# Semantic Enhanced Point-E: A Semantics-Driven Advancement on 3D Model Generation

Shangyang Min
Brown University
Providence, RI 02912
shangyang_min@brown.edu

Jiaying Cheng
Brown University
Providence, RI 02912
jiaying_cheng@brown.edu

## Abstract

*We present Semantic Enhanced Point-E (SEPE), an innovative advancement to the Point-E framework that seamlessly integrates image and text modalities at the embedding stage before generating 3D point clouds. Leveraging the efficiency of Point-E's diffusion-based architecture, our method employs early multi-modal fusion to harness the synthetic power of visual and linguistic in formation. Our approach significantly enhances semantic comprehension, resulting in 3D generations that are more attuned to human preferences and offer greater adjustability in multi-modal 3D modeling.*

## 1. Introduction

The rise of diffusion models has revolutionized generative modeling across modalities, with text-to-image synthesis models like GLIDE [1] and DALL·E [2] achieving impressive quality and diversity. In parallel, 3D object generation has seen rapid progress. Methods such as DreamFields [3], DreamFusion [4], and Point-E [5] tackle text-to-3D synthesis using various intermediate representations and optimization strategies.

Among these, Point-E stands out for its efficiency. It generates colored 3D point clouds quickly on a single GPU by employing a two-stage pipeline: first, a text-to-image model generates an image from a prompt; then, an image-to-3D diffusion model synthesizes a point cloud conditioned solely on the generated image. While effective, this two-stage design treats textual and visual information independently—disregarding the original text prompt during the 3D generation stage. As a result, critical semantic nuances encoded in language may be lost, limiting grounding and diversity in generated shapes.

To investigate the reliability of the initial stage in Point-E, we conducted a preliminary experiment using GLIDE to generate images from prompts. The sample demonstrations consistently reflect the intended semantics, confirming GLIDE's effectiveness in capturing and visualizing text-described concepts. While in the second process of image to point cloud diffusion, the generated output is struggled to handling the sematic information and can't produce ideal result solely based on the image. See Appendix8 for demonstrations.

Motivated by this observation, we come out some questions.

- How can we make the point cloud diffusion output more align with the image information from the input?
- From a user respective, can we add the compatibility to modify the generation results based with semantics?

Therefore, we propose **Semantic Enhanced Point-E (SEPE)** [1], an enhanced pipeline that preserves and fuses semantic information from both the original text and the generated image. The propose of this fuse process conditions every step of the point cloud diffusion process, allowing the model to leverage complementary multi-modal signals.

While recent work such as Wu et al. [6] explores sketch-and-text guidance for colored point cloud generation, their approach relies on abstract, ambiguous sketches and limited data scale. In contrast, we propose a model that:

- Uses rendered RGB views instead of hand-drawn sketches, enabling precise geometric and semantic guidance.
- Leverages CLIP [7] embeddings for both text and image, enabling a shared semantic space.
- Employs a simple but effective fusion strategy using an fusion module to produce conditioning vectors.

## 2. Related Work

CLIP [7] is being a powerful tool which learns a shared semantic embedding space between natural language and images using contrastive training. Given a batch of images

[1]The repository of SEPE can be found https://github.com/Greebbie/semantic-enhance-with-point-e

$\{i_k\}$ and text descriptions $\{t_k\}$, CLIP minimizes the following loss:

$$\mathcal{L} = -\sum_k \log \frac{\exp(\langle f_i(i_k), f_t(t_k)\rangle)}{\sum_j \exp(\langle f_i(i_k), f_t(t_j)\rangle)} \tag{1}$$

where $f_i$ and $f_t$ are image and text encoders respectively, and $\langle \cdot, \cdot \rangle$ denotes cosine similarity. Being able to extract and keep the semantics and image information in a same embedding space, it is become a powerful choice to handle the fusion.

In the original Point-E [5], it proposes a fast system for generating colored point clouds using diffusion models.

Similar to a standard diffusion model, the method predicts both the noise and the covariance of the denoised signal, leveraging classifier-free guidance to refine the generation process:

$$\boldsymbol{\epsilon}_{\text{guided}} = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \emptyset) + s(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, y) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \emptyset)) \tag{2}$$

where $s$ is the guidance scale.

