Learning 3D Object Spatial Relationships from Pre-trained 2D Diffusion Models ——Supplemental Material——

Sangwon Baik[1] Hyeonwoo Kim[1] Hanbyul Joo[1,2]
[1]Seoul National University [2]RLWRLD

https://tlb-miss.github.io/oor/

Figure S1. **Controllability of Diffusion Models Through Text Prompts.** Each is an output image of diffusion generated with the following prompt: **(a)** "A table with a teacup on top." **(b)** Adding "White background." to the end of the prompt. **(c)** Specifying the shape and texture of the table as "four-legged rectangular marble table." **(d)** Specifying the camera view via ", from a diagonal angle."

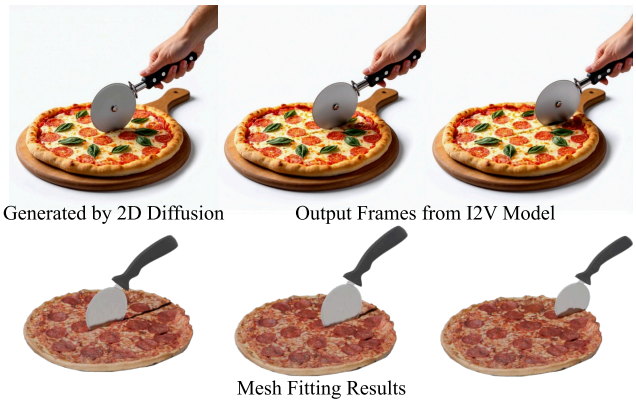

Figure S2. **Image Augmentation via Image-to-Video Model.** We diversify scenes with dynamic object-object spatial relationships using the image-to-video model.

A. Details on Dataset Generation

Text Prompts for Image Generation. The controllability of diffusion via text prompts offers advantages in learning OOR from synthetic images over real-world images by generating realistic OOR images while simultaneously enhancing their learnability through precise control. Fig. S1 illustrates this clearly. (a) shows an image generated with a simple prompt: "A table with a teacup on top." While the image is highly realistic, it poses challenges for learning the OOR between the "table" and "teacup" because the full shape of the table is not visible. (b) shows the result of adding "White background." to the end of the prompt, which directs focus to the two objects and ensures that their full shapes are captured within the image frame without the need for additional context. (c) demonstrates control over the object's shape and texture in the generated image. This facilitates the registration of template object meshes. Finally, (d) shows how to mitigate frame size constraints by controlling the camera view. This is useful for capturing OOR between objects with large size differences, such as a table and a teacup, or between objects positioned at some distance, such as a monitor and a keyboard. We use the FLUX.1-dev [1] in all image generation.

Synthetic Image Augmentation via Video Diffusion. We further augment 2D OOR images using the I2V model [11] for contexts where dynamic OOR can be generated by humans. The motivation for using the image-to-video model is to generate a broader range of relative object relationships within each context, as image diffusion models typically produce the most representative configuration (e.g., the pizza cutter tends to be in the center of a pizza). We then use each frame from the synthesized videos as additional synthetic 2D samples, disregarding their temporal information. See Fig. S2 for the example result.

Best Template Selection. If the shape of the object in the synthetic image and the template mesh are very different, mesh registration often fails. Therefore, for each category, we collect several template meshes as candidates and select the template that best matches the object in the image. To do this, we obtain DINO [2, 7] features for M pseudo multi-view images and N mesh multi-view renderings, and select the mesh with the highest value by calculating the average of cosine similarity for $N \times M$ pairs. We collect template meshes from Sketchfab [9]. There are 96 template meshes used for data generation.

Filtering Process. Our filtering process is automated in the following steps: First, we filter out all the bad quality images where segmentation and SfM fail. In the process from SfM to feature matching, we filter out if the number of points

corresponding to each base object and target object is less than 100. The cosine similarity threshold is set to 0.7 in most cases. There may still be misalignment between the registered mesh and the point clouds. We use the Chamfer Distance from the mesh to the point clouds. The threshold is adjusted according to the scale of the registered mesh. Most of the bad samples are filtered out through a series of processes, but some cases, such as flipped meshes, may remain. We use VLM [6] to filter out the last remaining bad samples. Specifically, we render combinations of base and target objects, and then ask VLM to judge whether the multi-view images align well with the text prompt, using the same criteria as when measuring the VLM score. The filtering ratio is 0.58 to 0.92. However, since our approach is based on fully synthetic data, we can iterate this process as needed to obtain a sufficient number of high-quality 3D outputs. We obtain 30 to 216 samples per context.

