then retain prompts that use careers as subjects, such as mechanics or musicians. Next, we standardize these prompts by converting all subject references to ‘person’ and eliminating duplicate HOIs to maintain the variety of our selected prompts. We enrich our dataset through augmentation sequentially, resulting in a collection of 465 diverse prompts.

Subject Augmentation of T2I-CompBench:

“person, man, woman, child”

6 MORE QUALITATIVE COMPARISON

In Figure S2, we present additional comparison between ReCorD and the methods A&E [7] and LMD [30]. ReCorD excels in these

methods by precisely generating human poses and object placements corresponding with text prompts, demonstrating its ability to render object interactions with great precision. On the other hand, A&E and LMD frequently misplace objects or struggle to grasp the intended actions’ subtleties fully. A&E, although capable of generating multiple objects, falls short in capturing complex interactions, which is evident in (a), (d), and (e). Furthermore, despite the help from the SDXL’s refiner module, LMD’s inclination to emphasize nouns rather than verbs undermines its ability to depict interactions accurately.

As demonstrated in Figure S3, we conduct the experiments between ReCorD and other baseline methods on T2I-CompBench. Considering inputs are solely textual, we exclude L2I-based methods in this experiment. For MultiDiffusion [4], we leverage its ability of text-to-panorama while restricting the image dimensions to 512×512 . However, A&E and MultiDiffusion have difficulties generating interactions derived from the erroneous priors inherent in pre-trained diffusion models, similar to those observed in SD [48]. Even with the support of LLM for layout generation, LayoutLLM-T2I [43] and LMD fail to adequately address the critical aspect of postures, focusing primarily on object creation within the generated layouts. Particularly, both SDXL [42] and DALL-E 3 struggle with precise object placement and managing the correct number of objects.

7 MORE ABLATION STUDIES

Impact of Inverse Mask \tilde{A}_m Figure S1 demonstrates the effectiveness of our correction mechanism in eliminating attention overlapping issues. In HOI scenes, the human and the object often share overlapping regions. As shown in the scenario without applying element-wise product operation on \tilde{A}_m (w/o \tilde{A}_m), the absence of \tilde{A}_m leads to attention overlapping issues, resulting in the fusion of human and object in the generated images. On the other hand, by applying element-wise product operation on \tilde{A}_m (w/ \tilde{A}_m), our correction mechanism ensures the successful generation of objects, as it mitigates the attention overlapping between humans and objects



Figure S1: Ablation study of inverse mask \tilde{A}_m . The element-wise product operation on \tilde{A}_m resolves the attention overlapping issues in HOI generation.

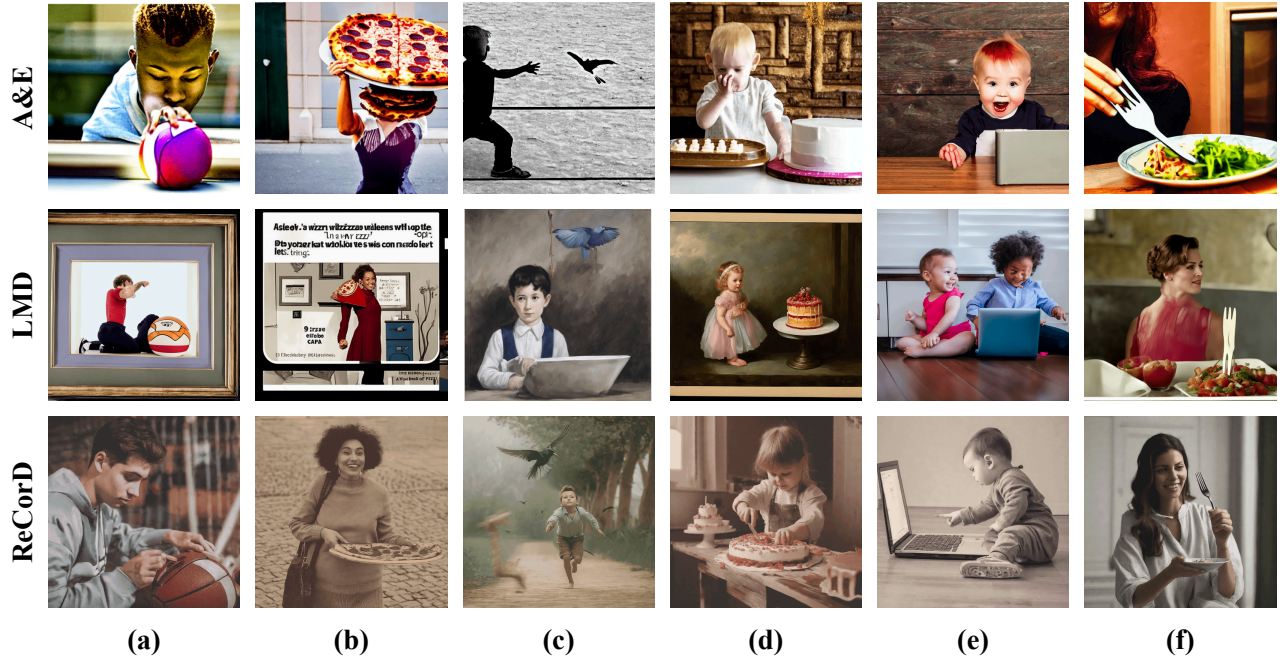


Figure S2: Visual comparison to other existing benchmarks for HICO-DET and VCOCO using various text prompts, ReCorD excels by providing clearer delineation of interaction and generating images that accurately reflect the given text instructions. (a) a young man is signing a sports ball. (b) a woman is carrying a pizza. (c) a boy is chasing a bird. (d) a child is cutting a cake. (e) a toddler is pointing at a laptop. (f) a woman is holding a fork.

during correcting process in ReCorD.

