GPA4.0: GRADUAL PROMPT-TO-3D ASCENSION FOR(4) VIBE DESIGN

Wei Huang*
lme-hw@berkeley.edu
Student ID: 3040935420

Gangfeng Hu*
tonyhu_ucb@berkeley.edu
Student ID: 3040826233

Chuyue Li*
selene_li@berkeley.edu
Student ID: 3040824218

Kaixing Zhang*
keson@berkeley.edu
Student ID: 3040937812

May 16, 2025

ABSTRACT

The increasing maturity of Large Language Models (LLMs) and diffusion-based models has significantly advanced text-to-3D generation capabilities. However, designing high-fidelity 3D products from textual descriptions remains an underexplored and challenging task. Despite these advances, 3D asset generation remains fundamentally difficult due to cross-modal semantic gaps between language and 3D structures, challenges in achieving controllable spatial precision, high computational demands, and a lack of large-scale paired datasets. To overcome these challenges, we propose a multi-agent architecture for transforming natural language descriptions into 3D digital prototypes, leveraging both layout reasoning and interactive refinement for real-world design workflows. The integration of agent-specific reasoning, interactive multimodal refinement, and structured diffusion modeling ensures both flexibility and fidelity in vibe designing. The design paradigm facilitates practical, human-in-the-loop workflows for product visualization and prototyping, while also providing a reproducible foundation for future extensions in interactive design automation.

1 Introduction

1.1 Motivation: The Need for Efficient MVP Design in 3D Content Creation

The increasing maturity of Large Language Models (LLMs) and diffusion-based models has significantly advanced text-to-image and text-to-3D generation capabilities. However, designing high-fidelity 3D product from textual descriptions remains an underexplored and challenging task. An MVP refers to a minimally functional version of a product, developed with limited resources and time, yet sufficient to validate ideas, gather feedback, and drive iterative improvements. Efficient 3d product generation is particularly crucial for early-stage startups and small companies, where time and budget constraints demand rapid and flexible design cycles.

In practice, small teams often cannot afford large-scale, computationally expensive text-to-3D workflows. Moreover, because a single text-to-3D generation attempt typically fails to fully match user expectations, repeated refinements are necessary. However, editing and localizing 3D structures is substantially harder than in 2D, requiring more sophisticated and accessible workflows to enable effective iterative design.

1.2 Challenges in Text-to-3D Generation

Recent surveys[1, 2, 3] have categorized the evolution of text-to-3D techniques into three main paradigms: feedforward generation, optimization-based generation, and view-reconstruction-based pipelines. Feedforward approaches offer

*Equal contribution by all authors. Authors are listed alphabetically by last name.

speed but often compromise on geometric fidelity[4], while optimization-based methods leveraging Score Distillation Sampling (SDS) achieve higher-quality results at the expense of heavy computational costs[5, 6]. View-reconstruction strategies attempt to generate intermediate 2D views and then reconstruct 3D, but struggle with coherence across views.

Despite these advances, text-to-3D generation remains fundamentally difficult due to cross-modal semantic gaps between language and 3D structures, challenges in achieving controllable spatial precision, high computational demands, and a lack of large-scale paired datasets[7]. Furthermore, existing evaluation metrics are insufficient to fully capture the perceptual quality and geometric accuracy of generated 3D assets[8].

1.3 Our Solution: Progressive Text-to-3D Refinement with GPA4.0

To overcome these challenges, we propose **GPA4.0** (Gradual Text-to-3D Product Ascending for MVP Design), an integrated platform that facilitates progressive refinement through a structured 1D-text $\rightarrow$ 2D-image $\rightarrow$ 3D-model pipeline. Our system leverages Chain-of-Thought (CoT) reasoning[9], LLM-driven prompt engineering[10], and state-of-the-art text-to-2D-to-3D diffusion to enable users to create, inspect, and iteratively refine 3D assets with minimal technical barriers.

Unlike traditional one-shot text-to-3D pipelines that often struggle with controlling fine-grained features[2], GPA4.0 introduces an explicit 2D intermediate stage for semantic clarification and localized adjustment, thereby improving final 3D fidelity. In addition, a multi-turn dialogue system powered by LLMs[11] guides users through iterative feedback and correction cycles, ensuring that design intent is progressively captured and reflected.

Our platform draws inspiration from emerging systems like Meshy, which lower the barrier for 3D content creation by simplifying user interaction. However, GPA4.0 extends this idea by embedding structured reasoning and refinement at every stage, targeting the specific needs of MVP-oriented 3D product design. By addressing both technical and usability challenges, GPA4.0 aims to democratize high-quality 3D asset generation for agile product development environments.

