SemGrasp: Semantic Grasp Generation via Language Aligned Discretization – *Supplementary Material* –

Kailin Li[1,2], Jingbo Wang[2], Lixin Yang[1], Cewu Lu[1,2]✉, and Bo Dai[2]

[1] Shanghai Jiao Tong University, Shanghai, China
{kailinli,siriusyang,lucewu}@sjtu.edu.cn
[2] Shanghai AI Laboratory, Shanghai, China
{wangjingbo,daibo}@pjlab.org.cn

A Experiments Details

A.1 Setting Details

Dataset Split Our dataset, **CapGrasp**, builds upon and extends the OakInk dataset [8]. As such, we adopt the same split as OakInk, with 80% of the data allocated for training, 10% for validation, and 10% for testing. We ensure that the test set includes a wide variety of objects and language instructions, thereby allowing us to evaluate the generalization capabilities of our **SemGrasp**.

MLLM prompt The prompt for our grasp-aware MLLM is specified in Tab. 2, guiding the model to generate coherent and plausible grasps from language instructions.

GPT-4 assisted evaluation We leverage the commercial GPT-4v [5] to evaluate the quality of our **SemGrasp**. This involves rendered images $\boldsymbol{I}$ of the generated grasps $\hat{\boldsymbol{G}}$, alongside the evaluation prompt outlined in Tab. 3 to GPT-4v. The model then returns a quality score reflecting both semantic similarity and physical reliability. To enhance the accuracy of GPT-4v, we take the following methods to improve the quality of the rendered images: 1) The grasp $\hat{\boldsymbol{G}}$ is mapped onto the differentiable NIMBLE model [3], which contains delicate muscle modeling and high-fidelity hand skin textures. 2) Images are rendered in Blender using the Cycles rendering engine, complemented by random lighting and camera positioning to ensure diversity.

Perceptual Score We ask 5 volunteers to rate the quality of the generated grasps $\hat{\boldsymbol{G}}$ on a 5-point Likert scale. We randomly sample 50 predicated grasps from the test set for each experiment. The evaluation indicators involve the following three aspects: 1) Semantic coherence with the provided language instructions, 2) Physical plausibility of the hand pose, and 3) Stability of the hand-object interaction. The perceptual score is the average of the ratings.

A.2 Representation Ablation Studies Details

This section elaborates on the ablation studies conducted to examine our discrete representation.

Single Token Contrary to our primary model's multi-token and hierarchical VQ-VAE structure, we explore a simplified model using a single VQ-VAE with one codebook to encapsulate the entire grasp representation.

The $<\texttt{o},\texttt{m}>$ Setting In this variant, we devise a dual-layer hierarchical VQ-VAE specifically for grasp representation that is trained from scratch. The first codebook is for the *orientation* and the second codebook is for the *manner*.

Multiple *refinement* Tokens This configuration introduces a delta VQ-VAE designed to refine the grasp pose by predicting incremental *refinement* tokens $<\texttt{r}>$ conditioned on the preceding hand grasp and object point cloud. Based on this setting, we can iteratively adjust the hand pose by applying the delta parameters to the previous grasp.

Single VQ-VAE Here, a unified VQ-VAE codebook is employed to simultaneously derive the three tokens ($<\texttt{o},\texttt{m},\texttt{r}>$), each decoded into the target pose through distinct decoding head.

