Human-object Interaction Image Generation

Peiye Zhuang

November 25th 2024

1 Motivation

Problem Statement. Human-Object Interaction Image Generation focuses on generating images that seamlessly integrate specified objects into given human avatar images. This task holds significant value in numerous applications, including personalized advertisement, e-commerce, and human-centric AI simulations. The ability to produce realistic and contextually appropriate interactions between humans and objects enhances user engagement and usability in these domains.

Challenges. The task presents several critical challenges: (1) **Preserving Identity.** It is essential to maintain the distinct identities of both the human avatars and the objects being inserted. Any loss or distortion of these identities can lead to unrealistic or unrecognizable outputs, limiting the practicality of the generated images. (2) **Photo-Realism and Coherence.** The generated images must exhibit a high level of visual fidelity, ensuring that the inserted objects appear naturally within the scene. This includes appropriate lighting, shadowing, and contextual placement to achieve coherence and avoid breaking the illusion of reality. (3) **Efficiency and Scalability.** Given the vast diversity of human avatars and objects, it is crucial to design a model that is efficient in terms of computational resources and adaptable to various scenarios. The model should be capable of generating high-quality results in a time-effective manner, supporting scalability for real-world applications.

To address the challenges of Human-object interaction image generation, several approaches have been proposed, which are detailed below:

- **Ip-adapters** are designed to enhance base image generation models like Stable Diffusion and Flux without requiring fine-tuning the base generative models. This approach is particularly attractive due to its generalizability across various inputs. Furthermore, IP-adapters simplify the integration process, enabling streamlined workflows for vary tasks.

  Cons: However, despite these advantages, IP-adapters currently fall short of meeting the production-level quality. Specifically, they struggle to consistently preserve the distinct identities of both human avatars and objects. For these reasons, their current limitations in identity preservation

and output quality make them less suitable for addressing this task at the current stage.

- **Suject-driven inpainting** presents a compelling approach for object insertion tasks by formulating the problem as an inpainting task. Representative methods such as AnyDoor[1] and SeedEdit[2] have demonstrated promising potential in preserving the identities of instances. These methods effectively integrate objects into a scene while maintaining their core characteristics, making them valuable for scenarios where identity preservation is a priority.

  Cons: (1) Location Guidance Requirement. These approaches rely on explicit location guidance to define where the object should be inserted. This dependency can reduce automation and scalability, especially in tasks requiring dynamic and adaptive placement. Some automatic locating approaches might be proposed to this end. **(2) Flexibility in Human-Object Interaction.** While preserving identity is critical for the proposed human-object interaction image generation task, other aspects, such as natural variations in human gestures, clothing, and object positioning, should also adapt naturally to the context. Subject-driven inpainting methods tend to constrain generated outputs, keeping them overly similar to the input images. This rigidity limits the creative flexibility needed to achieve realistic and contextually coherent interactions.

- **Dreambooth and its variants** [3] represent a state-of-the-art approach for generating high-quality images that preserve the identity of specific input instances, by learning a specific text embedding for each instance. To the best of my knowledge, no existing feed-forward model achieves DreamBooth-level quality. **Cons: Time-consuming.** The need for per-instance optimization is computationally expensive and time-consuming. This process can take up-to-hours per instance, making it less practical to use. **Difficult on composition tasks.** Moreover, it is difficult to use Dreambooth for composition tasks.

- **Unified image generation models** offer a flexible approach to image synthesis by considering free-form conditionings — such as text, images, human skeletons, or other modalities — as tokens in a unified framework. These models employ transformer-based architectures to process diverse input queries, effectively unifying the generation process across multiple conditioning types. A representative example of this approach is Omnigen[4]. Because of the flexibility, scalability, the inherent power of leveraging large language models (LLMs), I chose Omnigen to address the task.

[1] https://arxiv.org/pdf/2307.09481
[2] https://arxiv.org/pdf/2411.06686
[3] https://arxiv.org/pdf/2208.12242
[4] https://arxiv.org/pdf/2409.11340

Inputs Given Results Results

Figure 1: **Results of given examples**. The last 2 columns show results using different random seeds.

2.5/1.6 2.5/2.5 2.5/3.5 3.5/1.6 3.5/2.5 3.5/3.5

Figure 2: **Hyper-parameter analysis** with various guidescale/img_guidescale.

2 Proposed approach

Here, I present the results and analysis using OmniGen [5].

Results. The results for the given examples are shown in Figure 1.

Hyper-parameter analysis. OmniGen's performance is influenced by hyperparameters such as the guidance scale for text prompts and images. To evaluate this, I provide an ablation study in Figure 2, showing the outcomes with different guidance scale configurations.

Prompt engineering. Beyond this, I also experimented with prompt engineering; however, it did not lead to any noticeable improvement in quality. This direction might be worth further exploration to develop more effective prompts.

Additional results are shown in Figure 3.

[5] https://arxiv.org/pdf/2409.11340

Figure 3: **Additional results**.