Human-Object Interaction Detection Collaborated with Large Relation-driven Diffusion Models

Liulei Li[1], **Wenguan Wang**[2]*, **Yi Yang**[2]

[1]ReLER, AAII, University of Technology Sydney [2]CCAI, Zhejiang University

https://github.com/Oliliulei/DiffusionHOI

Abstract

Prevalent human-object interaction (HOI) detection approaches typically leverage large-scale visual-linguistic models to help recognize events involving humans and objects. Though promising, models trained via contrastive learning on text-image pairs often neglect mid/low-level visual cues and struggle at compositional reasoning. In response, we introduce DIFFUSIONHOI, a new HOI detector shedding light on text-to-image diffusion models. Unlike the aforementioned models, diffusion models excel in discerning mid/low-level visual concepts as generative models, and possess strong compositionality to handle novel concepts expressed in text inputs. Considering diffusion models usually emphasize instance objects, we first devise an inversion-based strategy to learn the expression of relation patterns between humans and objects in embedding space. These learned relation embeddings then serve as textual prompts, to steer diffusion models generate images that depict specific interactions, and extract HOI-relevant cues from images without heavy fine-tuning. Benefited from above, DIFFUSIONHOI achieves SOTA performance on three datasets under both regular and zero-shot setups.

1 Introduction

As a crucial topic in the field of visual scene understanding, human-object interaction (HOI) detection demands not only inferring the semantics and locations of entities but also should comprehend the ongoing events happening between them[1, 2]. Given the complexity and diversity of human activities in object-rich realistic scenes, this task presents challenges in long-tailed distributions and zero-shot discovery[3]. A set of studies seek to tackle these two issues by leveraging large-scale visual-linguistic models (*e.g.*, CLIP[4]) which show strong generalization ability on dozens of tasks. Though strides made, it has been observed that models trained by aligning high-level text-image semantics face difficulties in discerning spatial locations[5], and struggle at compositionality[6] which is a fundamental ability for human to capture new concepts by combining known parts. In fact, both middle-level visual cues (*e.g.*, spatial relation) and compositionality are essential facets for HOI detection. The former can help deduce feasible interactions according to locations between instances, while compositionality contributes significantly to zero-shot generalization. For example, we can easily understand `human-hold-horse` by composing `human-hold-dog` and class `horse` that have encountered previously.

In contrast, the text-to-image diffusion models[7–14] also pre-trained on large-scale image-text pairs, are demonstrating superior capabilities outperforming models like CLIP. Concretely, they are able to generate diverse high-quality images conditioned on textual inputs, showing proficiency in understanding ***high-level semantics***[15, 16]. In addition, the generated images convey reasonable shape, texture, layout, and structure, indicating the comprehension in ***mid/low-level visual concepts*** as generative models[17]. More importantly, the descriptions are typically organized in a compositional

*Corresponding Author: Wenguan Wang.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

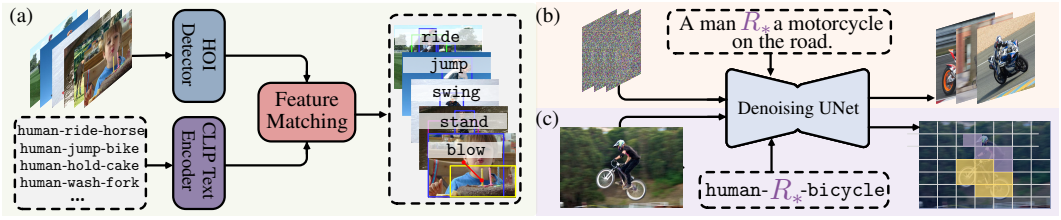

Figure 1: Existing solutions utilize mere linguistic knowledge (a). Our solution utilizes both text-prompt image generation (b) and conditioned feature extraction (c) abilities of diffusion models for knowledge transfer.

manner, with phrases such as "happy","near a bridge", or "hugged by a man" continually appended to objects like "a dog". This suggests that diffusion models inherently possess ***compositionality***, to systematically adapt to newly encountered user requirements by composing known visual concepts.

The above analysis motivates us to explore diffusion models for HOI detection. Nonetheless, to fully unlock the potential of diffusion models and accommodate the unique characteristic of HOI detection task, the following questions naturally arise: ❶ With diffusion models typically emphasizing instance generation, how to steer it to prioritize the relationships between humans and objects? ❷ How to transfer the extensive knowledge obtained from large-scale pre-training in diffusion models to assist the recognition of interactions? To address ❶, we harness textual inversion [18] which conceptualizes a user-provided object by inverting it to a text embedding. However, this method focuses solely on instance objects. To facilitate a smooth shift from object-centric to *relation-centric* modeling, we devise a human-object relation inversion strategy grounded in the disentanglement of HOI. Concretely, given the HOI latent describing `human-action-object`, we build a cycle-consistency objective to reconstruct it from a intermediate relation latent derived from the original HOI latent. This reconstruction process is guided by a set of learnable relation embeddings as text prompts, for which we use the placeholder R_* to denote the textual form before encoded into embedding space. These relation embeddings further involves in a relation-centric contrastive learning to enhance the awareness of high-level relational semantics. To answer ❷, we leverage both the text promoted image generation and conditioned feature extraction abilities of diffusion models. We realize *relation-driven* image generation by compositionally organizing R_* with other linguistic elements to formulate new text prompts (Fig. 1(b)). This allows for the generation of novel interactions with unseen objects, and extends the training set for HOI detectors. Moreover, we directly utilize diffusion models as backbone to extract HOI-relevant features conditioned on R_* (Fig. 1(c)). After a single noise-free forward step, features distinct for each interaction can be obtained. Finally, to establish a loop for mutual boosting between above *relation-inspired* HOI detection and relation modeling, we devise an online update strategy to facilitate the continual evolving of relation embeddings during HOI detection learning.

Benefited from controllable image generation and knowledge transfer from diffusion models, our method named DIFFUSIONHOI enjoys several appealing advantages: **First**, it steers diffusion models to focus on complex relationships rather than single objects in an efficient way. This offers a robust foundation for HOI modeling. **Second**, from the perspective of relation-driven, it unlocks the image generation power of diffusion models tailored for the HOI detection task. This enriches the pool of training samples, particularly for long-tailed/unseen interaction classes. **Third**, the relation-inspired prompting improves both the flexibility and accuracy of HOI detectors. It adapts to each individual image to extract action or object related cues, while CLIP-based methods[3, 19] produce action/object features merely from texts (*i.e.*, Fig. 1(a)), remaining static and unresponsive to image content.

By embracing text-to-image diffusion models as well as facilitating relation-driven image generation and prompting, our method demonstrates superior performance. It surpasses all top-leading solutions on HICO-DET [20] and V-COCO [21], and sets new state-of-the-arts. In addition, it yields up to **6.43%** mAP improvements on SWiG-HOI [22] under the zero-shot HOI discovery setup. These promising performance evidences the great potential of integrating diffusion models for visual relation understanding. We hope this work could foster the broader exploration of large-scale pre-trained diffusion models on more computer vision tasks beyond mere image generation.

2 Related Work

Human-Object Interaction Detection. According to the architecture design of networks, existing solutions for HOI detection can be broadly categorized into two groups: one-stage and two-stage.

The one-stage methods [23–26] typically employ a multi-task learning pipeline that jointly undertake the tasks of human-object detection and interaction classification in an end-to-end manner, therefore distinguished by fast inference. In contrast, two-stage methods [27–37] first detect entities with off-the-shelf detectors such as Faster R-CNN [38], and the predict the dense relationships among possible human-object pairs. This paradigm effectively disentangles the HOI detection process and results in improved performance. Inspired by DETR [39], recent advancements shift to adopt Transformer-based architectures [3, 40–46]. Several studies [3, 47–51] also supplement the Transformer-based HOI detectors with large-scale visual-linguistic models like CLIP [4] or visual knowledge [52, 53] to conduct logic-induced reasoning [54]. However, these models focus solely on aligning high-level semantics and overlooking mid/low-level visual cues. To tackle this, we redirect our attention to diffusion models, which perfectly address the aforementioned challenges and possess the capacity to handle previously unseen concepts through their strong compositionality.

Controllable Image Generation. To facilitate customized image generation with respect to predefined class, attribute, text or image [55], various approaches based on GANs [56] have been proposed. For instance, [57, 58] develop a photo realistic hairstyle transfer method through latent space optimization. However, these methods typically show limited diversity when compared to likelihood-based models [59]. In response, diffusion models [60–62] have emerged that not only demonstrate remarkable synthesis quality but also offer enhanced controllability. The core idea behind is to transform a simple and known distribution (*e.g.*, Gaussian) into the desired data distribution. These models have proven to be highly effective in various conditional scenarios. According to the conditional targets, the prevalent work can be grouped into class-driven [63, 64] text-driven [7, 8], exemplar image-driven [65, 66], *etc.*. These advances have found application in a wide range of domains such as super resolution [66, 67], image editing [13, 68]. Recently, a new approach achieves guided image generation by learning a single word embedding through a frozen text-to-image model to properly describe the desired target objects [18]. Take inspiration from it, we achieve relation-driven image generation by extending such object-centric concept modeling approach to relation-centric.