Without Semantic In our primary model, the reconstruction loss is composed of three parts: the *orientation* loss $\mathcal{L}_\texttt{o}$, the *manner* loss $\mathcal{L}_\texttt{m}$, and the *refinement* loss $\mathcal{L}_\texttt{r}$. Specifically, $\mathcal{L}_\texttt{o}$ only supervises the $\hat{\boldsymbol{T}}$ and $\mathcal{L}_\texttt{m}$ mainly focuses on the $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}$. We simply set the items that are not supervised to zero in the loss function Eq. (1).

$$\begin{aligned}\mathcal{L}_{\text{rec}} = &\quad \mathcal{L}_\texttt{o} + \mathcal{L}_\texttt{m} + \mathcal{L}_\texttt{r} \\ = &\quad \|\mathcal{M}(\boldsymbol{T},\mathbf{0},\mathbf{0}) - \mathcal{M}(\hat{\boldsymbol{T}},\mathbf{0},\mathbf{0})\|_2^2 \\ &+ \|\mathcal{M}(\boldsymbol{T},\boldsymbol{\theta},\boldsymbol{\beta}) - \mathcal{M}(\hat{\boldsymbol{T}},\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\beta}})\|_2^2 \\ &+ \|\mathcal{M}(\boldsymbol{T},\boldsymbol{\theta},\boldsymbol{\beta}) - \mathcal{M}(\Delta\hat{\boldsymbol{T}}\cdot\hat{\boldsymbol{T}},\Delta\hat{\boldsymbol{\theta}}+\hat{\boldsymbol{\theta}},\Delta\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\beta}})\|_2^2 \end{aligned} \tag{1}$$

In the *without semantic* scenario, the $<\texttt{o}>$ token is not exclusively constrained to represent *orientation.* We experimentally find that, during training, the three tokens collapse into a single token, which degrades performance.

Table 1: Quantitative scores of generated descriptions under different LoRA configurations. MLLM prompt: "Could you please describe this `<grasp>` with the `{object}`?".

	BLEU-1 ↑	BLEU-4 ↑	ROUGE ↑
LoRA $r = 128, \alpha = 32$	33.9	15.6	37.9
LoRA $r = 128, \alpha = 256$	34.1	14.9	40.0
LoRA $r = 256, \alpha = 32$	32.0	14.1	37.9
LoRA $r = 256, \alpha = 256$	34.8	15.4	40.3

B Exploratory Study

In our **SemGrasp**, we focus on training a grasp-aware MLLM to synchronize three distinct modalities—grasps, object models, and language instructions—within a unified representational space. We conduct an exploratory study to investigate the effectiveness of the alignment. Our findings indicate that the MLLM is not only capable of generating semantic grasps but also demonstrates promise in the grasp captioning task. Specifically, when provided with grasp tokens, the MLLM is able to produce corresponding language descriptions that capture both the low-level details and high-level intents of the grasps in some instances (as illustrated in Fig. 1). Additionally, we report linguistic metrics, namely BLEU [6] and ROUGE [4], in Tab. 1 to quantitatively evaluate the quality of the generated descriptions. It is important to admit that these generated language descriptions do not always achieve the same level of accuracy as the ground truth. The MLLM occasionally struggles to capture the details of interactions or to hallucinate details in its language descriptions. This discrepancy underscores the inherent complexity of the caption task, which necessitates a comprehensive understanding of point clouds, intent interpretation, interaction reasoning, and natural language generation capabilities. However, it is a promising direction to explore the potential of the MLLM in the future when scaling up the model and the dataset.

C Applications Details

C.1 Physical-Plausible Dynamic Grasp using Human-like Hand

In our VR/AR application, we employ the open-source D-grasp method [1] to synthesize dynamic, human-like grasps. As described in the main paper, the reference pose $\bar{\boldsymbol{G}}$, corresponding to the language instruction $\boldsymbol{L}$, is generated using our **SemGrasp**. This dynamic grasp policy is then applied to assess the