2 Related Work

Prompt engineering has emerged as a fundamental strategy for guiding the behavior of generative models. LLM-driven systems such as DiffusionGPT[12] and Prompt Engineering for X3D[10] dynamically adjust prompts to improve controllability and semantic alignment. LayoutLLM-T2I[11] enhances visual grounding by inferring scene layouts from text, while Chain-of-Thought techniques[9, 13] support step-wise semantic reasoning that enables structured, multi-stage generation. In the 2D domain, Blended Diffusion[14] demonstrates region-based image editing with text, preserving realism and enabling localized prompt control—capabilities that are desirable precursors for refining intermediate visual assets in a 3D pipeline.

Diffusion-based models such as Stable Diffusion have enabled high-fidelity **2D generation** from natural language. For **text-to-3D** tasks, early models like DreamFusion were succeeded by GET3D[15], PI3D[4], and BoostDream[5], each incrementally improving the fidelity and efficiency of 3D shape generation. Gaussian Splatting-based models[6] offer a lightweight and differentiable representation suitable for real-time rendering. Magic3D[16] further refines the two-stage paradigm by combining low-resolution NeRF optimization with high-resolution mesh refinement via latent diffusion, aligning well with GPA4.0's progressive 2D-to-3D refinement pipeline.

Recent works extend **LLM control to 3D content** by learning unified text-to-geometry representations. Uni3D-LLM[17] unifies point cloud perception, generation, and editing under one framework. VP-LLM[18] introduces a patchification strategy for 3D volume completion, bridging transformer architectures with volumetric generation. These methods reveal the potential of language models to directly mediate spatial reasoning and editing workflows, providing foundations for GPA4.0's multi-turn 3D refinement interface.

We exclude traditional multi-view geometry methods[7] and pose-estimation extensions such as NeRF–[19] as they lie outside our primary focus on LLM-conditioned, end-to-end generative modeling. Likewise, time-series-based applications such as shapelet networks[20] are orthogonal to our goal of structural visual generation.

In summary, GPA4.0 builds upon advances in prompt engineering, LLM-guided semantic decomposition, and progressive diffusion pipelines to enable scalable and controllable text-to-3D MVP creation. It extends recent trends in modular 2D-to-3D generation, leveraging both layout reasoning and interactive refinement for real-world design workflows.

3 Methodology

We proposed a modular, agent-based architecture for transforming 1D natural language descriptions into 3D digital prototypes, as shown in Figure 1. First, by introducing a 2D bridging mechanism, we decompose the large modality gap from 1D to 3D into two smaller steps: 1D-to-2D and 2D-to-3D generation, which significantly improves cross-modal feature alignment. Second, we utilize multiple large language models and leverage chain-of-thought (CoT) prompting and prompt engineering techniques to optimize prompt design and enhance visual representations. Finally, we build an end-to-end, integrated text-to-3D generation platform that allows users to guide the generation, editing, undoing, product description creation, and business planning assistance of 2D and 3D assets through textual interaction. In the following section, we will provide a detailed introduction to our approach.

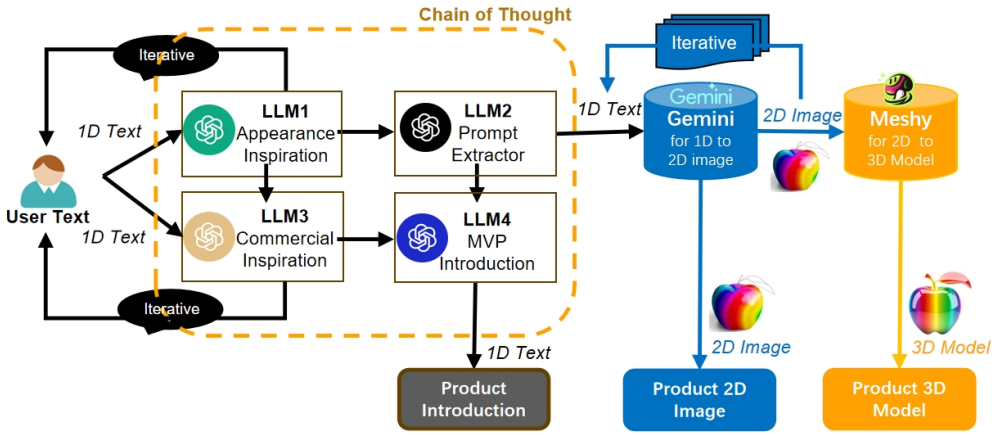

Figure 1: Prompt to 3D Framework for MVP Design:
LLM1 & LLM3: help users inspire their design and commercial thoughts.
LLM2: summarize users' description, and extract key words to pass proper prompt to the text-to-2D agent.
LLM4: based on users' description, write MVP introduction for users.