Impact of γ in Self-Attention Map Modulation As shown in Table S1, we compare the impact of parameter γ on VCOCO dataset. Parameter γ determines when to execute self-attention map modulation (once denoising steps $t > \gamma$). We observe that increasing γ enhances pose preservation, but it reduces the quality of images. Conversely, if the γ value is too small, the chosen pose cannot be maintained. Therefore, we select an appropriate γ value of 5, which allows ReCorD to preserve the pose while enabling the generation of HOI details.

Table S1: Impact of adjustment parameter γ .

	$S_{\text{CLIP}} \uparrow$	$S_{\text{CLIP}}^{\text{verb}} \uparrow$	PickScore \uparrow	FID \downarrow	HOI \uparrow
$\gamma = 5$	31.94	21.84	22.22	60.74	22.48
$\gamma = 10$	31.33	22.11	21.89	64.33	21.20
$\gamma = 15$	31.00	19.86	21.69	74.30	19.23

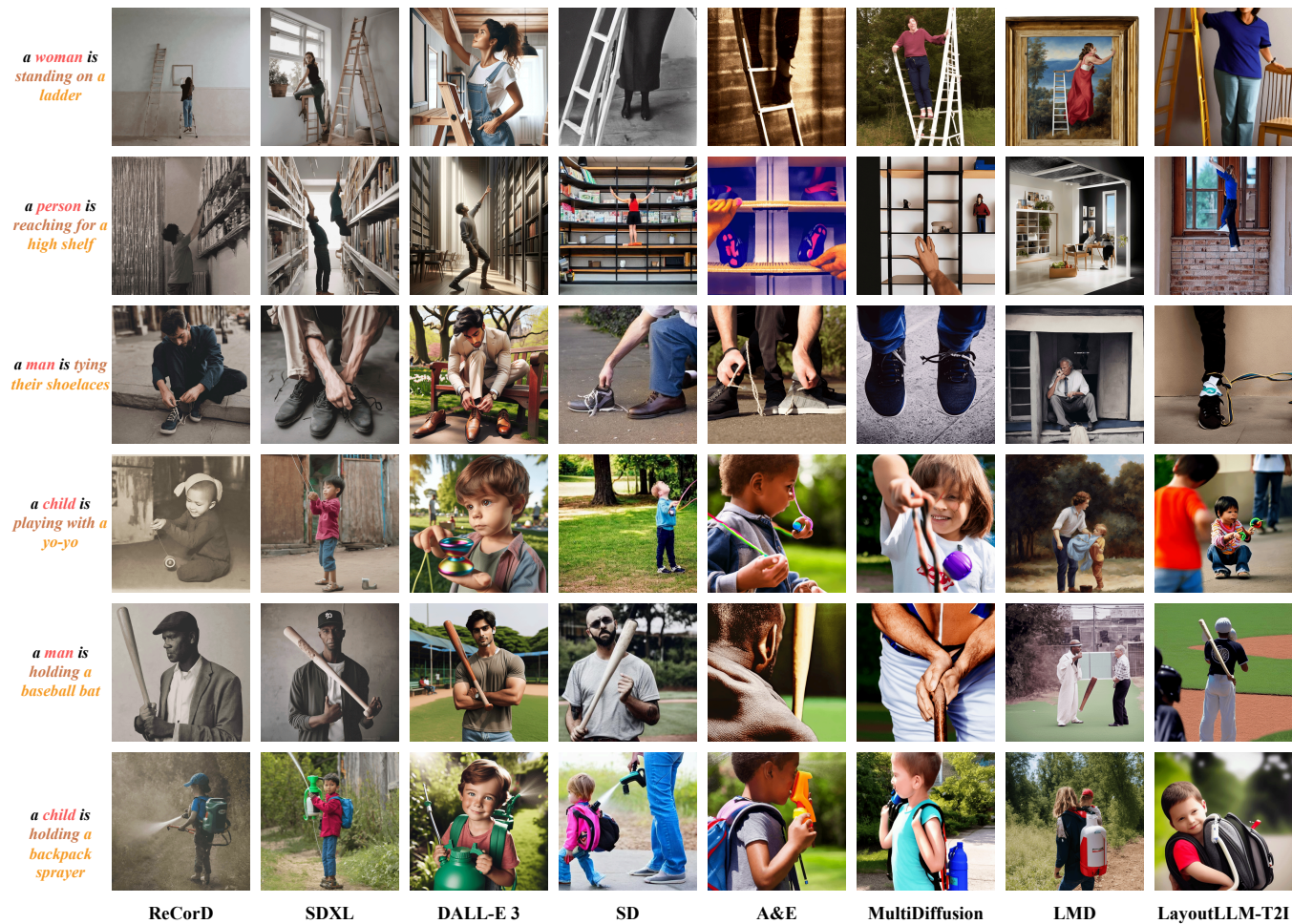


Figure S3: Visual comparison against existing benchmarks for T2I-CompBench using various text prompts, ReCorD achieves improved delineation of interaction and produces images that closely match the text instructions.