Knowledge Transfer from Diffusion Models. In light of the notable success achieved by diffusion models in applications, there is a growing interest in transferring knowledge acquired from large-scale pre-training to various tasks [17, 69–75]. For example, given the limited availability of data for constructing NeRFs and the unprecedented generalizability of diffusion models, researchers are motivated to explore generating 3D NeRFs via a 2D text-to-image diffusion model using diverse input text [69–71]. More recently, a notable trend has emerged where efforts are dedicated towards learning semantic representations from diffusion models by extracting intermediate feature maps. It finds diverse application in image segmentation [17], semantic correspondence learning [72–74], and general representation learning [75]. In this work, we extensively harness both the image generation and semantic representation abilities of diffusion models, by using relation-centric embeddings to control the generation and prompt semantic extraction from images with respect to specific interactions.

3 Methodology

3.1 Preliminary: Textual Inversion

Latent diffusion models [14] represent an evolution of diffusion models which offer significant enhancements in both computational and memory efficiency by executing denosing in the latent space. It comprises two primary components. The first is a pre-trained generator equipped with an encoder $\mathcal{E}$ to map the input image x into a latent vector $\boldsymbol{z} = \mathcal{E}(x)$, from which the original data can be reconstructed via a decoder $\mathcal{D}$ by $\hat{x} = \mathcal{D}(\boldsymbol{z}) \approx x$. The second is a diffusion model to generate latent codes $\boldsymbol{z}$ conditioned on user guidance y which can be text, image, *etc.*The latent codes then serve as inputs to $\mathcal{D}$ for image generation *w.r.t.* y. The training objective is given as:

$$\mathcal{L}_{\text{LDM}} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(\boldsymbol{z}_t, t, c_\theta(y))\|_2^2\Big], \tag{1}$$

where c_θ is a conditioning model to encode y, $\boldsymbol{z}_t$ is the noised latent at time t, ϵ is sampled noise, ϵ_θ is the denoising network. Based on latent diffusion models, inversion-based diffusion [18] seeks to learn a text embedding v_* that accurately describes novel concepts in user provided images. This is achieved by optimizing v_* with Eq. 1 to iteratively reconstruct the latent code $\boldsymbol{z}$ of user provided images with text prompts y like "an image of S_*", where S_* is the placeholder of new concept:

$$v_* = \arg\min_v \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(\boldsymbol{z}_t, t, c_\theta(y))\|_2^2\Big]. \tag{2}$$

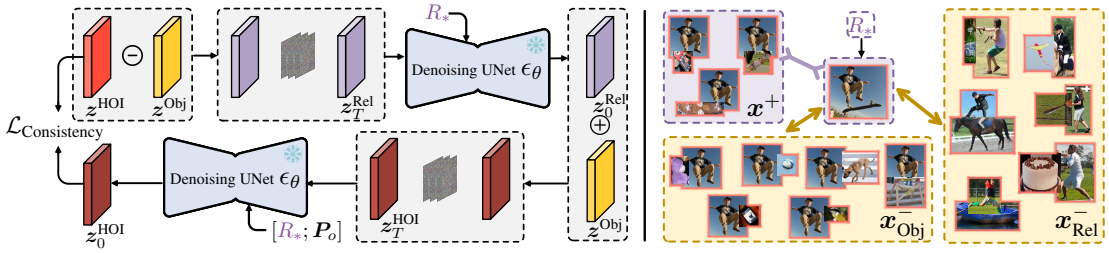

Figure 2: (Left) Disentanglement-based cycle-consistency learning. (Right) Relation-centric contrastive learning.

As such, it enables image generation *w.r.t.* target concepts in diverse scenes by using the learned embedding v_* to replace the tokenized placeholder S_* in text prompts.

3.2 Inversion-Based HOI Modeling

Disentanglement-based Relation Embedding Learning. To facilitate above inversion technology for relation modeling, two options present: **i)** directly optimizing embeddings describing interactions (*i.e.*, `human-action-object`), which risks overfitting with limited samples for long-tailed categories and cannot generalize to novel concepts, and **ii)** learning `action` embeddings with diverse images sharing a common action but different objects, which seems feasible but poses significant convergence issues due to the complex content, and the optimization target cannot be fixed to actions but not other unrelated elements. In contrast, drawn from the compositional nature of HOI, we adopt a disentangled solution (*i.e.*, Fig. 2) where HOI triplets are broken into `human-action` and `object`. Here `human-action` is considered to describe the relation between `human` and `object`, as `action` is executed by and strictly adheres to `human` involved. Then, denoting the text describing `human-action` as R_*, encoded relation embeddings as $v_*^{\text{Rel}} = c_\theta(R_*)$, and the latent of one happening HOI in image as $\boldsymbol{z}^{\text{HOI}}$, a relation latent $\boldsymbol{z}_0^{\text{Rel}}$ could be reconstructed (*i.e.*, denoising with ϵ_θ from time T to 0) by:

$$\epsilon_\theta((\boldsymbol{z}^{\text{HOI}} - \boldsymbol{z}^{\text{Obj}})_T, T, v_*^{\text{Rel}}) \rightarrow \boldsymbol{z}_0^{\text{Rel}}. \tag{3}$$

Here $(*)_T$ is the noised version at time T, and $\boldsymbol{z}^{\text{Obj}}$ is retrieved by encoding the cropped object from image with provided bounding box annotations. We consider $\boldsymbol{z}^{\text{HOI}} - \boldsymbol{z}^{\text{Obj}}$ is able to describe the `human-action` component by subtracting the `object` from `human-action-object`. Then, we can reconstruct the latent representing the complete HOI image by adding $\boldsymbol{z}^{\text{Obj}}$ back to $\boldsymbol{z}_0^{\text{Rel}}$:

$$\epsilon_\theta((\boldsymbol{z}_0^{\text{Rel}} + \boldsymbol{z}^{\text{Obj}})_T, T, [v_*^{\text{Rel}}; \boldsymbol{P}_o]) \rightarrow \boldsymbol{z}_0^{\text{HOI}}, \tag{4}$$

where $\boldsymbol{P}_o$ is the CLIP encoded text embedding of `object`, and it is combined with the relation embedding v_*^{Rel} to generate the prompt that describes the entire HOI image. In this way, with only one learnable relation embedding (*i.e.*, v_*^{Rel}), we build a cycle to generate relation latent $\boldsymbol{z}_0^{\text{Rel}}$ from the HOI image latent $\boldsymbol{z}^{\text{HOI}}$, and subsequently, the original HOI image latent is reconstructed from the generated relation latent. The learning of v_*^{Rel} can be supervised without human annotation, but just ensuring the consistency between the original HOI latent and the reconstructed one:

$$\mathcal{L}_{\text{Consistency}} = ||\ell_2(\boldsymbol{z}^{\text{HOI}}) - \ell_2(\boldsymbol{z}_0^{\text{HOI}})||_2^2, \tag{5}$$

where all latents are ℓ_2-normalized for improved training stability [76]. Through such a disentanglement-based relation modeling and cycle-consistency training, the optimization objective become clearer and easier to learn. It enables using same `action` from different interactions to enhance the comprehension of a relation, and generalizing to new interactions by combining it with other `object`.

Relation-Centric Contrastive Learning. Eq. 5 is a pixel-level reconstruction loss which prioritizes aligning low-level cues. We supplement it with a relation-centric contrastive loss to enhance the awareness of high-level semantics. Instead of directly engaging learning with relation latents, we combine them with object latents to form new HOI latents, thus significantly enriching the diversity of samples:

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{z}_0^{\text{Rel}} + \boldsymbol{z}^{\text{Obj}}, & \boldsymbol{x}^+ &= \boldsymbol{z}_0^{\text{Rel}} + \boldsymbol{p}^{\text{Obj}}, \\ \boldsymbol{x}^-_{\text{Obj}} &= \boldsymbol{z}_0^{\text{Rel}} + \boldsymbol{n}_k^{\text{Obj}}, & \boldsymbol{x}^-_{\text{Rel}} &= \boldsymbol{n}_{0,i}^{\text{Rel}} + \boldsymbol{s}_j^{\text{Obj}}, \end{aligned} \tag{6}$$

where $\boldsymbol{x}$ is the anchor sample, $\boldsymbol{x}^+$ is the positive sample composed of a different object latent $\boldsymbol{p}^{\text{Obj}}$ sharing the same class as $\boldsymbol{z}^{\text{Obj}}$. Conversely, $\boldsymbol{x}^-_{\text{Obj}}$ and $\boldsymbol{x}^-_{\text{Rel}}$ are negative samples, with $\boldsymbol{x}^-_{\text{Obj}}$ composed

of a different class object latent $\boldsymbol{n}^{\text{Obj}}$ compared to $\boldsymbol{x}$, and $\boldsymbol{x}^{-}_{\text{Rel}}$ composed of any other relation latent $\boldsymbol{n}^{\text{Rel}}_{0}$ and arbitrary object latent $\boldsymbol{s}^{\text{Obj}}$. The final optimization objective is given as:

$$\mathcal{L}_{\text{Contrastive}} = -\log \frac{\exp(\boldsymbol{x}\cdot\boldsymbol{x}^{+}/\tau)}{\exp(\boldsymbol{x}\cdot\boldsymbol{x}^{+}/\tau)+\sum_k \exp(\boldsymbol{x}\cdot\boldsymbol{x}^{-}_{\text{Obj}}/\tau)+\sum_i\sum_j \exp(\boldsymbol{x}\cdot\boldsymbol{x}^{-}_{\text{Rel}}/\tau)}, \tag{7}$$

to optimize v^{Rel}_{*} which involves in reconstructing $\boldsymbol{z}^{\text{Rel}}_{0}$. $\tau = 0.07$ is the temperature parameter.