3.1 2D Bridging for Cross-modal Feature Alignment

Currently, most mainstream text-to-3D generative models on the market employ direct text-to-3D diffusion generation (e.g., Meshy). However, due to the large modality gap, direct diffusion generation struggles to achieve fine-grained alignment of cross-modal features, as demonstrated in our experiment section.

Moreover, current text-to-3D products on the market do not support localized editing. As a result, users are forced to repeatedly generate entirely new models from scratch to match their intended outcomes, which leads to significant time and cost expenditures.

To address these problems, we introduce a 2D bridging mechanism, decomposing the large-span 1D-to-3D diffusion generation into two stages: 1D-to-2D generation using a transformer-based vision model, followed by 2D-to-3D diffusion based on the generated 2D images. In this way, we reduce the modality upscaling gap at each stage, enabling more precise, stable, editable and efficient cross-modal feature alignment and generation.

3.1.1 Model Selection

For the 1D-to-2D generation process, we incorporate both text-to-2D image synthesis and multi-turn dialog-based editing of 2D images. To fully leverage the state-of-the-art multimodal capabilities of current text-to-image models, we select specialized generative models for different tasks. Specifically, we utilize GPT-4o (gpt-image-1) [21] as the generative model and select Gemini (gemini-2.0-flash-preview-image-generation) [22] as the editing and local refinement model, integrating both into our 1D-to-2D framework.

Both models are capable of text-to-image synthesis, but each exhibits unique strengths and weaknesses in specific aspects, making them suitable for different stages of the 2D image generation workflow. GPT-4o excels in tasks requiring photorealism, complex visual alignment, or final presentation-grade images, making it an ideal choice for initial generation and final-stage refinement. In contrast, Gemini is adept at rapid and precise local editing, offering high responsiveness in multi-step interactive workflows. Its ability to maintain spatial and semantic consistency makes it a practical option for progressive image modification. A detailed comparison of the two models is provided in Table 1.

Aspect	GPT-4o	Gemini
Image Quality	High resolution, superior fidelity	Moderate resolution, lower fidelity
Aesthetic Coherence	Strong global visual aesthetics	Acceptable, may lack fine coherence
Edit Consistency	Limited consistency across edits	High consistency in localized edits
Latency	Higher (longer generation time)	Low (fast response)
Cost Efficiency	Higher computational cost	Lower cost

Table 1: Model Selection: Comparison between GPT-4o and Gemini in 1D-to-2D image generation.

3.1.2 Hybrid Generation-Editing Workflow

To maximize generation quality without sacrificing editing flexibility, we propose a hybrid workflow that leverages the complementary strengths of GPT-4o and Gemini.

Step 1: Initial Generation (GPT-4o) :

Given a 1D textual prompt, our hybrid system first employs GPT-4o to generate an initial image. The superior generative performance of this model ensures that the foundational visual structure possesses high fidelity and aesthetic quality.

Step 2: Iterative Refinement (Gemini):

For subsequent edits—including layout adjustments, color correction, local detail modifications, and undo operations—we employ Gemini to fully utilize its superior consistency and low-latency performance. This enables users to rapidly apply changes through simple natural language interactions while preserving the overall integrity of the image.

Step 3: Final Rendering (GPT-4o):

Once editing converges, the intermediate result is passed back to GPT-4o for automatic regeneration at higher resolution and visual precision. This final step repairs any quality loss introduced during editing and produces a 2D output suitable for direct presentation.

Based on this model selection and hybrid generation-editing workflow, we achieve fine-grained alignment, stable performance, and efficient, controllable text-to-2D functionality. The 2D images, which are highly aligned with the design objectives, provide a solid foundation and effective control for subsequent 3D diffusion generation.

3.2 1D Prompt Refinement via CoT

Through our experiments, we found that both 2D and 3D generative models are highly sensitive to the format and granularity of textual prompts; well-crafted and detailed prompts significantly contribute to high-quality generation results.

