Supplementary Materials of InterFusion

Sisi Dai[1] Wenhao Li[1] Haowen Sun[2] Haibin Huang[3] Chongyang Ma[3]
Hui Huang[2] Kai Xu[1⋆] Ruizhen Hu[2⋆]

[1] National University of Defense Technology
[2] Shenzhen University
[3] Kuaishou Technology
https://sisidai.github.io/InterFusion/

Outline

In this work, we present InterFusion[i], a novel zero-shot text-driven 3D human object interaction generation method. We now provide supplementary details in this document, which is arranged as follows:

(1) Sec. A illustrates the implementation details about the methods;
(2) Sec. B conducts more experiments to verify the superiority of InterFusion;
(3) Sec. C discusses the application potential, limitations and future work.

We also encourage readers to watch our supplementary videos on the project page, which provide more visual representations and perspectives to showcase the 3D properties of our generated human-object interactions.

A Implementation Details

We implement InterFusion with threestudio [**?**]. Specifically, we leverage the multi-resolution hash-grid implementation of implicit volumes in threestudio, along with a Multi-Layer Perceptron (MLP) for predicting density and color values.

Shading. We adopt Lambertian shading with randomly sampled point light during training. We consider three types of shading, including albedo, diffuse and textureless. During training, the shading types of H-NeRF and O-NeRF are enforced to be same for better convergence.

Prompting. We use one prefix and two suffixes in prompting. We empirically use the prefix "a photo of" to enhance optimization. Additionally, we use the first suffix "8K, HD" to improve the resolution and quality. The second suffix is view-dependent and based on the camera location sampled randomly, similar to that in [**?**]. Specifically, this view-dependent suffix is set to "overhead view" at elevation angles above 60°. For elevation angles below 60°, the corresponding

⋆ Corresponding authors: kevin.kai.xu@gmail.com; ruizhen.hu@gmail.com.
[i] Our code would be accessible at https://github.com/sisidai/InterFusion.

text embedding is a weighted interpolation of text embeddings attached with suffixes "front view", "side view", and "back view", where weights are dependent on the azimuth angle.

Regularizations. Similar to [**?**], several regularization terms are incorporated to enhance the optimization of H-NeRF and O-NeRF, constituting L_{reg}. We employ the orientation loss from Ref-NeRF [**?**] to encourage normal vectors, that of points along the ray when they are visible, to be forward-facing but not backward-facing to the camera:

$$\mathcal{L}_{orient} = \sum_i \text{stopgrad}(w_i)\max(\boldsymbol{n_i} \cdot \boldsymbol{v}, 0)^2. \tag{1}$$

To encourage the separation from the background and discourage unnecessary floating in empty space, there is also a regularization on the opacity (accumulated the alpha value along each ray):

$$\mathcal{L}_{opacity} = \sqrt{(\sum_i w_i)^2 + 0.01}. \tag{2}$$

Optimization. Recall that our total loss for optimization is:

$$\mathcal{L} = \mathcal{L}^H_{\text{SDS}} + \lambda_1 \mathcal{L}^O_{\text{SDS}} + \mathcal{L}^H_{\text{geo}} + \lambda_2 \mathcal{L}^O_{\text{geo}} + \lambda_3 L_{\text{reg}}. \tag{3}$$

λ_1, λ_2 and λ_3 are the corresponding loss weights, and we adopt weight annealing for them during the optimization process. Specifically, over a total of 10,000 iterations, the weight λ_1 linearly increases from 0 to 1, adding 0.1 every 1,000 iterations. At the outset, the SDS guidance of interaction plays a crucial role initially, providing a good initialization for the object. As the optimization progresses, confidence in the density of the object increases. The weight λ_1 continuously augments, ensuring that the generated components align with the semantic context of the object. As for the weight λ_2, it is empirically set to 0.001 during the initial and final 1,000 iterations, 0.01 during iterations 1,000-2,000 and 8,000-9,000, and 0.1 for the remaining iterations in between. As this weight corresponds to the anchor pose occupancy penalty for the object model, starting with a small value ensures the generation of well-initialized objects from the anchor. Adopting a larger value gradually aids in eliminating redundant human information introduced during initialization, coupled with the SDS guidance from the object. The subsequent decrease in value encourages the final object to contact the human sufficiently, thus aligning more closely with the semantic context of the interaction. The weight λ_3 for the regularization term is constant throughout the optimization process.