3.3 Relation-Driven Sample Generation

Text Prompts Preparation. We harness the captions provided in the MS COCO Caption dataset [77] to generate diverse prompts. Compared to text synthesized by GPT-4, these captions are more precise and closer to real visual scenes as they are annotated by human subjects. The preparation initiates with a filtration where captions not containing pronouns indicating `human` (*e.g.*, man, woman, boy) or `action` words are removed. To further enrich the diversity of prompts, given two randomly selected sentences that share the same `action`, we exchange the clauses following the `action` word. Prompts are exclusively generated with GPT-4 only when actions or objects not present in COCO Caption. This results in 33,834 text prompts in total. Finally, `action` words in prompts are replaced with placeholders corresponding to learned relation embeddings, so as to empower the diffusion model with enhanced awareness of relation patterns between `human` and `object` during generation.

Image and Annotation Generation. Denoting text prompts as $\mathcal{P} = \{\mathcal{P}_1, \cdots, \mathcal{P}_N\}$, we aim to construct a dataset $\mathcal{X} = \{(\mathcal{I}_1, \mathcal{A}_1), \cdots, (\mathcal{I}_N, \mathcal{A}_N)\}$ where $\mathcal{I}_i \in \mathbb{R}^{H\times W\times 3}$ represents the synthesized image and $\mathcal{A}_i = \{\mathcal{B}^h_i, \mathcal{B}^o_i, \mathcal{C}^o_i, \mathcal{C}^a_i\}$ is the pseudo annotation containing bounding boxes $\mathcal{B}^h_i$ for `human`, $\mathcal{B}^o_i$ for `object`, and class labels $\mathcal{C}^o_i$ for `object`, $\mathcal{C}^a_i$ for `action`. For the generation of $\mathcal{I}_i$, the text prompts $\mathcal{P}_i$ is first encoded by CLIP text encoder to obtain the conditioning vector $\boldsymbol{P}_i = c_\theta(\mathcal{P}_i) \in \mathbb{R}^d$, where the placeholder string is directly replaced with relation embedding v^{Rel}_{*}. Then, a random sampled noise tensor $\boldsymbol{z}_T \in \mathbb{R}^{h\times w\times d}$ is iteratively denoised to yield a new latent $\boldsymbol{z}_0$. $\mathcal{I}_i$ is generated by a single pass through $\mathcal{D}$, *i.e.*, $\mathcal{I}_i = \mathcal{D}(\boldsymbol{z}_0)$. For the generation of $\mathcal{A}_i$, $\mathcal{C}^o_i$ and $\mathcal{C}^a_i$ can be easily determined by referring to the `action` and `object` words in $\mathcal{P}_i$, while $\mathcal{B}^h_i$ and $\mathcal{B}^o_i$ are derived from the cross-attention maps computed within the U-shape denoising network ϵ_θ. Specifically, to effectively tackle various input modalities, ϵ_θ is equipped with cross-attention mechanisms in each layer to inject $\boldsymbol{P}_i$ into $\boldsymbol{z}$ conforming to the similarity between them. For the l-th layer at the last denoising step 0, the cross-attention map is computed as: $\boldsymbol{M}^l_{i,0} = \texttt{softmax}(\boldsymbol{z}_0 \cdot \boldsymbol{P}^\top_i/\sqrt{d}) \in \mathbb{R}^{h\times w}$. According to prior work [17, 78], here $\boldsymbol{M}^l_{i,0}$ signifies the correspondence between text prompt $\mathcal{P}_i$ and regions in generated image. Thus, we explicitly concatenate words describing `human` and `object` with $\mathcal{P}_i$ (*i.e.*, $[\mathcal{P}_i; word_{\texttt{human}}; word_{\texttt{object}}]$), resulting in a new text embedding $\hat{\boldsymbol{P}}_i \in \mathbb{R}^{d\times 3}$ and corresponding cross-attention maps $\hat{\boldsymbol{M}}^l_{i,0} \in \mathbb{R}^{h\times w\times 3}$ where the last two items along the third dimension channel are probability maps of `human` and `object`. Finally, we leverage the implementation in weakly supervised object localization [79] to outline bounding boxes from these probability maps.

3.4 HOI Knowledge Transfer from Diffusion Models

While prior studies [3, 48–50] have investigated knowledge transfer from visual-linguistic models such as CLIP, they utilize visual knowledge solely during training. The prediction relies on a confined set of CLIP encoded word embeddings, which leads to limited knowledge transfer and rigid inference unresponsive to image content. In contrast, we propose directly leveraging diffusion models as the feature extractor and build HOI detector on this basis. Moreover, given the conditioning property of diffusion model, relation embeddings can serve as text prompts to guide the retrieval of interaction-relevant visual cues from images, further benefiting HOI detection.

HOI Detector Built Upon Diffusion Models. Pioneering studies [17, 78, 80] have empirically demonstrated that the output of frozen text-to-image diffusion models possesses rich visual features to tackle complex perception tasks. Next we illustrate how to build a HOI detector on this basis. As shown in Fig. 3, our method is a one-stage solution composed of: a visual encoder with diffusion models serving as the backbone, and a HOI decoder consisting of two parallel decoders for instance and interaction detection. For the visual encoder, given an image $\mathcal{I}$, it is encoded into latent space with the encoder $\mathcal{E}$ of a pre-trained generator (*e.g.*, VQGAN): $\boldsymbol{z} = \mathcal{E}(\mathcal{I})$. Then, $\boldsymbol{z}$ is fed into ϵ_θ through a single noise-free forward pass to derive text conditioned features: $\epsilon_\theta(\boldsymbol{z}, T, c_\theta(y)) \rightarrow \{\boldsymbol{z}^l_T\}^4_{l=1}$. All

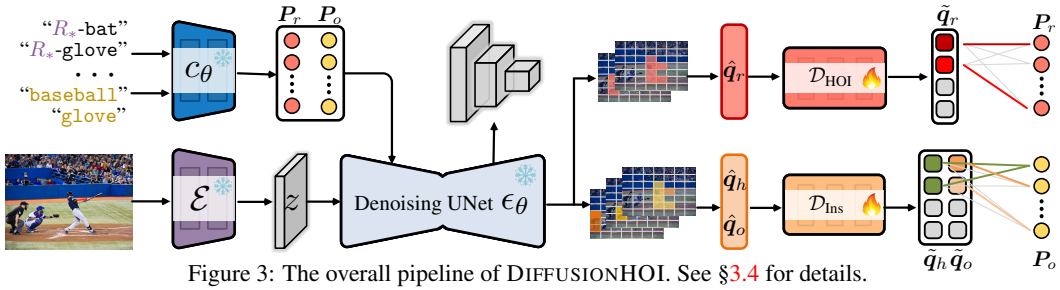

Figure 3: The overall pipeline of DIFFUSIONHOI. See §3.4 for details.

scales of features are aggregated with FPN [81], yielding $\boldsymbol{z}'_T$ in a downsampling factor of 32. The architecture of HOI decoder is similar to GEN-VLKT [3]. Concretely, the instance decoder $\mathcal{D}_{\text{Ins}}$ employs a set of human queries $\{\boldsymbol{q}_h^i\}_{i=1}^{N_q}$ and object queries $\{\boldsymbol{q}_o^j\}_{j=1}^{N_q}$, and considers those at the same index (*i.e.*, $i=j$) as a pair to initialize queries $\boldsymbol{q}_r$ for the interaction decoder $\mathcal{D}_{\text{HOI}}$. In fact, the HOI decoder can be replaced with any other one-stage models. We do not claim the detector architecture as the contribution, but focus on how to derive HOI-relevant feature to assist in HOI detection.