There are previous attempts to fuse the text and image embedding information. One of the fusion paradigm of concatenating the CLIP features and follow by a MLP [8] to extract the features in to a into a joint vector [6]:

In diffusion-based model, the final condition vector ($\mathbf{c}$) is passed to the UNet [9] backbone. This approach is inspired by prior work demonstrating that rendered RGB images, as opposed to sketches, provide richer geometric cues, enhancing the model's ability to capture detailed structures—particularly when trained on large-scale datasets for improved generalization.

Prior work in multimodal learning, particularly leveraging cross-attention to maintain semantic consistency between input conditions and generated outputs. Cross-attention, a mechanism that dynamically aligns features across modalities (e.g., text and images), has proven effective in various tasks. For instance, Xu demonstrated its significant impact on image and text matching tasks, where it enhanced the alignment of visual regions with textual descriptions, yielding more coherent results [10]. In our model, we adopt cross-attention within the diffusion framework to ensure that generated outputs reflect both the structural details and semantic intent of the input prompts. This approach not only improves accuracy but also enables the model to handle complex prompts efficiently, distinguishing our work in the context of generative modeling.

To further enhance the parameter efficiency of conditioning layers in our model, we adopt LoRA [11], a low-rank adaptation method that introduces trainable rank-decomposed updates to frozen layers. In our implementation, the final MLP used for multimodal fusion employs LoRALinear layers, which reduce memory and computation costs while preserving expressive capacity. [12].

# 3. Method

Our method builds on Point-E [5], aiming to improve semantic grounding by incorporating both image and text features in the 3D point cloud generation process. We describe our dataset, model architecture, and the multimodal conditioning mechanism in detail.

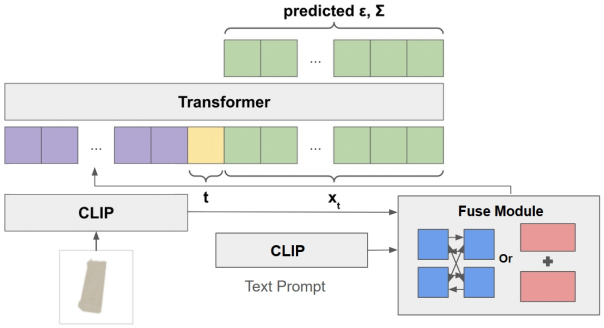



Figure 1. Overview of our Semantic Fusion Point-E pipeline. A text prompt and its corresponding rendered image are embedded using a shared CLIP encoder. These vectors are then fused through a multimodal fusion module—either through concatenation followed by an MLP or through cross-attention—to produce a joint embedding. This representations conditions the following diffusion model that generates colored 3D point clouds.

## 3.1. Fusion Module

To effectively condition the point cloud generation process on both text and image semantics, we propose a fusion module that jointly embeds the text prompt and rendered image into a unified conditioning vector.

Given a text prompt and its corresponding rendered image, we use a pre-trained CLIP encoder [7] to extract their embeddings:

$$\mathbf{t} \in \mathbb{R}^{768} \quad \text{(text)} \tag{3}$$

$$\mathbf{i} \in \mathbb{R}^{256\times1024} \quad \text{(image grid latent)} \tag{4}$$

Here, $\mathbf{t}$ is the global text feature, and $\mathbf{i}$ consists of $N$ image grid features.

We project both text and image features into a shared latent space of dimension $d_f = 512$ through two options.

**Option A: Concat Fusion** In our concat-based fusion, to keep the information in the embedding space consistent and do not lose the spatial information for reconstruction. we extend the text embedding from $[B, 768]$ to $[B, 256, 768]$ and keeping the text information are passed into every patch.

Then they are passed into a MLP [8] to obtain the final condition vector:

$$\mathbf{c} = \text{MLP}(\mathbf{c}_{\text{input}}) \in \mathbb{R}^{512} \tag{5}$$

**Option B: Cross-Attention Fusion** To fuse image and text embeddings using cross-attention [13]. Using image to text as an example, we first apply cross-attention from the projected image tokens $\mathbf{i} \in \mathbb{R}^{256\times1024}$ to the projected text embedding $\mathbf{t} \in \mathbb{R}^{1\times768}$:

$$\mathbf{F}_{\text{img2text}} = \text{CrossAttn}(\mathbf{i}, \mathbf{t}) \in \mathbb{R}^F$$