B. Details on OOR Diffusion.

Our OOR diffusion is trained for 20,000 epochs, taking about 10 hours on an RTX 6000 48GB.

Architecture Implementation. We follow the implementation of ScoreNet in GenPose [14] for our score-based OOR diffusion. However, we take text as a condition instead of point clouds. For this, we introduce the T5 text encoder [8]. Also, unlike Genpose, which only deals with the scores of rotation, translation, we also consider the 3-dimensional scale. For this, in the inference reverse ODE process, we add guidance to make the scale positive. Following GenPose, we consider a 6D representation [15] for rotation. Therefore, our OOR diffusion is learned in a 15-dimensional space (6D target object rotation, 3D target object translation, 3D target object scale, 3D base object scale).

Text Context Augmentation. As proposed in Sec. 3, we perform text context augmentation to increase the generality of OOR diffusion. Through the guided prompts in Fig. S3, LLM generates various text prompts that describe a given context $\mathbf{c}$. Object categories are augmented by asking the LLM to present categories with similar shape and scale that could replace $\mathcal{B}$ and $\mathcal{T}$ in the given text context $\mathbf{c}$.

Inconsistency Loss. The inconsistency loss introduced in Sec. 3.3 is computed as the average of the following three inconsistency parts: (1) The scale variance of a global base object; (2) The pose and scale variance in the global coordinate system derived from different parents; (3) The variance of each component's ratio between the scale in the global coordinate system and the base scale in pairwise OOR, measured for parent nodes that are not global bases. Specifically, (1) corresponds to the part related to the desk in Fig. 5. OOR diffusion generates different $\mathbf{s}^{\mathcal{B}}$ for each pair, (desk, monitor), (desk, keyboard), and (desk, mouse), within the batch. In this case, (1) takes the variance of three $\mathbf{s}^{\mathcal{B}}$ as a loss term. (2) is the part corresponding to the keyboard in Fig. 5. The

```
Provide {N} English prompts in a single-line list format that describe the situation of "{prompt}". Follow these steps sequentially to complete the task:

Step 1

Carefully check the following conditions:
1) Each sentence must start with "A", "An", or "The".

2) Each sentence must include both objects: a {obj1} and a {obj2}.

3) Either object can be the subject of the sentence (each can be the subject individually, or both can be the subject together).

4) Include a variety of sentence structures, such as active voice, passive voice, and noun phrases(e.g., 'A baseball bat hitting a baseball.').

5) Do not include any objects other than the main two ({obj1}, {obj2}). Keep the sentences simple and avoid unnecessary embellishments.

6) If there are verbs that describe the same situation, use different words instead of using the same word in every sentence (increase variety). However, you must not force the use of words just to increase diversity. Choose words that are appropriate for the given situation of "{prompt}".


Step 2
After creating each sentence, carefully check whether it meets all the conditions. If any sentence does not satisfy the conditions, rewrite it.


Step 3
Once all {N} sentences are created, output them in a single-line list format.
```

Figure S3. **Guided Text prompt Provided to LLM for Text Context Augmentation.** LLM augments on text context $\mathbf{c}$ via the following guided prompt.

pose and scale of each object in the global coordinate system are obtained as many times as the number of parent nodes of the corresponding node in the scene graph. Thus, the variance of three different poses and scales in global coordinate systems of the keyboard obtained from the desk, monitor, and mouse is the loss term in (2). (3) relates to the scale ratio consistency of monitor and mouse, which are parent nodes but not the global base. For example, the monitor should maintain consistency between its scale in the global coordinate system obtained from paths in the scene graph and the $\mathbf{s}^{\mathcal{B}}$ of the (monitor, keyboard) OOR sample. They do not have to be equal, but the ratio of each component should be constant. For example, if the obtained monitor scale in the global coordinate system is $(0.5, 0.4, 0.2)$, then the $\mathbf{s}^{\mathcal{B}}$ in the (monitor, keyboard) OOR sample should be $(1.0, 0.8, 0.4)$. The variance of the ratio of each component of the relevant scales is the loss term in (3).