Relation-Inspired HOI Detection. As the relation embeddings are optimized towards modeling the interactions between `human` and `object`, we use them as conditions to inspire the extraction of HOI-relevant cues. This can eliminate the potential domain gap between general-purpose diffusion models and the downstream HOI detection task. Specifically, all feasible HOI phrases (*e.g.*, "human feed horse") are encoded with CLIP text encoder into embedding space and concatenated together, with the `human-action` component replaced with learned relation embeddings (*e.g.*, "R_* horse"). This results in HOI prompts $\boldsymbol{P}_r \in \mathbb{R}^{N_r \times d_l}$ which further participates into the cross-attention in ϵ_θ via:

$$\boldsymbol{z}_T^l = \boldsymbol{z}_T^l + \boldsymbol{M}_r^l \cdot \boldsymbol{V}_{\boldsymbol{P}_r} \in \mathbb{R}^{h \times w \times d_l}, \quad \boldsymbol{M}_r^l = \texttt{softmax}(\boldsymbol{z}_T^l \cdot \boldsymbol{K}_{\boldsymbol{P}_r}^\top / \sqrt{d}) \in \mathbb{R}^{h \times w \times N_r}, \tag{8}$$

where $\boldsymbol{K}_{\boldsymbol{P}_r}$ and $\boldsymbol{V}_{\boldsymbol{P}_r}$ are key and value embeddings projected from $\boldsymbol{P}_r$. As seen, $\boldsymbol{P}_r$ contributes to: **i)** encourage the denoising network ϵ_θ to extract visual features $\boldsymbol{z}_T^l$ *w.r.t.* HOI prompts, and **ii)** guide the derivative of cross-attention maps $\boldsymbol{M}_r^l$ in response to ongoing interactions in $\mathcal{I}$. The final interaction maps are computed as the average value of $\{\boldsymbol{M}_r^l\}_{l=1}^4$. we also derive cross-attention maps for `human` $\boldsymbol{M}_h$ and `object` $\boldsymbol{M}_o$ in a similar way as $\boldsymbol{M}_r$, by without update to $\boldsymbol{z}_T^l$. Then, these cross-attention maps are used to initialize queries from the aggregated visual feature $\boldsymbol{z}'_T$ via mask pooling:

$$\hat{\boldsymbol{q}}_r = \texttt{MaskPooling}(\boldsymbol{z}'_T, \boldsymbol{M}_r^k), \quad \hat{\boldsymbol{q}}_o = \texttt{MaskPooling}(\boldsymbol{z}'_T, \boldsymbol{M}_o), \quad \hat{\boldsymbol{q}}_h = \texttt{MaskPooling}(\boldsymbol{z}'_T, \boldsymbol{M}_h). \tag{9}$$

Note we conduct Hungarian matching between $\hat{\boldsymbol{q}}_r^k$ and $\hat{\boldsymbol{q}}_o^i + \hat{\boldsymbol{q}}_h^j$, so as to arrange HOI, and combined human-object queries that are most similar to the same index in their respective query lists. Following [17], the classification for interaction and instance are jointly supervised by:

$$\begin{aligned} \mathcal{L}_{\text{HOI}} &= \texttt{CE}(\texttt{softmax}(\tilde{\boldsymbol{q}}_r \cdot \boldsymbol{P}_r / \tau_r), y_r) + \texttt{CE}(\texttt{softmax}(\texttt{FFN}(\tilde{\boldsymbol{q}}_r)), y_r), \\ \mathcal{L}_{\text{Ins}} &= \texttt{CE}(\texttt{softmax}(\tilde{\boldsymbol{q}}_o \cdot \boldsymbol{P}_o / \tau_o), y_o) + \texttt{CE}(\texttt{softmax}(\texttt{FFN}(\tilde{\boldsymbol{q}}_o)), y_o), \end{aligned} \tag{10}$$

where y_r and y_o are ground truth for interaction and object categories, $\tilde{\boldsymbol{q}}_r$ and $\tilde{\boldsymbol{q}}_o$ are queries after decoding through $\mathcal{D}_{\text{HOI}}$ and $\mathcal{D}_{\text{Ins}}$, `CE` and `FFN` denote the cross entropy loss and feed-forward network. Beyond the score delivered by conventional linear classifier (*i.e.*, $\texttt{softmax}(\texttt{FFN}(\tilde{\boldsymbol{q}}_o))$), here $\texttt{softmax}(\tilde{\boldsymbol{q}} \cdot \boldsymbol{P}/\tau)$ with learnable parameters τ_r and τ_o computes the similarity between decoded queries and conditioning prompts, thereby facilitating the recognition for unseen categories.

Online Update for Relation Embedding. To enable the continual evolution of relation embeddings v_*^{Rel} throughout the supervised HOI detection learning, an additional loss considering the compositional nature of HOI is devised. Specifically, we concatenate all N_a relation embeddings into a new prompt $\boldsymbol{P}_a$, from which a set of relation query embeddings $\hat{\boldsymbol{q}}_a$ can be initialized in the same way as $\hat{\boldsymbol{q}}_o$ (*c.f.*, Eq.9). In addition, another set of embeddings describing relations can be derived from $\tilde{\boldsymbol{q}}_r$ and $\tilde{\boldsymbol{q}}_o$ by: $\tilde{\boldsymbol{q}}_a = \tilde{\boldsymbol{q}}_r - \tilde{\boldsymbol{q}}_o$. The goal is to align $\hat{\boldsymbol{q}}_a$ directly derived from visual features with relation embeddings as conditions, and $\tilde{\boldsymbol{q}}_a$ computed from interaction and object queries after decoding:

$$\mathcal{L}_{\text{Rel}} = ||\ell_2(\hat{\boldsymbol{q}}_a) - \ell_2(\tilde{\boldsymbol{q}}_a)||_2^2. \tag{11}$$

Here $\mathcal{L}_{\text{Rel}}$ solely optimizes v_*^{Rel} to render a mutual boost between HOI detection and relation embedding learning. Concretely, enhanced relation embeddings inspires improved HOI feature discovery, and in turn, the more precise query decoding benefits the update of relation embeddings.

Table 1: Quantitative results for regular HOI detection on HICO-DET[20] and V-COCO[21].

Method	Backbone	VL Pretrain	Default			Known Object			V-COCO	
			Full	Rare	Non-Rare	Full	Rare	Non-Rare	AP^{S1}_{role}	AP^{S2}_{role}
iCAN[83] [BMVC18]	R50	-	14.84	10.45	16.150	16.26	11.33	17.73	45.3	-
PPDM[24] [CVPR20]	HG104	-	21.73	13.78	24.10	24.58	16.65	26.84	-	-
HOTR[41] [CVPR21]	R50	-	23.46	16.21	25.60	-	-	-	55.2	64.4
QPIC[42] [CVPR21]	R101	-	29.90	23.92	31.69	32.38	26.06	34.27	58.3	60.7
CDN[44] [NeurIPS21]	R101	-	32.07	27.19	33.53	34.79	29.48	36.38	63.9	65.9
CPChoi[84] [CVPR22]	R50	-	29.63	23.14	31.57	-	-	-	63.1	65.4
STIP[85] [CVPR22]	R50	-	32.22	28.15	33.43	35.29	31.43	36.45	66.0	70.7
UPT[86] [CVPR22]	R101	-	32.62	28.62	33.81	36.08	31.41	37.47	61.3	67.1
Iwin[87] [ECCV22]	R101	-	32.79	27.84	35.40	35.84	28.74	36.09	60.9	-
MCPC[88] [ECCV22]	R50	-	35.15	33.71	35.58	37.56	35.87	38.06	63.0	65.1
PViC[89] [ICCV23]	R50	-	34.69	32.14	35.45	38.14	35.38	38.97	62.8	67.8
PViC†[89] [ICCV23]	Swin-L	-	44.32	44.61	44.24	47.81	48.38	47.64	64.1	70.2
GEN-VLK[3] [CVPR22]	R101	CLIP	34.95	31.18	36.08	38.22	34.36	39.37	63.6	65.9
HOICLIP[19] [CVPR23]	R50	CLIP	34.69	31.12	35.74	37.61	34.47	38.54	63.5	64.8
CQL[90] [CVPR23]	R50	CLIP	35.36	32.97	36.07	38.43	34.85	39.50	66.4	69.2
ViPLO[91] [CVPR23]	ViT-B	CLIP	37.22	35.45	37.75	40.61	38.82	41.15	62.2	68.0
AGER[92] [ICCV23]	R50	CLIP	36.75	33.53	37.71	39.84	35.58	40.2	65.7	69.7
RmLR[93] [ICCV23]	R101	MobileBERT	37.41	28.81	39.97	38.69	31.27	40.91	64.2	70.2
ADA-CM[94] [ICCV23]	ViT-L	CLIP	38.40	37.52	38.66	-	-	-	58.6	64.0
DIFFUSIONHOI	VQGAN	Stable Diffusion	**38.12**	**38.93**	**37.84**	**40.93**	**42.87**	**40.04**	**66.8**	**70.9**
DIFFUSIONHOI	ViT-L	Stable unCLIP	**42.54**	**42.95**	**42.35**	**44.91**	**45.18**	**44.83**	**67.1**	**71.1**

†: Models built upon advanced object dector, *i.e.*, $\mathcal{H}$-Deform-DETR[95].