Instead of averaging $\mathbf{F}_{\text{img2text}}$ across the token dimension, we employ an attention-based pooling mechanism to preserve spatial information. A learnable query vector $\mathbf{q} \in \mathbb{R}^F$ computes attention weights for each token:

$$\alpha_n = \frac{\exp\left(\mathbf{q} \cdot \mathbf{F}^{(n)}_{\text{img2text}}/\sqrt{F}\right)}{\sum_{m=1}^{N} \exp\left(\mathbf{q} \cdot \mathbf{F}^{(m)}_{\text{img2text}}/\sqrt{F}\right)}$$

The fused context vector $\mathbf{c}$ is then obtained as:

$$\mathbf{c} = \sum_{n=1}^{N} \alpha_n \mathbf{F}^{(n)}_{\text{img2text}} \in \mathbb{R}^F$$

Finally, $\mathbf{c}$ is processed through a linear transformation and layer normalization:

This approach ensures that the fused representation $\mathbf{f}$ captures spatially-aware features from the image tokens, weighted by their relevance to the task.

### 3.2. Point Cloud Generation

We adopt a two-stage diffusion framework similar to Point-E [5], where each point cloud is represented as a tensor of shape $K \times 6$ (XYZ coordinates and RGB color), and all values are normalized.

**Stage 1: Coarse Point Cloud Diffusion.** A Transformer-based UNet model predicts both noise $\boldsymbol{\epsilon}$ and variance $\boldsymbol{\Sigma}$ conditioned on timestep $t$, the noised point cloud $\mathbf{x}_t$, and the multimodal fused embedding $\mathbf{c}$ [14]. The model starts from Gaussian noise and progressively denoises the sample via:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I) \tag{6}$$

Input points are projected into a feature space of dimension $D$, and timestep embeddings are prepended. The context also includes CLIP-based embeddings projected into the same space. The model does not use positional encoding, thus remains permutation-invariant.

**Stage 2: Point Cloud Upsampling.** To enhance geometry, a smaller UNet is trained to upsample from 1024 to 4096 points. It takes as input both the coarse point cloud and the same fused condition $\mathbf{c}$, generating 3072 new points to augment the initial shape [15]. The coarse points are embedded using a separate projection layer to distinguish them from the points being generated.

## 4. Experiments

We conduct our training based on pre-trained *base40M point-e* model.l, extending it to the Semantic Enhanced Point-E. Below, we detail the data preparation, training setup, backbone adaptation strategies, and model extension process.

### 4.1. Dataset

We use the Cap3D-ABO subset [12]. It includes rendered RGB images, autogenerated captions by BLIP [16] and 3D colored point clouds. 90% of the data around 7000 samples are being used for training, the rests are kept for validation and test purpose.

### 4.2. Training

We modified the input pipeline to accept both image and text data simultaneously. In contrast, the original Point-E model processes either an image or an embedding, depending on the training or sampling stage. We retained the time encoding embedding logic but adjusted it to better suit our needs. Specifically, we added a condition check to ensure that both text and image inputs are present during training and sampling. This ensures that the model learns only from the fused information of both modalities.

To improve adaptability, we unfroze the last two layers of the Transformer backbone, enabling full weight updates. While this increases computational cost, it enhances performance. Experiments with unfreezing additional layers caused instability. As an alternative, we tested LoRA with a rank of 8, keeping the original weights frozen. Both approaches proved effective, but for the final version, we chose to unfreeze the last two layers. We set different learning rates for the fusion methods and a lower learning rate for the unfrozen backbone layers.

Given the limited computational resources and small dataset, a full cross-attention mechanism may not be feasible. Instead, we adopt a one-way attention mechanism, which has proven effective in image and text tasks [17]. This approach is better suited to our current resource constraints while still delivering reasonable performance.