C. Experiments Details

Baseline Methods Details for Pairwise OOR Generation. SceneMotifCoder (SMC) [10] is an example-driven visual program learning method. It takes text prompts as input and produces 3D object alignments by selecting and arranging meshes from a mesh pool. Given an example of a GT text prompt and mesh alignments in training, SMC analyzes patterns within the input, generates programs, and updates the program when new examples are introduced. During the inference process, when an input text comes in, LLM maps it to an appropriate task, and the program produces 3D object arrangements with the retrieved meshes from the candidate mesh pool. We convert our template mesh pairs

User Study: Evaluating Object-Object Relationships in Images

In this study, you will be presented with 30 questions, each featuring three sets of multi-view images. Your task is to select the set of multi-view images that best depicts the **object-object spatial relationship** described by the given **text prompt**. Please focus solely on the spatial and relational aspects of the objects as described in the text prompt, and **do not consider** other image qualities such as texture or artistic details. Your responses will help us better understand how well these relationships are conveyed visually.

This study should take approximately 5 minutes... (text omitted)

1. **Which set of multi-view images best illustrates the object-object relationship described in the following text prompt: "A knife cuts an apple"?**
   Please select the set that most accurately represents the spatial and relational aspects as described.

   1) multi-view images generated by method A

   2) multi-view images generated by method B

   3) multi-view images generated by method C

   ⋮

Figure S4. **Questionnaire for User Study.** Participants select the multi-view image set that best captures the given OOR as instructed in the questionnaire.

obtained in the dataset generation process (Sec. 3.2) into the SMC format. Since SMC is not concerned about the relative scale between objects, we use meshes with scales in our OOR dataset as a mesh pool during inference.

SceneTeller [5] leverages LLMs for in-context learning by providing the LLM with pairs of (*GT text prompt, GT scene layout*), enabling it to generate the appropriate scene layout for a test prompt. However, existing methods only focus on the layout for placement on the plane. To generate (*GT text prompt, GT OOR*) pairs, we instruct the LLM with our world coordinate system and object canonical space. Then, we provide the GT OORs for generating corresponding text prompts. It also allows LLM to generate additional prompts for inference based on the generated GT prompts. For rotation, Euler angles format is chosen as the representation because LLMs tend to generate incomplete SO(3) matrices.

VLM Score. As described in Sec. 4, we propose the VLM score, inspired by GPTEval3D [12], to evaluate the alignment between the text context of OOR and multi-view images. We use VLM (specifically GPT-4o [6]) to compare two sets of multi-view images, each containing 10 images. These image sets are generated using our method or baselines. VLM is tasked with selecting the image set that better represents the spatial relationship between objects described in the text prompt. To ensure a fair comparison, we instruct VLM to ignore texture quality and focus solely on OOR. Fig. S6 illustrates the guided prompt we provided to VLM along with an example response from VLM.

User Study. For the user study, we randomly select one scene from each of the object pairs per method to create a total of 30 questions for pairwise OOR generation. Similar to the VLM score evaluation, participants are instructed to disregard factors such as texture quality and focus solely on the OOR. For each question, multi-view image sets generated by methods A, B, and C are presented, and participants are asked to select the method that best represents the OOR described in the text prompt. To prevent bias, the order of A, B, and C is randomized for each question. We collect responses from 92 participants in total(81 participants for multi-OOR evaluation). The detailed questionnaire structure is illustrated in Fig. S4.

Figure S5. **Ablation Study for Data Generation.** We show that applying PCA to features of points and separating the base object and target object through segmentation for matching in better results.

Methods		Fréchet distance (FD) ↓
PCA	Segmentation	
✓		1.87
	✓	1.50
✓	✓	**1.43**

Table S1. **Ablation Study for Data Generation.** We demonstrate the superiority of our data generation method through an ablation study. We compare the similarity between real data's OOR distributions and synthetic OOR distributions produced by our approach.