3.5 Implementation Details

Network Architecture. DIFFUSIONHOI is built upon Stable Diffusion v1.5 with xFormers[82] installed. The denoising UNet ϵ_θ receives input latents at a downsampling factor of 1/8, with four encoder blocks output feature at a size of $1/2^{l+3}$ where l is the block index. For the final visual feature $\boldsymbol{z}'$ after FPN aggregation, it is interpolated to a size of 1/32 and then projected to 256 channels to enhance computing efficiency. Both $\mathcal{D}_{\text{HOI}}$ and $\mathcal{D}_{\text{Ins}}$ consist of six Transformer decoding layers with hidden dimension of 768. The query number N^q is uniformly set to 64 for both $\mathcal{D}_{\text{HOI}}$ and $\mathcal{D}_{\text{Ins}}$.

Training Objective. The inversion-based HOI modeling is jointly optimized by two embedding learning losses: $\mathcal{L}_{\text{Inversion}} = \mathcal{L}_{\text{Consistency}} + \lambda_1 \mathcal{L}_{\text{Contrastive}}$, where $\lambda_1 \in [0, 0.2]$ is scheduled following a cosine annealing policy. For HOI detection learning, we follow DETR[39] to match predictions and ground truths with Hungarian algorithm. Denoting the bounding box detection loss as $\mathcal{L}_{\text{Det}}$, the final training objective is given as: $\mathcal{L} = \mathcal{L}_{\text{HOI}} + \mathcal{L}_{\text{Ins}} + \mathcal{L}_{\text{Det}} + \lambda_2 \mathcal{L}_{\text{Rel}}$ where λ_2 is fixed to 0.5.

4 Experiment

4.1 Experimental Setup

Datasets. We conduct extensive experiments on three datasets.

- HICO-DET[20] is a large-scale HOI detection benchmark with 38,118/9,658 images for training/testing, respectively. This dataset includes 80 object categories as in MS-COCO[96] and 117 action categories, formulating a rich vocabulary of 600 human-object interactions in total.
- V-COCO[21] is a curated subset of MS-COCO[96] including 2,533/2,867/4,946 images in `train`/`val`/ `test` sets. It also contains 80 object categories from MS-COCO[96] and a much smaller set of 29 action classes, resulting in a total of 263 human-object interactions.
- SWiG-HOI[22] is assembled from SWiG[97] and DOH[98] with about 45,000/14,000 for training/testing. This dataset covers 406 human actions and 1,000 object categories.

Zero-Shot HOI Discovery. In accordance with prior research[3, 19, 99–102], the zero-shot HOI discovery on HICO-DET[20] uses four setups: Rare First Unseen Combination (RF-UC), Non-rare First Unseen Combination (NF-UC), Unseen Verb (UV), and Unseen Object (UO). The RF-UC and NF-UC configurations excluded the 120 most frequent/infrequent interaction categories from the training sets for testing purposes only. The UV and UO setups reserve 20 verb classes and 12 object classes never encountered during training for testing. For SWiG-HOI[22], the test set includes approximately 5,500 interactions, with around 1,800 of them not present in the training set.

Table 2: Zero-shot generalization on HICO-DET [20].

Method	Type	Unseen	Seen	Full
ATL[101] [CVPR21]	RF-UC	9.18	24.67	21.57
FCL[100] [CVPR21]	RF-UC	13.16	24.23	22.01
SCL[102] [ECCV22]	RF-UC	19.07	30.39	28.08
GEN-VLKT[3] [CVPR22]	RF-UC	21.36	32.91	30.56
OpenCat[103] [CVPR23]	RF-UC	21.46	33.86	31.38
HOICLIP[19] [CVPR23]	RF-UC	25.53	34.85	32.99
DIFFUSIONHOI	RF-UC	**32.06**	**36.77**	**35.89**
ATL[101] [CVPR21]	NF-UC	18.25	18.78	18.67
FCL[100] [CVPR21]	NF-UC	18.66	19.55	19.37
SCL[102] [ECCV22]	NF-UC	21.73	25.00	24.34
GEN-VLKT[3] [CVPR22]	NF-UC	25.05	23.38	23.71
OpenCat[103] [CVPR23]	NF-UC	23.25	28.04	27.08
HOICLIP[19] [CVPR23]	NF-UC	26.39	28.10	27.75
DIFFUSIONHOI	NF-UC	**30.04**	**30.29**	**30.25**
ATL[101] [CVPR21]	UO	5.05	14.69	13.08
FCL[100] [CVPR21]	UO	15.54	20.74	19.87
GEN-VLKT[3] [CVPR22]	UO	10.51	28.92	25.63
OpenCat[103] [CVPR23]	UO	23.84	28.49	27.72
HOICLIP[19] [CVPR23]	UO	16.20	30.99	28.53
DIFFUSIONHOI	UO	**22.37**	**32.03**	**31.12**
GEN-VLKT[3] [CVPR22]	UV	20.96	30.23	28.74
HOICLIP[19] [CVPR23]	UV	24.30	32.19	31.09
DIFFUSIONHOI	UV	**28.05**	**33.24**	**32.67**

Table 3: Zero-shot generalization on SWiG-DET [22].

Method	Non-rare	Rare	Unseen	Full
QPIC[42] [CVPR21]	16.95	10.84	6.21	11.12
THID[48] [CVPR22]	17.67	12.82	10.04	13.26
CMD-SE[104] [CVPR24]	21.46	14.64	10.70	15.26
DIFFUSIONHOI	**25.59**	**20.61**	**18.93**	**21.69**

Table 4: Comparison of parameters and running efficiency. * means applying accelerated technology.

Method	Backbone	Trainable Params (M)	FPS	HICO-DET
Two-stages Detectors:				
iCAN[83] [BMVC18]	R50	39.8	6.23	14.84
DRG[105] [ECCV20]	R50	46.1	6.05	19.26
STIP[85] [CVPR22]	R50	50.4	7.12	32.22
ViPLO[91] [CVPR23]	ViT-B	118.2	5.66	37.22
ADA-CM[94] [ICCV23]	ViT-L	6.6	3.24	38.40
One-stages Detectors:				
PPDM[24] [CVPR20]	HG104	194.9	17.58	21.73
HOTR[41] [CVPR21]	R50	51.2	15.92	23.46
QPIC[42] [CVPR21]	R50	41.9	17.41	29.07
CDN[44] [NeurIPS21]	R50	42.1	16.24	31.78
GEN-VLKT[3] [CVPR22]	R50	42.8	18.23	33.75
DIFFUSIONHOI	VQ-GAN	27.6	9.49	38.12
*DIFFUSIONHOI	VQ-GAN	27.6	24.77	38.12

Evaluation Metric. Following conventions [3, 41, 42], we adopt mAP as metrics. For HICO-DET, we report performance according to Default and Known Object two setups. The former computes mAP across all testing images, while the latter is tailored for each object class. For each setup, the scores are reported in Full/Rare/Non-Rare three types. For V-COCO, we evaluate the performance under scenario 1 (S1) which contains all 29 actions and scenario 2 (S2) which excludes 4 actions interact with no objects. For zero-shot setup, the evaluation is divided into Seen/Unseen/Full three sets for HICO-DET, and Non-Rare/Rare/Unseen/Full four sets for SWiG-HOI.

Training and Testing. The diffusion model and CLIP text encoder are kept frozen during training. For inversion-based HOI modeling, the only learnable parameters are relation embeddings, which are updated for 40,000 steps using images sampled from HICO-DET. Following [18], we employ a base learning rate of $8e^{-2}$ with a batch size of 32. For HOI detection learning, we train the interaction decoder $\mathcal{D}_{\text{Ins}}$ and object decoder $\mathcal{D}_{\text{HOI}}$ for 60 epochs with a base learning rate of $1e^{-4}$ and batch size of 16, using both synthesized data and the target dataset. Subsequently, the model is trained only on the target dataset for an additional 30 epochs with a base learning rate of $1e^{-5}$. During inference, no data augmentation is used to ensure fair comparison. Following [3, 103], the inputs are resized to maximum of 1,333 pixels on long sides, and the shortest sides falls between 480 and 800 pixels.

Reproducibility. DIFFUSIONHOI is implemented in PyTorch and trained on 8 Tesla A40 GPUs with 48GB memory per card.