To support users without design or modeling backgrounds, we leverage the specialized capabilities of multiple large language models and integrate a chain-of-thought (CoT) pipeline to optimize prompt design. Additionally, by adjusting the temperature parameter, we provide functionalities such as inspiration references, design suggestions, commercial summaries, and product description text generation.

3.2.1 Model Selection

Specifically, we leverage the distinct strengths of different large language models (LLMs) to professionally accomplish various language tasks. As illustrated in Figure 1, we integrate four different LLMs.

LLM1 interacts with users through conversational dialogue to optimize the product's appearance design. This component forms the core of "vibe designing." We utilize GPT-4o's rapid response, creativity, and coherence in multi-turn dialogues to help users efficiently refine product appearance goals and clarify design requirements.

LLM2 takes the chat history output from LLM1 as input and is responsible for information extraction and formatted prompt generation. We exploit GPT-4o's capabilities in information summarization, complex context understanding, and multimodal vision-text processing to extract key content from the dialogue and generate prompts in a format suitable for visual generation models. Details regarding prompt construction are provided in Section **??**.

LLM3 engages in dialogue with users to provide auxiliary business design innovation functions. Similar to LLM1, we choose the GPT-4o model for this stage.

LLM4 receives the outputs from LLM2 and LLM3 to summarize and generate product descriptions. We employ GPT-4.1 to leverage its superior contextual understanding, complex reasoning, and professional report generation capabilities.

By integrating different specialized LLMs for distinct tasks, we fully utilize the chain-of-thought (CoT) approach to achieve better prompt generation, thereby delivering a high-quality "vibe designing" user experience and generation results. This enables us to optimize design details, generate structured prompts, and produce product descriptions within a unified, multifunctional framework.

3.3 Integrated 1D to 3D Generation Platform

As shown in Figure 1, we have ultimately developed an end-to-end, fully integrated, multifunctional prompt-to-3D generation platform. The system consists of the following modules:

1. 1D-to-1D Module

This module integrates multiple LLMs to enable iterative conversational interactions for vibe designing. Users interact with GPT models through a unified UI, achieving multi-functional integration under the control of different independent LLMs, including appearance design goal refinement, iterative optimization, prompt extraction and structured generation, as well as business design document generation.

2. 1D-to-2D Module

In this module, the LLM outputs are fed into GPT and Gemini 2D generation models. Users can perform multi-turn natural language interactions to locally edit and modify 2D images, undo modifications, and generate 2D images based on photos or existing images. Through iterative refinement and leveraging the feature alignment capability from 1D to 2D, fine-grained feature-controlled generation is achieved.

3. 2D-to-3D Module

Finally, the images output by the 2D models are input into Meshy for diffusion-based 3D model generation, which is displayed directly on the same platform. Guided by the prompts optimized through 1D iterative refinement and the reference control of 2D images, the generated 3D models achieve effective feature alignment.

Our end-to-end model integrates the above three modal interaction and alignment modules, realizing a unified, multifunctional platform for high-quality text-to-3D generation.

4 Experiments

In this section, we will provide a detailed description of our experimental setup and performance evaluation, including the definition of evaluation metrics, prompt formats, experimental settings, results, and functionality demonstrations.

4.1 Datasets and Evaluation Metrics

The evaluation of generative models in the text-to-3D domain presents unique challenges, primarily due to the scarcity of standardized, established datasets. Meanwhile, unlike other more mature fields in generative AI, publicly available benchmarks for assessing the quality and fidelity and quality of models are also not available. Consequently, to rigorously evaluate our proposed methodology, we have developed a customized dataset (prompts) and a tailored suite of evaluation metrics.

Datasets: In the experiment, our dataset consists of 100 prompts, which are entirely generated by SoTA LLM (GPT4o) with few shots. Each prompt is structured to define a specific item with K distinct features, including but not limited to its color palette, specific patterns or markings, and overall stylistic characteristics, which comprehensively describe various attributes of the targeted item.

PFCR: We evaluate the effectiveness of introducing an intermediate 2D stage by comparing our full pipeline (text-image-3D) with a baseline that maps directly from text to 3D (text-to-3D). For this study, a diverse set of product design prompts was constructed, each consisting of a list of semantic features describing visual and functional aspects.

To quantitatively assess alignment between prompt and final output, we define three metrics: Prompt-Feature Capture Ratio (PFCR), Comparative Capture Ratio (CCR), and Preference Rate (PR). PFCR measures the proportion of semantic features from the prompt that are correctly manifested in the generated 3D model, defined as:

$$\text{PFCR}(m) = \frac{1}{K}\sum_{i=1}^{K} g_{m,i}, \quad g_{m,i} \in \{0,1\}$$

where K is the total number of features in the prompt, and $g_{m,i} = 1$ if feature f_i is correctly realized in model m, and 0 otherwise.