Training details. During training, images are rendered under randomly sampled camera views at the resolution of 64×64. We use DeepFloyd[ii], a pre-trained

[ii] `https://github.com/deep-floyd/IF`

diffusion model, with time steps from $t \sim \mathcal{U}(0.02, 0.98)$, and set the weighting function of the time step $\omega(t)$ as 1 consistently. The classifier-free guidance strength is set to 20. We use Adam optimizer [**?**] with a learning rate of 0.01. For each 3D scene, the optimization is performed on a single Tesla V100 GPU with 10,000 iterations, requiring approximately 1.5 hours.

B Experiments

B.1 Additional Comparisons

Additional qualitative comparisons. We have presented qualitative comparisons with several baseline methods, including DreamFusion [**?**], Magic3D [**?**], and TextMesh [**?**]. Qualitative comparisons of additional interaction types with them are shown in Figure 1. For fairness, the inputs of baselines are also prompted with the same prefix and suffixed as ours. Note that there are two stages in Magic3D: the first NeRF-based [**?**] stage as a coarse stage, and the second DMTet-based [**?**] stage as a refinement stage for higher quality results. We compare our method with its first NeRF-based stage, as ours can be also integrated with a refinement stage.

Fig. 1: Additional qualitative comparison results with baseline methods.

We now provide qualitative comparisons with our designed object-centric baseline (Ours-OC). With object priors, the object-centric baseline more easily generates complete interaction scenes than other baseline methods that start

Table 1: Quantitative comparisons of more baselines and metrics.

Method	DreamFusion [?]	Magic3D [?]	TextMesh [?]	MVDream	ProlificDreamer	Ours
R-Precision(%)	68.8	73.8	47.5	77.0	67.2	**83.6**
FID_{CLIP}(%)	68.4	70.0	69.8	65.5	64.8	**63.7**

from scratch. Nevertheless, the lack of sufficient human body priors still hampers the ability to achieve complete interaction generation. As seen in Figure 2, the object-centric baseline still struggles to generate the full human body, with noticeable absences of body parts involved in interactions and the presence of redundant artifacts.

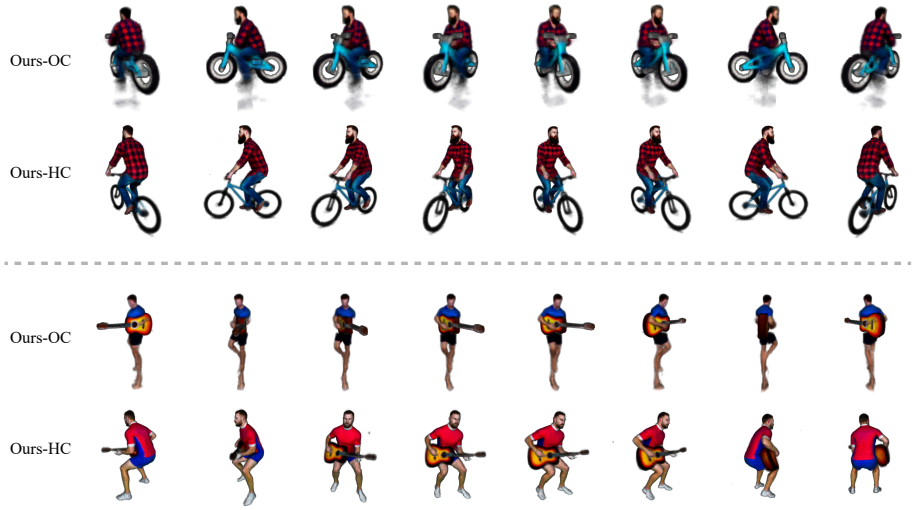

Fig. 2: Comparison between the object-centric baseline (Ours-OC) and InterFusion (Ours-HC) across multiple views, given the text prompt "a man with a full beard wearing a flannel shirt riding a bike" (top) and "a man in a rugby jersey and cotton shorts playing the guitar" (bottom).

Moreover, We further compare our method with recent avatar generation methods, including DreamAvatar [?] and AvatarCraft [?]. Visual comparisons are shown in Figure 3 and InterFusion achieves competitive quality.