4.2 Comparison with State-of-the-Arts

Regular Setup. We first compare DIFFUSIONHOI with top-leading solutions on HICO-DET [20] and V-COCO [21] under the regular setup. As shown in Table 1, for HICO-DET, our method achieves the best performance on both Default and Known Object setups. Notably, with the encoder of VQGAN as the backbone, it surpasses the previous SOTA, RemLR [93], which employs a similar level backbone (*i.e.*, ResNet-50) by **1.19%** and **2.64%** on the Full categories. Benefited from synthesized data and comprehensive knowledge transfer from diffusion models, the performance on Rare categories improves significantly, achieving higher scores than on Non-Rare categories for the first time. Finally, with a more powerful VL model (*i.e.*, Stable unCLIP [11]) and backbone (*i.e.*, ViT-L), the performance is boosted to **42.54%** under the Default setup, surpassing nearly all existing work by a considerable margin. Please note that PViC [89] with Swin-L as the backbone leverages $\mathcal{H}$-Deform-DETR [95] as the detector which achieves 48.7 mAP on MS COCO [96] by running merely 12 epochs, significantly higher than DETR [39] which achieves 36.2 mAP by running 50 epochs.

Zero-Shot Setup. Next we investigate the effectiveness of DIFFUSIONHOI under the zero-shot generalization setup. As shown in Table 2, our method yields remarkable performance across all four setups on HICO-DET. In particular, it surpasses the previous SOTA (*i.e.*, HOICLIP [19]) by

Table 5: Detailed analysis of essential components of DIFFUSIONHOI on HICO-DET[20].

Algorithm Component	Default			RF-UC		
	Full	Rare	Non-Rare	Full	Unseen	Seen
BASELINE	33.24	30.25	34.32	30.47	20.63	33.09
+ Synthesized Data *only*	35.49	36.27	35.02	33.79	28.22	34.85
+ Relation Prompting *only*	36.45	35.78	36.71	34.25	26.57	35.58
DIFFUSIONHOI	38.12	38.93	37.84	35.89	32.06	36.77

Table 6: Analysis of conditioning input for relation-inspired HOI detection on HICO-DET[20].

Conditioning Input	Default			RF-UC		
	Full	Rare	Non-Rare	Full	Unseen	Seen
-	33.24	30.25	34.32	30.47	20.63	33.09
Textual Description	33.71	30.98	34.73	30.72	21.29	33.24
Relation Embedding	36.45	35.78	36.71	34.25	26.57	35.58

Table 7: Analysis of relation embeddings with different learning strategies for relation-inspired HOI detection.

Learning Strategy	Default			RF-UC		
	Full	Rare	Non-Rare	Full	Unseen	Seen
Textual Inversion	34.03	32.17	34.61	30.93	21.55	33.45
Cycle-Consistency	35.23	34.56	35.46	32.96	24.24	34.54
+ Relation-Centric CL	35.94	35.06	36.32	33.72	25.73	35.17
+ Online Update	36.45	35.78	36.71	34.25	26.57	35.58

Table 8: Analysis of prompts for dataset generation. *TD*: textual description, *RE*: relation embedding.

Training Set	Default			RF-UC		
	Full	Rare	Non-Rare	Full	Unseen	Seen
HICO-DET	33.24	30.25	34.32	30.47	20.63	33.09
+ *TD* Synthesized Data	32.54	30.04	33.49	30.12	20.55	32.57
+ *RE* Synthesized Data	35.49	36.27	35.02	33.79	28.22	34.85

2.90% under the RF-UC setup. This setup emphasizes compositional generalization which requires models to comprehend new types of interactions using known actions and objects. It aligns well with the strengths of text-to-image diffusion models to generate images conditioned on compositionally organized textual descriptions. Moreover, due to the effective knowledge transfer, DIFFUSIONHOI also achieves satisfactory improvement under the UV and UO setups which focus on the recognition of novel actions and objects. Table 3 further confirms the exceptional ability of our method, showing **5.97%/8.23%** mAP improvements over CMD-SE[104] under Rare and Unseen two categories.

Model Efficiency. We compare the trainable parameter number and inference time in Table 4. As seen, DIFFUSIONHOI demonstrates significantly fewer trainable parameters compared to the one-stage counterparts. This is attributed to our inversion-based HOI modeling, which avoids fine-tuning diffusion models like previous work[78], while effectively capturing task-specific properties. Regarding inference speed, even with stable diffusion for feature extraction, our method still achieves 9.49 FPS, a rate similar to two-stage models. This is due to the inference involving only one single forward pass, and the downsampling factor of stable diffusion from 1/8 to 1/64 is smaller than conventional backbones typically from 1/4 to 1/32. Moreover, thank to the flourishing community of stable diffusion, a variety of optimized inference solutions have emerged. By running at fp16 precision and using traced UNet, the FPS increases to 24.77, surpassing most one-stage methods.

4.3 Diagnostic Analysis

Key Component Analysis. We first examine the essential components of DIFFUSIONHOI in Table 5. Here BASELINE denotes HOI detector built upon stable diffusion without text prompting. Through jointly training with the synthesized data, both Default and RF-UC setups observe notable improvements (*e.g.*, up to **2.25%** and **3.32%** on Full categories). This verifies the effectiveness of our relation-driven HOI image generation strategy. in addition, after imposing relation embedding to prompt the feature extraction and HOI detection processes, the performance boosts to **36.45%** and **34.25%** under two setups. Finally, after combining these two core components together, our DIFFUSIONHOI delivers consistent improvements and sets new SOTA across all setups.

Conditioning Input. To assess the effectiveness of learned relational embeddings, we present the experimental results using different conditional inputs to stimulate HOI detection in Table 6. As seen, though action words offer limited improvement, they are far surpassed by relation embeddings which enables HOI-oriented feature extraction and enhance query initialization through cross-attention.

Relation Embedding Learning. Next we probe the impact of different strategies for relation embedding learning. The results regarding relation-inspired HOI detection are summarized in Table 7. It can be observed that textual inversion, which directly uses different images sharing the same action for relation embedding learning, is inferior to our cycle-consistency learning strategy that considers the disentanglement nature of HOI interactions. On this basis, the relation-centric contrastive learning and online update strategies consistently bring improvement in both setups.

Prompt for Dataset Generation. Finally we study the impact of data synthesized by different types of textual prompts in Table 8. As observed, data generated with purely textual description using plain action words like "The man at bat readies to swing at the pitch" gives negative improvement over baseline. The potentially indicates that diffusion models cannot understand the relations between human-object pairs and generate meaningful images when provided with straightforward textual

Table 9: Comparison of total training time on HICO-DET[20].

Method	CDN[44]	HOTR[41]	UPT[86]	STIP[85]	GEN-VLKT[3]	HOICLIP[19]	CQL[90]	DIFFUSIONHOI
Time (Hour)	25.2	23.6	17.9	16.4	28.4	29.1	29.7	**5.7+11.5**

description. In contrast, through relation modeling, data generated with relation embeddings to replace the plain action words provides high-quality samples for the training of HOI detectors.

Analysis on Training Cost. For our inversion-based HOI modeling to learn relation-centric embeddings, unlike the original textual inversion technology that learns text embeddings within the image space, we optimize relation embeddings within the latent space by reconstructing interaction features. This lead to reduced training costs. Consequently, the 117 relation embeddings in HICO-DET[20] can be learned within **5.7** hours (23 minutes per relation embedding) which is more efficient than textual inversion (*i.e.*, 32 minutes per embedding). For the main training of HOI detection on HICO-DET, since our method utilizes significantly fewer trainable parameters compared to existing work (*e.g.*, 27.6M *v.s.* 50.4M for STIP[85], 41.9M for QPIC[42], and 42.8M for GEN-VLKT[3] in Table 4), the training process can be completed in just **11.5** hours. The comparison of the whole training time with some representative work is summarized in Table 9, with all experiments conducted on 8 Tesla A40 cards. It can be observed that DIFFUSIONHOI requires less training time than most existing work.

5 Conclusion

We present DIFFUSIONHOI, a new HOI detector built upon diffusion models. By explicitly modeling the relations between humans and objects in an inversion-based manner, we enable effective knowledge transfer from diffusion models while adapting unique characteristics of the HOI detection task. This is achieved in two aspects: **i)** *relation-driven* image generation using diffusion models to enrich the training set with more HOI-oriented samples, and **ii)** *relation-inspired* HOI detection with learned relation embeddings as prompts to retrieve task-specific features from images, thereby enhancing the recognition of ongoing interactions. Extensive experiments demonstrate that DIFFUSIONHOI excels in both regular or zero-shot setups and sets new SOTAs. We believe this work provides insights to unleash the power of diffusion models for downstream visual perception tasks in an efficient manner.