LLM PR: LLM Preference Rate (LLM PR) is a metric quantifies how often an LLM, acting as a judge, selects models from our GPA4.0 method over a text-to-3d baseline in pairwise comparisons. This approach, inspired by the LLM-as-a-judge paradigm[23][24], provides a direct measure of comparative model quality as determined by the LLM.

Human PR: Human Preference Rate (Human PR) is obtained from a human evaluation in which human evaluators are presented with both models' outputs and asked to select the more satisfying result. It is calculated as the proportion of times our model was preferred.

Monetary Cost: Monetary Cost is the total amount spent by the user in producing the final model. For our method (GPA4.0), the cost associated with the 1D text component is not included, as this feature serves as an additional utility for summarization and presentation purposes.

These metrics collectively quantify both objective alignment and subjective quality.

Figure 2: Final Comparisons

4.2 Quantitative Comparisons

As shown in Table 2, our method achieves significantly higher PFCR (0.90 vs. 0.50) and PR (0.79 vs. 0.21), demonstrating that the intermediate 2D stage improves textual alignment and user subjective satisfaction. These results highlight the benefit of decomposing the 1D-to-3D generation into manageable and interpretable subproblems.

Method	PFCR↑	Human PR↑	LLM PR↑	Monetary Cost↓
Baseline (Text-to-3D)	0.5	0.21	0.18	6.4
Ours (GPA4.0)	**0.9**	**0.79**	**0.82**	**2.43**

Table 2: Comparison of Prompt Feature Capture Rate (PFCR), Human Preference Rate, LLM Preference Rate and Monetary Cost (dollars/model) between the baseline text-to-3D pipeline and our GPA method.

4.3 Qualitative Comparisons

We also conduct qualitative comparisons 2 to illustrate the advantages of our method (GPA4.0) over direct text-to-3D method. Our GPA4.0 method, particularly through its 2D stage, yields significant improvements in model generation. For phone cases, this stage facilitates models with enhanced practical realism. For backpack, it produces designs with greater aesthetic appeal, demonstrating the 2D guided approach's versatile benefits.

4.4 Key Characteristics

The superior performance of our method is primarily due to the inherent flexibility of the image generation stage, which supports both modification and reshaping 3. This flexibility enables users to fine-tune visual elements and better align the output with the intended features, thereby improving overall quality and prompt fidelity.

Figure 3

5 Limitation

Although our GPA4.0 method is able to generate exquisite 3D MVP models from textual descriptions, it still faces several challenges in rendering textual information into 3D models.

Rendering Textual Information: In spite of the increasing proficiency of SOTA 2D text-to-image models in accurately rendering textual information, our method, which utilizes these 2D images as an intermediate step, still suffers from consistently producing correct textual information in the final 3D models.

Extended processing times: Due to the inherent sequential pipeline of GPA4.0, including text-to-2D and 2D-to-3D generation, the overall processing time is significantly longer than that of one-shot text-to-3D generation. The cumulative duration of entire pipeline may lead to relatively poor user experience. Future work may consider parallelizing the pipeline to reduce processing time.

6 Conclusion

In summary, our system is a structured system for iterative text-to-3D product design. By decomposing the generation process into interpretable 1D, 2D, and 3D stages, our framework mitigates semantic ambiguity, enhances controllability, and reduces user burden.

In our comparative evaluation against a direct Meshy pipeline, GPA4.0 shows encouraging results in both semantic alignment and user preference. We observe higher prompt faithfulness and preference scores in our limited-scale study, suggesting that introducing intermediate stages may help improve clarity and controllability in 3D generation. While these findings are preliminary, they provide support for the potential value of structured guidance in text-to-3D workflows.

References

[1] Jian Liu, Xiaoshui Huang, Tianyu Huang, Lu Chen, Yuenan Hou, Shixiang Tang, Ziwei Liu, Wanli Ouyang, Wangmeng Zuo, Junjun Jiang, et al. A comprehensive survey on 3d content generation. *arXiv preprint arXiv:2402.01166*, 2024.

[2] Chenghao Li, Chaoning Zhang, Joseph Cho, Atish Waghwase, Lik-Hang Lee, Francois Rameau, Yang Yang, Sung-Ho Bae, and Choong Seon Hong. Generative ai meets 3d: A survey on text-to-3d in aigc era. *arXiv preprint arXiv:2305.06131*, 2023.