Additional quantitative comparisons. We additionally incorporate CLIP R-Precision and FID_{CLIP} into our evaluation metrics, and conduct evaluation to include recent advancements in text-to-3D generation, i.e. MVDream [?] and ProlificDreamer [?]. The CLIP R-Precision metric [?], from the text-to-image generation literature, is the retrieval accuracy with which CLIP [?] retrieves

the matching caption among rendered images, evaluates the relevance of the retrieved 3D models to the textual queries. FID_{CLIP} assesses the visual fidelity of our generated scenes within the CLIP feature space. These metrics, as shown in Table 1, underscore our method's robustness, with our approach outperforming all the methods across all these dimensions.

Assessment details for GPT-4V selection. Though the CLIP score is designed to measure how closely an image aligns with the input text, it falls short in capturing finer details, thus resulting in less pronounced differences in metrics. Inspired by the powerful image understanding capabilities of GPT-4V[iii], we further evaluate the performance of baselines and InterFusion over 61 text prompts, using GPT-4V for selection, named GPT-4V select. Specifically, we ask GPT-4V to select one from all generated results with the most 3D justifiability such as full human body, complete object, and correct physical interaction, and then return the index. Note that no in-context examples are given for guidance. Meanwhile, the given order of generated results is randomly shuffled. The answers are summarized in Table **??**. We also encourage readers to utilize GPT-4V for evaluating the results we have presented, where readers would receive more detailed responses.

B.2 Additional Ablations

We provide additional visual examples for loss terms of pose-guided generation in Figure 4, where multiple views of generated results are also provided. As for details of GPT-4V selection, we similarly employ GPT-4V to evaluate the efficiency of loss terms over 61 text prompts. Differently, the object view and the interaction view are both given to GPT-4V in ablations (given object-only in the upper half and human-object in the lower half of the image). We then ask GPT-4V to select one from all generated results with the most 3D justifiability, considering both the complete object and correct physical interaction, and then return the index. No in-context examples are given and the given order of generated results is also randomly shuffled. The answers are summarized in Table **??**. We also recommend readers use GPT-4V for evaluating the results of our ablations.

In general, results generated by our full pipeline are mostly selected, showcasing the collective efficacy of all loss terms. As seen in the 7th and 8th column in Figure 4, results of the absence of SDS from object are mixed with noise from the human body, thus are rarely selected by GPT-4V when considering both the object view and the interaction view. Without the geometric constraint, generations are unstable, resulting in object degeneration and flawed interactions. In some scenarios, the generated object penetrates the human body, with semantically inconsistent interactions (top of the 3rd and 4th column). In rare cases, though the object also intersects, the final interaction remains plausible (bottom

[iii] `https://chat.openai.com/`

of the 3rd and 4th column). Sometimes, such cases would be selected by GPT-4V due to its stochastic nature.

Fig. 3: Comparisons with recent avatar generation methods, given the text prompt "a man with blond hair wearing a brown leather jacket".

C Application Potential, Limitations and Future Work

C.1 Application Potential

Controls for the generated 3D content are challenging and desired. Our InterFusion supports controllable text-conditioned editing, providing users more control over the generated 3D models. Following DreamFusion, we conduct the control by refining the generated 3D model under new given text conditioning. While general text-conditioned editing would modify the geometry and texture in all differing spatial locations, our representation with decomposed human and object enables editing for human-only or object-only within controlled spatial locations. The resulting model preserves the complex spatial relations consistent with the interaction type.

In Figure 5, we show the model trained with the base prompt for <push, shopping cart>. Results show that we can refine the human part of the scene model only, e.g. changing the "hipster man" to "elderly hipster man" or "hipster man with a brown leather jacket". Meanwhile, we can also tune the object part of the scene model only, e.g. changing the "shopping cart" to "red shopping cart" or "shopping cart full of fruits and vegetables". Both geometry and texture are supported to be edited under new given text conditioning, with the interaction relationship maintained.

C.2 Limitations and Future Work

Generating high-fidelity 3D HOI, especially in a zero-shot text-to-3D manner without 3D supervision, is an extremely challenging problem. Our current method primarily focuses on optimizing the global spatial relationship for full-body interactions, thus some inaccuracies in local may still exist, e.g. penetrations at hands. The additional module for hands could be induced in the future.