Acknowledgement. This work was supported by the National Science and Technology Major Project (No. 2023ZD0121300), the National Natural Science Foundation of China (No. 62372405), the Fundamental Research Funds for the Central Universities 226-2024-00058, National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University (No. HMHAI-202403), and CIPSC-SMP-Zhipu Large Model Cross-Disciplinary Fund.

References

[1] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 1

[2] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *ICCV*, 2019. 1

[3] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *CVPR*, 2022. 1, 2, 3, 5, 6, 7, 8, 10, 17

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 3

[5] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *ACL*, 2022. 1

[6] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *CVPR*, 2023. 1

[7] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1, 3

[8] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022. 3

[9] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *WACV*, 2023.

[10] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022.

[11] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 8

[12] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *ICLR*, 2022.

[13] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 3

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 3

[15] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*, 2022. 1

[16] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *ICLR*, 2022. 1

[17] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 1, 3, 5, 6

[18] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2022. 2, 3, 8

[19] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *CVPR*, 2023. 2, 7, 8, 10, 17

[20] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 2, 7, 8, 9, 10, 17

[21] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2, 7, 8, 17

[22] Suchen Wang, Kim-Hui Yap, Henghui Ding, Jiyan Wu, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In *ICCV*, 2021. 2, 7, 8, 17

[23] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *ECCV*, 2020. 3, 17

[24] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, 2020. 7, 8, 17

[25] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *CVPR*, 2020.

[26] Hao-Shu Fang, Yichen Xie, Dian Shao, and Cewu Lu. Dirv: Dense interaction region voting for end-to-end human-object interaction detection. In *AAAI*, 2021. 3

[27] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 3

[28] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, 2019.

[29] Hai Wang, Wei-shi Zheng, and Ling Yingbiao. Contextual heterogeneous graph network for human-object interaction detection. In *ECCV*, 2020.

[30] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. In *NeurIPS*, 2020.

[31] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, 2020.

[32] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *ICCV*, 2021.

[33] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018.

[34] Tianfei Zhou, Wenguan Wang, Siyuan Qi, Haibin Ling, and Jianbing Shen. Cascaded human-object interaction recognition. In *CVPR*, 2020.

[35] Tianfei Zhou, Siyuan Qi, Wenguan Wang, Jianbing Shen, and Song-Chun Zhu. Cascaded parsing of human-object interaction recognition. *IEEE TPAMI*, 44(6):2827–2840, 2021.

[36] Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Hydra-sgg: Hybrid relation assignment for one-stage scene graph generation. *arXiv preprint arXiv:2409.10262*, 2024.

[37] Jianan Wei, Tianfei Zhou, Yi Yang, and Wenguan Wang. Nonverbal interaction detection. In *ECCV*, 2024. 3

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 3

[39] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3, 7, 8

[40] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, 2021. 3, 17

[41] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021. 7, 8, 10, 17

[42] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021. 7, 8, 10, 17

[43] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *CVPR*, 2021.

[44] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. In *NeurIPS*, 2021. 7, 8, 10, 17

[45] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In *CVPR*, 2022.

[46] Xubin Zhong, Changxing Ding, Zijian Li, and Shaoli Huang. Towards hard-positive query mining for detr-based human-object interaction detection. In *ECCV*, 2022. 3

[47] ASM Iftekhar, Hao Chen, Kaustav Kundu, Xinyu Li, Joseph Tighe, and Davide Modolo. What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions. In *CVPR*, 2022. 3, 17

[48] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *CVPR*, 2022. 5, 8

[49] Xian Qu, Changxing Ding, Xingao Li, Xubin Zhong, and Dacheng Tao. Distillation using oracle queries for transformer-based human-object interaction detection. In *CVPR*, 2022. 17

[50] Leizhen Dong, Zhimin Li, Kunlun Xu, Zhijun Zhang, Luxin Yan, Sheng Zhong, and Xu Zou. Category-aware transformer network for better human-object interaction detection. In *CVPR*, 2022. 5, 17

[51] Guikun Chen, Jin Li, and Wenguan Wang. Scene graph generation with role-playing large language models. In *NeurIPS*, 2024. 3

[52] Liulei Li, Jianan Wei, Wenguan Wang, and Yi Yang. Neural-logic human-object interaction detection. In *NeurIPS*, 2023. 3

[53] Wenguan Wang, Yi Yang, and Yunhe Pan. Visual knowledge in the big model era: Retrospect and prospect. *Frontiers of Information Technology & Electronic Engineering*, 2024. 3

[54] Liulei Li, Wenguan Wang, and Yi Yang. Logicseg: Parsing visual semantics with neural logic learning and reasoning. In *ICCV*, 2023. 3

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3

[56] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3

[57] Rohit Saha, Brendan Duke, Florian Shkurti, Graham W Taylor, and Parham Aarabi. Loho: Latent optimization of hairstyles via orthogonalization. In *CVPR*, 2021. 3

[58] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 3

[59] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *NeurIPS*, 2019. 3

[60] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3

[61] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[62] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 3

[63] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 3

[64] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. In *NeurIPS*, 2021. 3

[65] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021. 3

[66] Max Daniels, Tyler Maunu, and Paul Hand. Score-based generative neural networks for large-scale optimal transport. 2021. 3

[67] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 45(4):4713–4726, 2022. 3

[68] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2021. 3

[69] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2022. 3

[70] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, 2023.

[71] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023. 3

[72] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *NeurIPS*, 2023. 3

[73] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023.

[74] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. In *NeurIPS*, 2023. 3

[75] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *ICCV*, 2023. 3

[76] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. In *ICLR*, 2022. 4

[77] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5

[78] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *ICCV*, 2023. 5, 9

[79] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *CVPR*, 2020. 5

[80] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2022. 5

[81] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 6

[82] Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. xformers: A modular and hackable transformer modelling library. https://github.com/facebookresearch/xformers, 2022. 7

[83] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018. 7, 8, 17

[84] Jihwan Park, SeungJun Lee, Hwan Heo, Hyeong Kyu Choi, and Hyunwoo J Kim. Consistency learning via decoding path augmentation for transformers in human object interaction detection. In *CVPR*, 2022. 7, 17

[85] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *CVPR*, 2022. 7, 8, 10, 17

[86] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *CVPR*, 2022. 7, 10, 17

[87] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. Iwin: Human-object interaction detection via transformer with irregular windows. In *ECCV*, 2022. 7, 17

[88] Xiaoqian Wu, Yong-Lu Li, Xinpeng Liu, Junyi Zhang, Yuzhe Wu, and Cewu Lu. Mining cross-person cues for body-part interactiveness learning in hoi detection. In *ECCV*, 2022. 7, 17

[89] Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting of human-object interactions. In *ICCV*, 2023. 7, 8, 17

[90] Chi Xie, Fangao Zeng, Yue Hu, Shuang Liang, and Yichen Wei. Category query learning for human-object interaction classification. In *CVPR*, 2023. 7, 10, 17

[91] Jeeseung Park, Jin-Woo Park, and Jong-Seok Lee. Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In *CVPR*, 2023. 7, 8, 17

[92] Danyang Tu, Wei Sun, Guangtao Zhai, and Wei Shen. Agglomerative transformer for human-object interaction detection. In *ICCV*, 2023. 7, 17

[93] Yichao Cao, Qingfei Tang, Feng Yang, Xiu Su, Shan You, Xiaobo Lu, and Chang Xu. Re-mine, learn and reason: Exploring the cross-modal semantic correlations for language-guided hoi detection. In *ICCV*, 2023. 7, 8, 17

[94] Ting Lei, Fabian Caba, Qingchao Chen, Hailin Jin, Yuxin Peng, and Yang Liu. Efficient adaptive human-object interaction detection with concept-guided memory. In *ICCV*, 2023. 7, 8, 17

[95] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *CVPR*, 2023. 7, 8, 17

[96] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 7, 8

[97] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *ECCV*, 2020. 7

[98] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020. 7

[99] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020. 7

[100] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, 2021. 8

[101] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *CVPR*, 2021. 8

[102] Zhi Hou, Baosheng Yu, and Dacheng Tao. Discovering human-object interaction concepts via self-compositional learning. In *ECCV*, 2022. 7, 8

[103] Sipeng Zheng, Boshen Xu, and Qin Jin. Open-category human-object interaction pre-training via language modeling framework. In *CVPR*, 2023. 8

[104] Ting Lei, Shaofeng Yin, and Yang Liu. Exploring the potential of large foundation models for open-vocabulary hoi detection. In *CVPR*, 2024. 8, 9

[105] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *ECCV*, 2020. 8, 17

[106] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *ACM MM*, 2020. 17

[107] Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In *CVPR*, 2022. 17

[108] Xinpeng Liu, Yong-Lu Li, Xiaoqian Wu, Yu-Wing Tai, Cewu Lu, and Chi-Keung Tang. Interactiveness field in human-object interactions. In *CVPR*, 2022. 17

[109] Guangzhi Wang, Yangyang Guo, Yongkang Wong, and Mohan Kankanhalli. Chairs can be stood on: Overcoming object bias in human-object interaction detection. In *ECCV*, 2022. 17

A Discussion

A.1 Limitation

One potential limitation of our method is that it does not run as quickly as previous one-stage methods due to the adoption of diffusion models. A detailed comparison is provided in Table 4. As shown, DIFFUSIONHOI achieves inference speeds similar to two-stage methods. However, currently there is a trend towards leveraging large-scale pre-trained models to assist in various perception tasks. These models offer substantial enhancements in performance and accuracy. For instance, after utilizing diffusion models, our method has demonstrated an 8.23% improvement on unseen categories in the SWiG-HOI benchmark. Though it is inevitable that these advancements introduce additional computational overhead, the trade-off is generally considered acceptable given the significant improvements in task performance and the ability to generalize better to previously unseen scenarios. Therefore, an important consideration in future research is the balance between computational cost and the enhanced capabilities provided by large-scale pre-trained models. In this work, we implement acceleration technology for Stable Diffusion which successfully boosts the inference speed of DIFFUSIONHOI to 24.77 FPS, surpassing most one-stage methods.