[3] Chenhan Jiang. A survey on text-to-3d contents generation in the wild. *arXiv preprint arXiv:2405.09431*, 2024.

[4] Ying-Tian Liu, Yuan-Chen Guo, Guan Luo, Heyi Sun, Wei Yin, and Song-Hai Zhang. Pi3d: Efficient text-to-3d generation with pseudo-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19915–19924, 2024.

[5] Yonghao Yu, Shunan Zhu, Huai Qin, and Haorui Li. Boostdream: efficient refining for high-quality text-to-3d generation from multi-view diffusion. *arXiv preprint arXiv:2401.16764*, 2024.

[6] Zilong Chen, Feng Wang, Yikai Wang, and Huaping Liu. Text-to-3d using gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21401–21412, 2024.

[7] Yuandong Niu, Limin Liu, Fuyu Huang, Siyuan Huang, and Shuangyou Chen. Overview of image-based 3d reconstruction technology. *Journal of the European Optical Society-Rapid Publications*, 20(1):18, 2024.

[8] Jia-Mu Sun, Tong Wu, and Lin Gao. Recent advances in implicit representation-based 3d shape generation. *Visual Intelligence*, 2(1):9, 2024.

[9] Jiacheng Ye, Shansan Gong, Liheng Chen, Lin Zheng, Jiahui Gao, Han Shi, Chuan Wu, Xin Jiang, Zhenguo Li, Wei Bi, et al. Diffusion of thoughts: Chain-of-thought reasoning in diffusion language models. *arXiv preprint arXiv:2402.07754*, 2024.

[10] Nicholas Polys, Ayat Mohammed, and Ben Sandbrook. Prompt engineering for x3d object creation with llms. In *Proceedings of the 29th International ACM Conference on 3D Web Technology*, pages 1–7, 2024.

[11] Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. pages 643–654, 2023.

[12] Jie Qin, Jie Wu, Weifeng Chen, Yuxi Ren, Huixia Li, Hefeng Wu, Xuefeng Xiao, Rui Wang, and Shilei Wen. Diffusiongpt: Llm-driven text-to-image generation system. *arXiv preprint arXiv:2401.10061*, 2024.

[13] Yutaro Yamada, Khyathi Chandu, Yuchen Lin, Jack Hessel, Ilker Yildirim, and Yejin Choi. L3go: Language agents with chain-of-3d-thoughts for generating unconventional objects. *arXiv preprint arXiv:2402.09052*, 2024.

[14] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18208–18218, 2022.

[15] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems*, 35:31841–31854, 2022.

[16] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 300–309, 2023.

[17] Dingning Liu, Xiaoshui Huang, Yuenan Hou, Zhihui Wang, Zhenfei Yin, Yongshun Gong, Peng Gao, and Wanli Ouyang. Uni3d-llm: Unifying point cloud perception, generation and editing with large language models. *arXiv preprint arXiv:2402.03327*, 2024.

[18] Jianmeng Liu, Yichen Liu, Yuyao Zhang, Zeyuan Meng, Yu-Wing Tai, and Chi-Keung Tang. Vp-llm: Text-driven 3d volume completion with large language models through patchification. *arXiv preprint arXiv:2406.05543*, 2024.

[19] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. 2021.

[20] Guozhong Li, Byron Choi, Jianliang Xu, Sourav S Bhowmick, Kwok-Pan Chun, and Grace Lai-Hung Wong. Shapenet: A shapelet-neural network approach for multivariate time series classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8375–8383, 2021.

[21] OpenAI. Model - openai api: Gpt image 1. `https://openai.com/index/image-generation-api/`, 2025. Accessed: 2025-04-23.

[22] Google Developers. Create and edit images with gemini 2.0 in preview. `https://developers.googleblog.com/en/generate-images-gemini-2-0-flash-preview/`, 2025. Accessed: 2025-05-07.

[23] Lianmin Zheng, Siyuan Chen, Xiang Ren, Zhuohan Lin, Eric P. Li, Xinyang Song, Zixuan Wang, Zhe Yan, Haotian Zhang, Xuehai Li, Yuhui Li, Hao Li, Wei Wang, Jiayi Zhou, Zhiruo Liu, Yuhui Wang, Tianjun Yu, Zhen Zhang, Zefeng Ma, Yiming Wang, Yizhou Zhang, and Dawn Song Li. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

[24] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*, 2024.