### 4.3. Loss Function

A key factor influencing our design is that the quality of diffusion models scales with dataset size [18]. Given our limited computational resources and data, we've tailored the loss function specifically for this experiment. Our training loss combines diffusion objectives with constraints, where would be a reasonable choices when generation is shift to image while only image to text attention is used, the lambda values are changable and intend to keep low to avoid nega-

tive influence for the core diffusion loss:

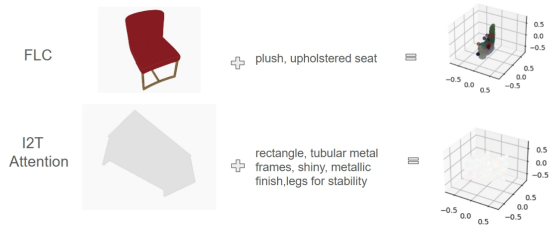
$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{\text{diffusion}} + \lambda_{\text{CD}}\mathcal{L}_{\text{CD}} + \lambda_{\text{color}}\mathcal{L}_{\text{color}} \\ &+ \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{uniform}}\mathcal{L}_{\text{uniform}}\end{aligned} \tag{7}$$

- $\mathcal{L}_{\text{diffusion}}$: Core DDPM [19] loss.
- $\mathcal{L}_{\text{CD}}$: Chamfer Distance [20] between predicted and ground truth 3D point clouds.
- $\mathcal{L}_{\text{color}}$: Extra L2 penalty over RGB values.
- $\mathcal{L}_{\text{smooth}}$: Laplacian loss promoting surface smoothness [21]:

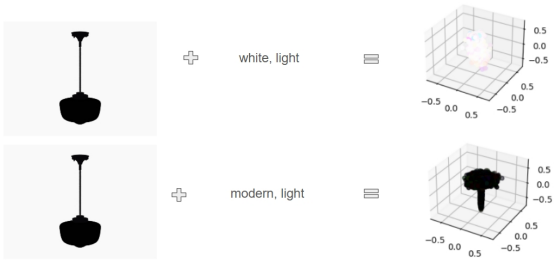
$$\mathcal{L}_{\text{smooth}} = \sum_i \left\| x_i - \frac{1}{|N(i)|} \sum_{j \in N(i)} x_j \right\|^2$$

- $\mathcal{L}_{\text{uniform}}$: Encourages uniform point spacing:

$$\mathcal{L}_{\text{uniform}} = \sum_i \left( \frac{1}{k} \sum_{j \in N_k(i)} (\|x_i - x_j\| - r)^2 \right)$$




(a) Generation results from concatenation and cross-attention of text and image embeddings




(b) Generation results from input image fused with text prompt

## 5. Results

Figure 2a shows that when text and image embeddings are concatenated, the model successfully learns the overall shape, such as a chair, but struggles with other properties. For example, the color red is confined to the middle area, while other regions display mixed or lost color information. By applying image-to-text attention, the model can adjust the generated point cloud's appearance to better align with the text prompt. As illustrated in Figure 2b, the input of a black light can be fused with the text prompt, enabling the model to change its color to white. However, the quality of the generated point clouds remains lower compared to the original.

## 6. Discussion

In our experiment, we chose the Cap3D-ABO subset over larger datasets like ShapeNet [22] for its smaller size, anticipating faster training and easier handling. While its diversity and strong image-text alignment enhance generalization and multimodal learning [23], the limited sample size restricted our diffusion model's ability to learn robust representations, yielding lower generation quality compared to models trained on larger datasets. This highlights a critical trade-off: while smaller datasets reduce training demands, they impair the model's capacity to effectively fuse image and text embeddings. As a result, it remains difficult to determine whether the reduced quality arises from the fusion process disrupting the information in the original embeddings or from the inherent challenges of training diffusion models on limited data. Although our CLIP-based text and image fusion aims to preserve complementary semantic information, it may also introduce conflicting signals, especially when text and image embeddings emphasize different aspects of the object (e.g., material vs. geometry). From related work [6], naive fusion strategies may result in suboptimal alignment between modalities, affecting the quality and consistency of the generated 3D shapes.

## 7. Conclusion

While our method successfully fuses text and image inputs, the generated outputs exhibit lower quality compared to those from single-input modalities. Future directions for improvement include exploring advanced fusion techniques, jointly fine-tuning the CLIP and diffusion backbone to better align features and improve generation quality, and incorporating multi-view consistency constraints during training, as demonstrated in DreamFusion [4], to increase the robustness of geometry generation.