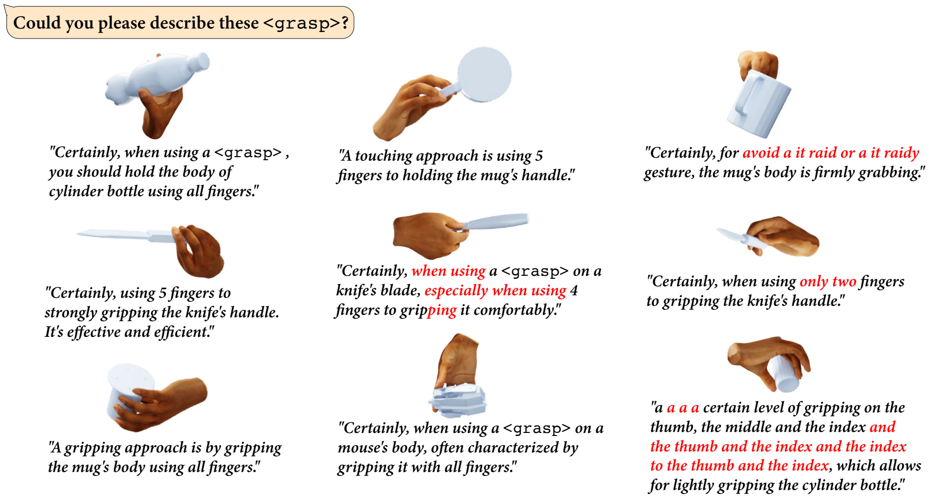

Could you please describe these <grasp>?
"Certainly, when using a <grasp> , you should hold the body of cylinder bottle using all fingers."
"A touching approach is using 5 fingers to holding the mug's handle."
"Certainly, for avoid a it raid or a it raidy gesture, the mug's body is firmly grabbing."
"Certainly, using 5 fingers to strongly gripping the knife's handle. It's effective and efficient."
"Certainly, when using a <grasp> on a knife's blade, especially when using 4 fingers to gripping it comfortably."
"Certainly, when using only two fingers to gripping the knife's handle."
"A gripping approach is by gripping the mug's body using all fingers."
"Certainly, when using a <grasp> on a mouse's body, often characterized by gripping it with all fingers."
"a a a a certain level of gripping on the thumb, the middle and the index and the thumb and the index and the index to the thumb and the index, which allows for lightly gripping the cylinder bottle."

Fig. 1: Results on the grasp caption task. The red text indicates the mistakes or the hallucinations of our model.

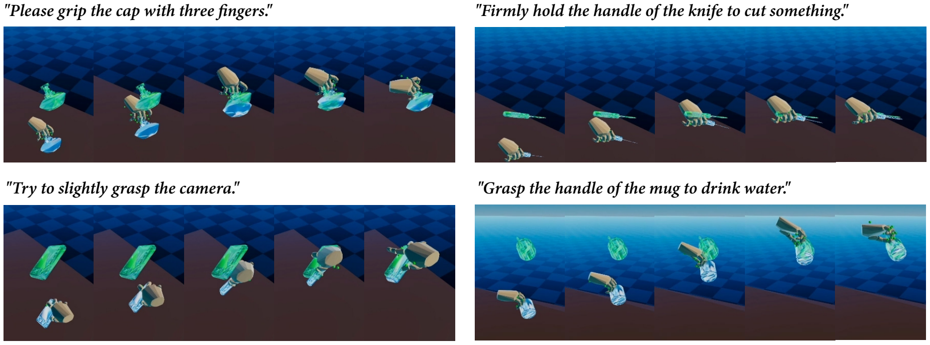

"Please grip the cap with three fingers."
"Firmly hold the handle of the knife to cut something."
"Try to slightly grasp the camera."
"Grasp the handle of the mug to drink water."

Fig. 2: Synthesis of human-like grasps motion in AR/VR application.