D. Ablation Study

We compare our OOR distribution to the real data distribution and perform an ablation study to provide further justification for our OOR data generation pipeline.

Dataset. We use the ParaHome DB [4], which captures dynamic 3D movements of humans and objects in a home environment. We extract three OOR distributions: ('cutter board', 'knife'), ('teacup, 'teapot'), and ('pan', 'salt shaker'). To exclude the approach and departure of humans relative to an object, we use the middle 70% of sequences. Since there is only one instance for each category, the scale is constant. Therefore, we construct a joint distribution for $(\mathbf{R}^{\mathcal{T}\to\mathcal{B}}, \mathbf{t}^{\mathcal{T}\to\mathcal{B}})$.

Baseline Methods. We ablate our mesh registration pipeline in Sec. 3.2 by removing point cloud separation and PCA on semantic features (SD+DINO [13]), comparing them with our full pipeline.

Metric. We use the Fréchet distance (FD) [3] to measure distribution similarity. When two distributions p and q are approximated by a multivariate Gaussian, the FD score d of

System Prompt

You are an evaluator who examines multi-view images and determines whether they align well with the text prompt.

User Prompt

You are tasked with reviewing and evaluating multi-view images generated from a text prompt describing the relationship between two objects, '{object_tuple[0]}' and '{object_tuple[1]}'.
Prompt: '{prompt}

There are a total of 20 images provided. The first 10 images are multi-view images generated using Method A, while the next 10 images are generated using Method B.

You are to carefully review and determine which method better satisfies the following criteria:
1. Both objects, '{object_tuple[0]}' and '{object_tuple[1]}', should be present in 3D space. Because these are multi-view images, there may be occlusions in some views. You need to think in 3D space, combining multi-view images rather than just looking at one image at a time.
2. The spatial relationship between the two objects, as described in the text prompt, should be clearly represented in the multi-view images.
3. The texture of the images or the geometry of each object is not a consideration. As long as the objects are recognizable, you can ignore texture and geometry. Focus solely on the object-to-object relationship.

Once again, the first 10 images are generated using method A, the next 10 images are generated using method B.
The text prompt specifies the relationship between the two objects '{object_tuple[0]}' and '{object_tuple[1]}':
'{prompt}'

Indicate which method better satisfies the criteria above and provide a brief explanation for your choice.

Please make your choice clear by including the phrase 'Choose: Method _' at the end. \nMultiple selection is not allowed. However, if it is difficult to determine which is the best, do not choose anything and just print the word 'Cannot choose'.

Example Response

To evaluate which method better satisfies the criteria, let's analyze the images:

Method A:
All images clearly show both the apple and the knife. The knife is consistently shown cutting into the apple, accurately representing the spatial relationship described in the prompt.

Method B:
All images include both the apple and the knife. However, the knife is not consistently shown cutting the apple. In some images, the knife is merely near the apple without cutting it. Based on the criteria, Method A better satisfies the requirement of depicting the knife cutting the apple.

Choose: Method A

Figure S6. **Guided Prompt Provided to VLM for VLM Score Evaluation.** Using the guided prompt above, VLM selects the preferred multi-view image set between the two generated by the different methods.

the two distributions is given by:

$$d^2 = \|\mu_p - \mu_q\|^2 + \mathrm{tr}\left(\Sigma_p + \Sigma_q - 2\left(\Sigma_p\Sigma_q\right)^{1/2}\right), \quad (1)$$

where μ_p is mean of p, μ_q is mean of q, Σ_p is covariance matrix of p, and Σ_q is covariance matrix of q. Since rotation and translation have different units in each context, we train a 3-layer MLP encoder-decoder on 50M randomly sampled rotation matrix and translation vector. Then we compute FD in the learned 128D feature space.

Results. Tab. S1 shows that our method produces closer OOR distributions to real data than baselines, validating our full pipeline. Fig. S5 further demonstrates the advantage of our segmentation and PCA modules. Without segmentation, registration often misaligns objects, and PCA enhances accuracy, yielding more realistic OOR samples.

Figure S7. **Applying Our OOR Diffusion Samples to Unseen Instances.** Our OOR diffusion still works when applied to instances other than the template meshes used to generate the dataset.