In evaluating our model, we encounter challenges distinct from those in the original Point-E framework, as our approach prioritizes user preference judgment over direct reconstruction. Currently, we lack a robust method for qualitative evaluation. The original Point-E paper employs PointNet-based metrics [24] P-FID and P-IS, to assess reconstruction quality, alongside CLIP R-Precision to measure text alignment. However, since our generation objective does not aim to replicate the ground truth, these metrics are not directly applicable. Moving forward, it is critical to develop a new evaluation metric that assesses the quality of the generated results while also incorporating a semantics match score to ensure alignment with user preferences.

## References


[1] Alex Nichol and Prafulla Dhariwal. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1

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

[3] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Zero-shot text-guided 3d shape generation from a single view. *arXiv preprint arXiv:2112.01455*, 2021. 1

[4] Ben Poole, Ajay Jain, Ben Mildenhall, Matthew Tancik, and Pieter Abbeel. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 4

[5] Alex Nichol et al. Point-e: A system for generating 3d point clouds from complex prompts. *OpenAI*, 2022. 1, 2, 3

[6] Firstname Wu and Anothername Others. Sketch-and-text guided 3d point cloud generation. *arXiv preprint arXiv:2301.12345*, 2023. 1, 2, 4

[7] 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 *International Conference on Machine Learning (ICML)*. PMLR, 2021. 1, 2

[8] Xu Wei, Yunchao Liu, Zheng-Jun Zha, Zechao Wang, and Yanbei Li. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10941–10950, 2020. 2

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015. 2

[10] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. Cross-modal attention with semantic consistence for image–text matching. *IEEE transactions on neural networks and learning systems*, 31(12):5412–5425, 2020. 2

[11] Edward Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[12] Tiange Luo, Justin Johnson, and Honglak Lee. View selection for 3d captioning via diffusion ranking. *arXiv preprint arXiv:2404.07984*, 2024. 2, 3

[13] Jiacheng Li, Yiming Chen, and Bowen Zhang. Cross attention for text and image multimodal data fusion. https://web.stanford.edu/class/cs224n/final-reports/256711050.pdf, 2021. Stanford CS224n Final Project Report. 3

[14] Xiao Zheng, Xiaoshui Huang, Guofeng Mei, Yuenan Hou, Zhaoyang Lyu, Bo Dai, Wanli Ouyang, and Yongshun Gong. Point cloud pre-training with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22935–22945. IEEE, 2024. 3

[15] Wentao Qu, Yuantian Shao, Lingwu Meng, Xiaoshui Huang, and Liang Xiao. A conditional denoising diffusion probabilistic model for point cloud upsampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20786–20795. IEEE, 2024. 3

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 3

[17] Chen He and Haifeng Hu. Image captioning with text-based visual attention. *Neural Processing Letters*, 49:177–185, 2019. 3

[18] Shiyuan Feng, Tongfeng Weng, Xiaolu Chen, Zhuoming Ren, Chang Su, and Chunzi Li. Scaling law of diffusion processes on fractal networks. *Physica A: Statistical Mechanics and its Applications*, 640:129704, 2024. 3

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020. 4

[20] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 605–613, 2017. 4

[21] Shichen Liu, Tianye Li, Weikai Chen, Hao Li, and Yebin Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *IEEE International Conference on Computer Vision (ICCV)*, pages 7708–7717, 2019. 4

[22] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 4

[23] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018. 4

[24] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation, 2017. 4

# Semantic Enhanced Point-E: A Semantics-Driven Advancement on 3D Model Generation

Supplementary Material

## 8. Appendix

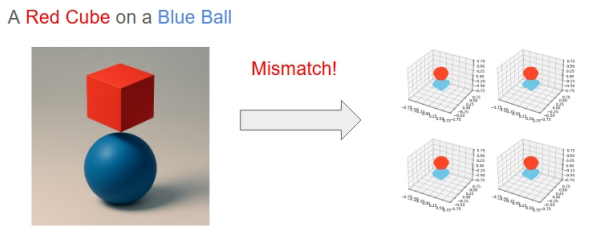



Figure 3. An simple color task during the point cloud diffusion process to generate a "red cube on a blue ball". The point-e mismatched between the blue and red color between the cube and the ball.

(a) Prompt: "a golden toilet"

(b) "a pink jelly-like toilet in club"

(c) "a clear toilet in a fancy hotel"

(d) "a 2050's toilet in the future"

Figure 4. GLIDE-generated images for various text prompts. These results confirm that the text-to-image stage reliably captures visual semantics before conditioning the 3D generation.