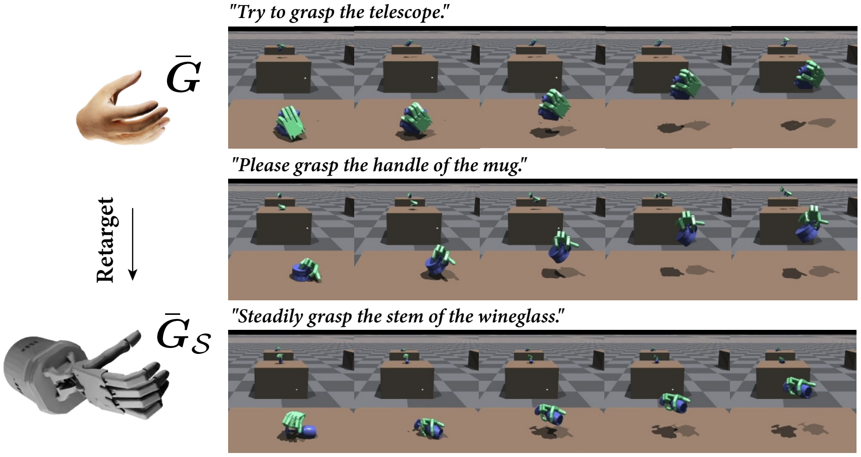

"Try to grasp the telescope."
"Please grasp the handle of the mug."
"Steadily grasp the stem of the wineglass."
$\bar{G}$
Retarget
$\bar{G}_{\mathcal{S}}$

Fig. 3: Application in robotics. We verify our results by retargeting the generated grasp to the dexterous ShadowHand. The left image displays the adaptation outcomes, while the right image illustrates the grasp execution process.

feasibility of the generated grasps. To ensure alignment with real-world scenarios, we rotate the hand-object grasp pair so the palm faces toward the table. Given that the OakInk dataset does not provide object weight information, we assign a hypothetical weight of 300g to each object for the purpose of this evaluation. Samples of the generated dynamic grasps are illustrated in Fig. 2. In our analysis, any relative sliding between the hand and the object exceeding 4cm is classified as a failure. Based on this criterion, we report a success rate of 62.9% for our generated grasps on the test set.

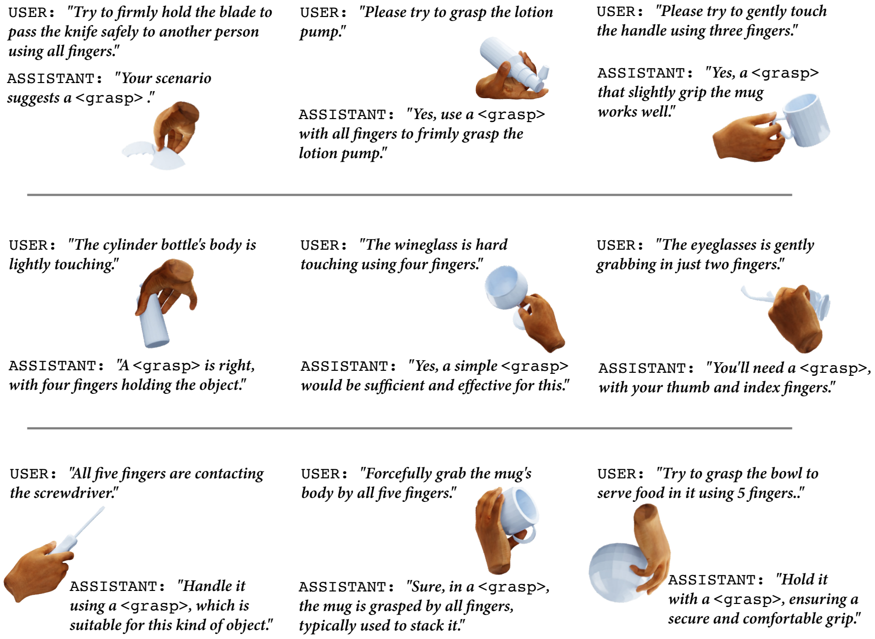
C.2 Physical-Plausible Dynamic Grasp using ShaodowHand