Table S2: Our full prompt for the Interaction Template.

```

1 # Your Role: Expert Human Pose Analyst
2
3 ## Objective: Think step by step, your task is analyzing keypoints of human pose in square images according to the
4 user's prompt and manipulating the bounding boxes of the object to the correct locations while maintaining visual
5 accuracy.
6
7 ## Human Pose Key Points and Bounding Box Specifications and Analysis
8 1. Image Coordinate: Define square images with top-left at [0, 0] and bottom-right at [512, 512].
9 2. Annotations of Key Points: ["nose", "left eye inner", "left eye", "left eye outer", "right eye inner", "right eye",
10 "right eye outer", "left ear", "right ear", "mouth left", "mouth right", "left shoulder", "right shoulder", "left
11 elbow", "right elbow", "left wrist", "right wrist", "left pinky", "right pinky", "left index", "right index",
12 "left thumb", "right thumb", "left hip", "right hip", "left knee", "right knee", "left ankle", "right ankle",
13 "left heel", "right heel", "left foot index", "right foot index"]
14 3. Box Format: [Top-left x, Top-left y, Bottom-right x, Bottom-right y]
15 4. Object Size: The object's bounding box size is represented as a fraction of the image size.
16 5. Results of Analysis: Pose Types: ["Static", "Dynamic"], Body Orientation: ["Frontal", "Backward", "Profile",
17 "Angled"], Facial Direction: ["Directly at Viewer", "Looking Upwards", "Looking Downwards", "Looking Sideways",
18 "Looking at objects"], Object Relationship: ["Above human", "Under human", "Beside human", "In front of human",
19 "Behind human", "On human", "Near human"], Object Size: ["one-tenth", "one-fifth", "three-tenths", "two-fifths",
20 "one-half", "three-fifths", "seven-tenths", "four-fifths", "nine-tenths", "one"], Object Location: [,,,]
21
22 ## Key Guidelines
23 1. Alignment: Follow the user's prompt, keeping the attributes of the specified object.
24 2. Boundary Adherence: Keep the bounding box coordinate within [0, 512].
25 3. Visual Accuracy: Ensure the object's bounding box size is visually accurate and aligned with the human pose.
26 4. Minimal Modifications: Change the bounding box of the object only if it doesn't match the scene affordances.
27 5. Human Location Constraints: The human bounding box should not be altered.
28 6. Overlap Reduction: Minimize intersections of all the bounding boxes.
29
30 ## Process Steps
31 1. Interpret Prompts: Read and understand the user's prompt.
32 2. Key Points Analysis: Identify all key points of the person and perform pose estimation to understand the spatial
33 relationships between different key points.
34 3. Implement Changes: Review and adjust the current bounding box of object while considering the interaction and scene
35 affordances.
36 4. Explain Adjustments: Justify the reasons behind the alteration and ensure the adjustment abides by the key
37 guidelines.
38 5. Output the Results: Present the analysis and predict the updated absolute coordinates of the object's bounding box,
39 which should include a list of bounding boxes in Python format.
40
41 ## Examples
42 - Example 1
43 User Prompt: a woman is sitting on a horse. Key Points: [[228, 87], [232, 81], [234, 81], [235, 82], [228, 80],
44 [228, 79], [227, 78], [242, 83], [232, 80], [232, 95], [228, 94], [254, 118], [233, 113], [264, 173], [234, 163],
45 [236, 202], [223, 194], [233, 219], [219, 200], [227, 211], [217, 201], [230, 206], [220, 200], [264, 201], [243,
46 200], [233, 246], [206, 263], [239, 309], [169, 373], [244, 331], [168, 392], [226, 352], [145, 396]], Original
47 Human Location: [200, 43, 278, 358], Original Object Location: [120, 46, 452, 389].
48 Reasoning: Here, there is one woman and one horse.
49 Pose Types: "Dynamic"
50 Body Orientation: "Profile"
51 Facial Direction: "Looking Sideways"
52 Object Relationship: "Under human"
53 Object Size: "two-fifths"
54 Updated Object Location: [101, 131, 433, 474]
55
56 Your Current Task: Carefully follow the provided guidelines and steps closely to accurately identify the human pose
57 based on the given prompt and adjust the bounding boxes in accordance with the user's prompt. Ensure adherence to
58 the above output format.
59
60 User Prompt: {{ GT_prompt }}. Key Points: {{ kps }}, Original Human Location: {{ ori_h_bbox }}, Original Object
61 Location: {{ ori_o_bbox }}.
62 Reasoning:

```