E. Generality of Our methods

Generality for Unseen Mesh Instances. Fig. S7 demonstrates the generality of our OOR modeling for unseen mesh instances. Our OOR diffusion generates appropriate relative poses and scales even for instances other than the template meshes used to generate the dataset. We consider the follow-

Figure S8. **Our OOR Diffusion Sampling Results Under Unseen Text Prompt Condition.** Our OOR diffusion also works on text prompts that are not explicitly seen in training (including new categories and spatial relations).

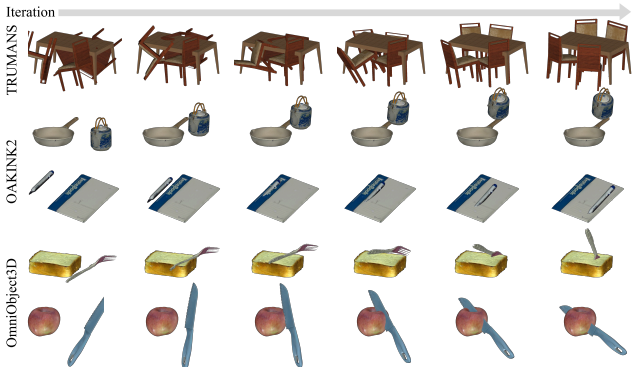

Figure S9. **Scene Editing Results on TRUMANS, OAKINK2 and OmniObject3D.** Our scene editing algorithm works on a variety of real and synthetic datasets.

ing scales to maintain the aspect ratio of each instance for both the base object and the target object:

$$\mathbf{s}' := \text{Mean}(\mathbf{s}/\text{BBOX}(\mathcal{M})) \cdot \text{BBOX}(\mathcal{M}), \qquad (2)$$

where $\mathbf{s}$ is 3-dimensional scale from OOR diffusion, and $\mathcal{M}$ is an instance mesh.

Generality for Unseen Text Inputs. Fig. S8 shows that our OOR diffusion still produces plausible outputs even for text prompts that are not seen during training. In the first example, the spatial relation "pour" is learned, but the object categories "moka pot" and "mug" are not seen during training. In the second example, both "steak" and "knife" are categories seen during training, but the spatial relationship of "cutting steak with a knife" is not learned. The last example shows a case of multi-object OOR. The spatial relation of placing a "cutting board" somewhere is not seen during training, but thanks to the generality of OOR diffusion, it is correctly placed on the "table".

Generality for 3D Scene Editing. Since our OOR can be applied regardless of objects' textures or shape details, the synthetic-to-real gap is minimal. Fig. S9 demonstrates the effectiveness of our scene editing algorithm on additional datasets which are TRUMANS, OAKINK2, and OmniObject3D, yielding convincing results.

References

[1] black-forest labs. Flux.1-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024. 1

[2] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 1

[3] DC Dowson and BV666017 Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3), 1982. 3

[4] Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. *2401.10232.arXiv.cs.CV*, 2024. 3

[5] Başak Melis Öcal, Maxim Tatarchenko, Sezer Karaoglu, and Theo Gevers. Sceneteller: Language-to-3d scene generation. In *ECCV*, 2024. 3

[6] OpenAI. Chatgpt: Chat-based ai language model. https://chat.openai.com, 2024. 2, 3

[7] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. In *TMLR*, 2023. 1

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21:1–67, 2020. 2

[9] Sketchfab. Sketchfab - publish & find 3d models online. https://sketchfab.com, 2024. 1

[10] Hou In Ivan Tam, Hou In Derek Pun, Austin T. Wang, Angel X. Chang, and Manolis Savva. Scenemotifcoder: Example-driven visual program learning for generating 3d object arrangements. In *3DV*, 2025. 2

[11] Kuaishou Technology. Kling ai. https://www.klingai.com, 2024. 1

[12] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation. In *CVPR*, 2024. 3

[13] Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *CVPR*, 2024. 3

[14] Jiyao Zhang, Mingdong Wu, and Hao Dong. Generative category-level object pose estimation via diffusion models. In *NeurIPS*, 2024. 2

[15] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 2