Given the distinct morphological features and DoFs between the human hand and the ShadowHand, we devise a specialized pipeline for adapting the generated grasp $\bar{\boldsymbol{G}}$ to the ShadowHand model $\bar{\boldsymbol{G}}_{\mathcal{S}}$. We manually select several corresponding keypoints on both the MANO and ShadowHand models, with a particular emphasis on the fingertips. For each MANO-based grasp $\bar{\boldsymbol{G}}$, the corresponding ShadowHand grasp $\bar{\boldsymbol{G}}_{\mathcal{S}}$ is optimized by aligning these selected keypoints. To mitigate issues such as unnatural finger movements and potential finger collisions, we introduce a loss function that imposes angular constraints on the ShadowHand's joints, thereby promoting physically plausible adaptations. The outcome of this fitting process is illustrated in Fig. 3 left. Following this adaptation, the refined grasp $\bar{\boldsymbol{G}}_{\mathcal{S}}$ is executed using UniDexGrasp's pretrained policy. To enhance the fidelity of collision detection, object meshes are preprocessed using Manifoldplus [2], followed by convex decomposition algorithms [7]. The results of these grasp executions are displayed in Fig. 3 right, showcasing the practicality and effectiveness of our methodology in the field of embodied robotics.

D CapGrasp collection

Prompts As mentioned in our main paper, we craft a set of prompts to direct both GPT-4 and GPT-4v in generating high-quality annotations automatically. For high-level details concerning manipulation intent and grasp status, two specialized prompts are utilized. These are detailed in Tabs. 4 and 5, designed to annotate high-level insights based on the contact information and images respectively. Additionally, to foster the generation of conversational content, another prompt is crafted to steer GPT-4 in creating conversation templates. This specific prompt, intended to enrich our dataset with conversation annotations, is outlined in Tab. 6.

Dataset Visualization Our **CapGrasp** dataset is showcased in Fig. 4, featuring an array of objects spanning various categories, shapes, and functionalities. Accompanying language instructions encompass a broad spectrum of grasp intentions and object interactions, with each object annotated with at least 8-10 distinct manipulation intents. This rich diversity is essential for the training of our **SemGrasp**, enabling the generation of grasps that are not only semantically coherent but also physically plausible across diverse scenarios.



USER: *"Try to firmly hold the blade to pass the knife safely to another person using all fingers."*

ASSISTANT: *"Your scenario suggests a* <grasp> *."*

USER: *"Please try to grasp the lotion pump."*

ASSISTANT: *"Yes, use a* <grasp> *with all fingers to frimly grasp the lotion pump."*

USER: *"Please try to gently touch the handle using three fingers."*

ASSISTANT: *"Yes, a* <grasp> *that slightly grip the mug works well."*

USER: *"The cylinder bottle's body is lightly touching."*

ASSISTANT: *"A* <grasp> *is right, with four fingers holding the object."*

USER: *"The wineglass is hard touching using four fingers."*

ASSISTANT: *"Yes, a simple* <grasp> *would be sufficient and effective for this."*

USER: *"The eyeglasses is gently grabbing in just two fingers."*

ASSISTANT: *"You'll need a* <grasp>*, with your thumb and index fingers."*

USER: *"All five fingers are contacting the screwdriver."*

ASSISTANT: *"Handle it using a* <grasp>*, which is suitable for this kind of object."*

USER: *"Forcefully grab the mug's body by all five fingers."*

ASSISTANT: *"Sure, in a* <grasp>*, the mug is grasped by all fingers, typically used to stack it."*

USER: *"Try to grasp the bowl to serve food in it using 5 fingers.."*

ASSISTANT: *"Hold it with a* <grasp>*, ensuring a secure and comfortable grip."*

Fig. 4: Visualization of our CapGrasp dataset.

Table 2: System Prompt of MLLM.

- A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. The assistant can understand the information of the three-dimensional object model provided by the user, and combine the knowledge of human hand grasping to assist the user. The following is the object model information: $\boldsymbol{O}$.