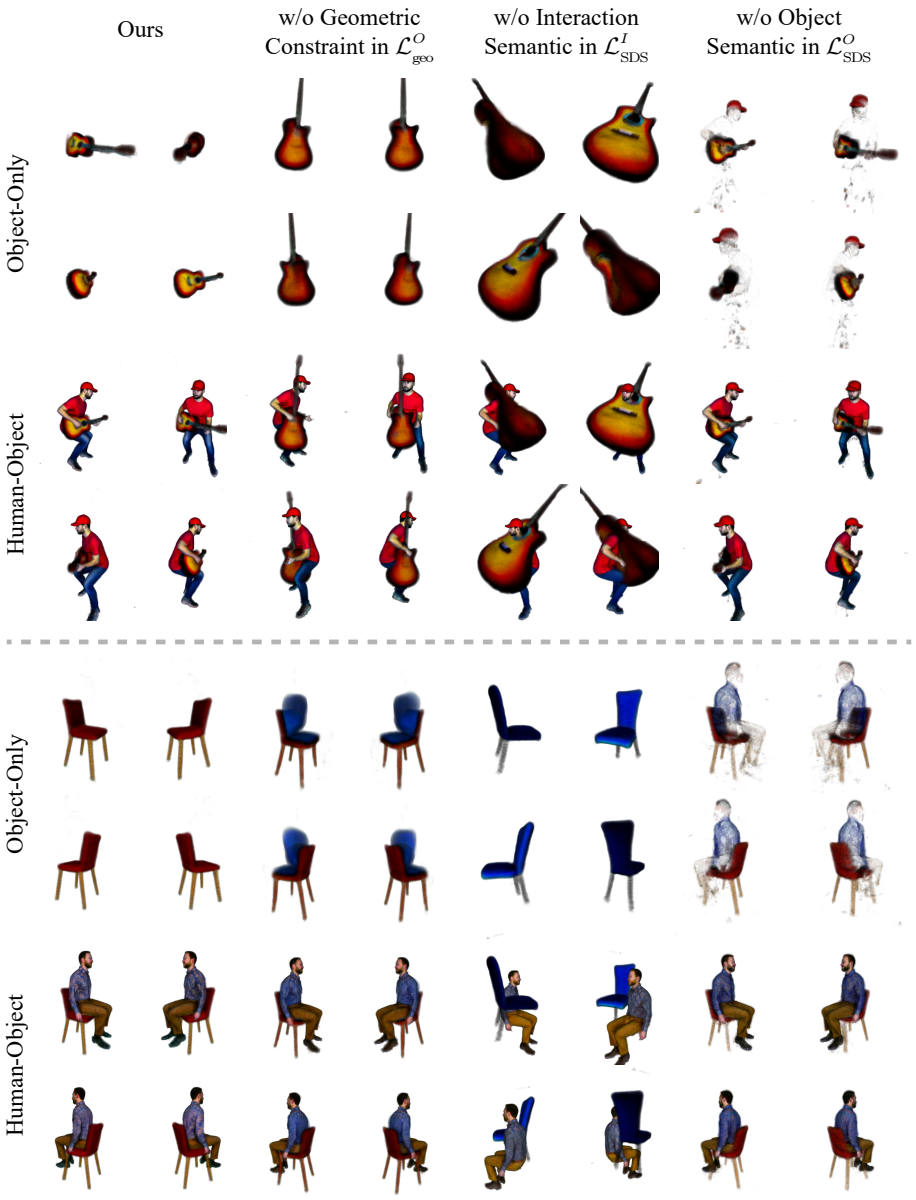

Fig. 4: Visual ablations across multiple views for loss terms during the pose-guided generation process, given the text prompt "a man wearing a red baseball cap playing the guitar" (top) and "a person in a paisley print shirt and corduroy pants sitting on a chair" (bottom).

Fig. 5: InterFuison provides a flexible way for controllable editing of human-object interactions, enabling geometry and texture manipulations for either humans or objects through simple adjustments in the corresponding text prompts.

Our method is also limited by the capabilities of currently used visual language models (VLMs). The progression of VLMs would benefit our method directly. Additionally, we are interested in employing large language models (LLMs) to further enhance our method. Meanwhile, the human-object interaction results generated by our current method are static, we believe extending our framework to incorporate dynamic HOI motions is a good direction for future work.