A.2 Failure Case

As shown in Fig. S1, we found that failure cases primarily manifest in the following scenarios: i) scenes featuring only partial human bodies, such as arm or leg, which introduces challenges for person detection; and ii) chaotic scenes teeming with people, which causes occlusion and difficulties in identifying interactions. Despite these challenges, DIFFUSIONHOI has shown remarkable improvement over existing approaches. Additionally, the patterns of these failure cases provide valuable insights for future research.

A.3 Broader Impact

On the positive side, this work proposes a more powerful solution for recognizing complex interactions between humans and objects in a scene. It is particularly effective for few- and zero-shot setups that are common in real-world scenarios. This advancement can significantly contribute to a range of applications, such as healthcare and assistive systems, smart homes and IoT (Internet of Things), security, and more. However, there are also potential negative aspects. Our method carries the risk of being used for continuous monitoring, which could raise concerns over intrusive surveillance and the unauthorized collection of personal data. Additionally, the inversion-based modeling strategy for adjusting diffusion models could be misused to create harmful or false information about individuals. Hence, it is essential to rigorously consider ethical standards and legal regulations to address privacy concerns, so as to avoid potential negative societal impacts.

B Detailed Comparison

We present a more detailed comparison of DIFFUSIONHOI with other methods in Table S1. As demonstrated, DIFFUSIONHOI consistently achieves state-of-the-art performance.

C Qualitative Results for Image Generation

We present the qualitative results of *relation-driven* sample generation in Fig. S2. It can be observed that the synthesized images exhibit realistic shapes and textures for individual instances, as well as coherent structure and layout for entire scenes. Additionally, the generated images accurately depict the interactions between humans and objects, demonstrating the effectiveness of our proposed inversion-based HOI modeling strategies.

Figure S1: Typical failure case on HICO-DET[19]. Actions highlighted in red indicate missing predictions that should be detected, while text with ~~strikethrough~~ means wrong predictions that should be removed.

D License

The V-COCO[21] and SWiG-HOI[22] datasets are released under the MIT license. The HICO-DET[20] dataset is released under the CC0: Public Domain license. The weight of Stable Diffusion is released under CreativeML Open RAIL M License.

Table S1: Detailed comparison for regular HOI detection on HICO-DET[20] and V-COCO[21].

Method	Backbone	VL Pretrain	Default			Known Object			V-COCO	
			Full	Rare	Non-Rare	Full	Rare	Non-Rare	AP^{S1}_{role}	AP^{S2}_{role}
iCAN[83] [BMVC18]	R50	-	14.84	10.45	16.150	16.26	11.33	17.73	45.3	-
UnionDet[23] [ECCV20]	R50	-	17.58	11.72	19.33	19.76	14.68	21.27	47.5	56.2
DRG[105] [ECCV20]	R50-FPN	-	19.26	17.74	19.71	23.40	21.75	23.89	51.0	-
PPDM[24] [CVPR20]	HG104	-	21.73	13.78	24.10	24.58	16.65	26.84	-	-
HOTR[41] [CVPR21]	R50	-	23.46	16.21	25.60	-	-	-	55.2	64.4
ConsNet[106] [MM20]	R50	-	24.39	17.10	26.56	-	-	-		
AS-Net[40] [CVPR21]	R50	-	28.87	24.25	30.25	31.74	27.07	33.14	53.9	-
QPIC[42] [CVPR21]	R50	-	29.07	21.85	31.23	31.68	24.14	33.93	58.8	61.0
CDN[44] [NeurIPS21]	R50	-	31.78	27.55	33.05	34.53	29.73	35.96	62.3	64.4
CPChoi[84] [CVPR22]	R50	-	29.63	23.14	31.57	-	-	-	63.1	65.4
MSTR[107] [CVPR22]	R50	-	31.17	25.31	32.92	34.02	28.83	35.57	62.0	65.2
UPT[86] [CVPR22]	R50	-	31.66	25.94	33.36	35.05	29.27	36.77	59.0	64.5
STIP[85] [CVPR22]	R50	-	32.22	28.15	33.43	35.29	31.43	36.45	66.0	70.7
IF-HOI[108] [CVPR22]	R50	-	33.51	30.30	34.46	36.28	33.16	37.21	63.0	65.2
ODM[109] [ECCV22]	R50-FPN	-	31.65	24.95	33.65	-	-	-	-	-
Iwin[87] [ECCV22]	R50-FPN	-	32.03	27.62	34.14	35.17	28.79	35.91	60.5	-
MCPC[88] [ECCV22]	R50	-	35.15	33.71	35.58	37.56	35.87	38.06	63.0	65.1
PViC[89] [ICCV23]	R50	-	34.69	32.14	35.45	38.14	35.38	38.97	62.8	67.8
PViC†[89] [ICCV23]	Swin-L	-	44.32	44.61	44.24	47.81	48.38	47.64	64.1	70.2
CTAN[50] [CVPR22]	R50	CLIP	31.71	24.82	33.77	33.96	26.37	36.23	60.1	
SSRT[47] [CVPR22]	R50	CLIP	30.36	25.42	31.83	-	-	-	63.7	65.9
DOQ[49] [CVPR22]	R50	CLIP	33.28	29.19	34.50	-	-	-	63.5	-
GEN-VLK[3] [CVPR22]	R50	CLIP	33.75	29.25	35.10	37.80	34.76	38.71	62.4	64.4
HOICLIP[19] [CVPR23]	R50	CLIP	34.69	31.12	35.74	37.61	34.47	38.54	63.5	64.8
CQL[90] [CVPR23]	R50	CLIP	35.36	32.97	36.07	38.43	34.85	39.50	66.4	69.2
ViPLO[91] [CVPR23]	ViT-B	CLIP	37.22	35.45	37.75	40.61	38.82	41.15	62.2	68.0
AGER[92] [ICCV23]	R50	CLIP	36.75	33.53	37.71	39.84	35.58	40.2	65.7	69.7
RmLR[93] [ICCV23]	R50	MobileBERT	36.93	29.03	39.29	38.29	31.41	40.3	63.8	69.8
ADA-CM[94] [ICCV23]	ViT-L	CLIP	38.40	37.52	38.66	-	-	-	58.6	64.0
DIFFUSIONHOI	VQGAN	Stable Diffusion	**38.12**	**38.93**	**37.84**	**40.93**	**42.87**	**40.04**	**66.8**	**70.9**
DIFFUSIONHOI	ViT-L	Stable unCLIP	**42.54**	**42.95**	**42.35**	**44.91**	**45.18**	**44.83**	**67.1**	**71.1**

†: Models built upon advanced object dector, *i.e.*, $\mathcal{H}$-Deform-DETR[95].

human-swing-baseball_bat

human-walk-dog

human-operate-microwave

human-hold-surfboard

human-jump-skateboard

human-hug-sheep

Figure S2: Qualitative results for relation-driven image generation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contributions have been claimed in abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The computational efficiency has been discussed in §4.2. The limitation of the proposed algorithm has been discussed in Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: There is no theoretical result.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The implementation details have been provided in §3.5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code is available at https://github.com/Oliliulei/DiffusionHOI.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setup, including data splits, training and testing detailed, are provided in §4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The prior work included into comparison does not provide error bar. It could be too expensive and time-consuming to run the experiments on large-scale datasets for multiple times.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computer resources are described in §4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This work conforms the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The border impacts is provided in Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The proposed method uses pre-trained models. This proposed methods is safe under the safeguards of adopted pre-trained models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The V-COCO and SWiG-HOI datasets are released under the MIT license. The HICO-DET dataset is released under the CC0: Public Domain license. The weight of Stable Diffusion is released under CreativeML Open RAIL M License.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There is no new assets released in this work.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: There is no research with human subjects in this work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: There is no research with human subjects in this work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.