Table 3: Prompt of evaluation.

- Your task is to evaluate the alignment between a hand pose and its corresponding textual description, focusing on the grasp's intent, specific object parts engaged by the hand, and the contact dynamics between fingers and the object. Disregard the background and any textures on both the hand and object. Assessments should consider the physical feasibility of the grasp. Scores will range from 0 to 100, where 100 signifies perfect alignment and a physically plausible grasp, while 0 indicates a significant misalignment or a grasp that seriously defies physical principles. Just directly give the final score, such as: 95. Here is the description $\boldsymbol{L}$ and the grasp image $\boldsymbol{I}$:

Table 4: Prompt of contact based on high-level annotations.

- Given a formatted input describing which part of an object is grasped by a human hand and how many fingers are contacting the object, your task is to generate a set of grasp purposes or add more grasping details, such as the grasping method, the grasping force, etc. Please provide a comprehensive list (8 - 10 phrases) of potential grasp purposes or additional grasping details for each object and its corresponding grasped part.
- For example: #input: [{OBJ: "mug", PART: "handle", FINGERS: 5}, {OBJ: "pen", PART: "", FINGERS: 3}] #output: [{"to drink", "to make a toast"}, {"to write", "to underline or highlight text"}] Please note that some PART may be empty, indicating that the entire object is grasped.

Table 5: Prompt of image-based high-level annotations.

- Deduce the manipulation intent and the grasp force status from an image depicting hand-object interaction. Your response should focus on the object's contact part, affordance, hand grasp types, and hand-finger status to support your answer. Please ignore the background and object texture when deducing. Provide a clear and concise answer of the manipulation intent and grasp force status, following the example: {"intent": "to cut something", "status": "firmly grasping"}. Your analysis should be thorough and accurate, considering all relevant aspects of the hand-object interaction to support your deductions effectively.

Table 6: Prompt of conversation templates generation.

- Please provide conversation templates related to the topic of "human grasping the object." The conversation should incorporate the optional elements: 1. {finger}: e.g. "four fingers", "only one finger" 2. {status}: e.g. "firmly contacting", "softly touching" 3. {intent}: e.g. "pour water", "cut something", "toast", "transfer food onto a plate" 4. {object}: e.g. "mug", "bottle" Along with the required element of <grasp> as a state noun. Build one round of conversation using these elements, allowing for flexibility and creativity in the conversation templates. You should focus on creating dialogue that reflects human interaction related to grasping objects, considering various scenarios and details provided by the optional elements. Here is some examples: {"USER": "Try to {status} the {object} using {finger}". "ASSISTANT": "Sure, here is the <grasp>."} {"USER": "{finger} are {status} the {object} {intent}". "ASSISTANT": "This sounds like a typical <grasp>."}

References

1. Christen, S., Kocabas, M., Aksan, E., Hwangbo, J., Song, J., Hilliges, O.: D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In: CVPR (2022)
2. Huang, J., Zhou, Y., Guibas, L.: Manifoldplus: A robust and scalable watertight manifold surface generation method for triangle soups. arXiv preprint arXiv:2005.11621 (2020)
3. Li, Y., Zhang, L., Qiu, Z., Jiang, Y., Li, N., Ma, Y., Zhang, Y., Xu, L., Yu, J.: Nimble: a non-rigid hand model with bones and muscles. ACM TOG (2022)
4. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out (2004)
5. OpenAI: Gpt-4v(ision) system card (2023), `https://openai.com/research/gpt-4v-system-card`
6. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002)
7. Wei, X., Liu, M., Ling, Z., Su, H.: Approximate convex decomposition for 3d meshes with collision-aware concavity and tree search. ACM TOG (2022)
8. Yang, L., Li, K., Zhan, X., Wu, F., Xu, A., Liu, L., Lu, C.: Oakink: A large-scale knowledge repository for understanding hand-object interaction. In: CVPR (2022)