Michelangelo: Conditional 3D Shape Generation based on Shape-Image-Text Aligned Latent Representation

Zibo Zhao[1,2*] **Wen Liu**[2*] **Xin Chen**[2] **Xianfang Zeng**[2]
Rui Wang[2] **Pei Cheng**[2] **Bin Fu**[2] **Tao Chen**[3]
Gang Yu[2] **Shenghua Gao**[1,4,5†]
[1]ShanghaiTech University [2]Tencent PCG, China
[3]School of Information Science and Technology, Fudan University, China
[4]Shanghai Engineering Research Center of Intelligent Vision and Imaging
[5]Shanghai Engineering Research Center of Energy Efficient and Custom AI IC

https://github.com/NeuralCarver/Michelangelo

Figure 1: Visualization of the 3D shape produced by our framework, which splits into triplets with a conditional input on the left, a normal map in the middle, and a triangle mesh on the right. The generated 3D shapes semantically conform to the visual or textural conditional inputs.

Abstract

We present a novel *alignment-before-generation* approach to tackle the challenging task of generating general 3D shapes based on 2D images or texts. Directly learning a conditional generative model from images or texts to 3D shapes is prone to producing inconsistent results with the conditions because 3D shapes have an additional dimension whose distribution significantly differs from that of 2D images and texts. To bridge the domain gap among the three modalities and facilitate multi-modal-conditioned 3D shape generation, we explore representing 3D shapes in a shape-image-text-aligned space. Our framework comprises two models: a Shape-Image-Text-Aligned Variational Auto-Encoder (SITA-VAE) and a conditional Aligned Shape Latent Diffusion Model (ASLDM). The former model encodes the 3D shapes into the shape latent space aligned to the image and text and reconstructs

∗Contributed equally and work done while Zibo Zhao was a Research Intern with Tencent PCG.
†Corresponding author.

the fine-grained 3D neural fields corresponding to given shape embeddings via the transformer-based decoder. The latter model learns a probabilistic mapping function from the image or text space to the latent shape space. Our extensive experiments demonstrate that our proposed approach can generate higher-quality and more diverse 3D shapes that better semantically conform to the visual or textural conditional inputs, validating the effectiveness of the shape-image-text-aligned space for cross-modality 3D shape generation.

1 Introduction

Conditional generative model-based 3D shaping generations, such as GAN [8, 30, 58], VAE [7, 31, 5], Auto-Regressive model [62, 34, 64], and Diffusion-based model [63, 36, 13, 11, 27, 37, 24], have great potential to increase productivity in the asset design of games, AR/VR, film production, the furniture industry, manufacturing, and architecture construction. However, two obstacles limit their ability to produce high-quality and diverse 3D shapes conforming to the conditional inputs: 1) diverse 3D shape topologies are complicated to be processed into a neural network-friendly representation; 2) since generating a high-quality 3D shape from a 2D image or textual description is an ill-pose problem, and also the distribution between the 3D shape space and image or text space is quite different, it is hard to learn a probabilistic mapping function from the image or text to 3D shape.

Recently, the neural fields in terms of occupancy [32, 40], Signed Distance Function (SDF) [38], and radiance field [33] have been driving the 3D shape representation in the computer vision and graphics community because their topology-free data structure, such as global latent [38], regular grid latent [40, 12], and point latent [63, 64], are easier to process for neural networks in an implicit functional manner. Once arrive at a compatible space to represent different topological 3D shapes, in light of the great success of auto-regressive and diffusion-based models in audio [25, 26], image [45, 46, 44, 48, 3], video [57, 53, 18, 6], and 3D human motions [66, 55, 60], a conditional auto-regressive or diffusion-based generative model [13, 63, 64] is learned to sample a 3D shape in latent from an image or text. However, generating a high-quality 3D shape from a 2D image or textual description is an ill-posed problem, and it usually requires more prior information for 3D shapes. In contrast, the distribution of the 3D shape space is significantly different from the 2D image or text space, and directly learning a probabilistic mapping function from the image or text to the 3D shape might reduce the quality, diversity, and semantic consistency with the conditional inputs. Prior research [63, 37] has endeavored to mitigate this concern through a coarse-to-fine approach, whereby the initial step involves generating coarse point clouds as an intermediary representation, followed by the regression of a neural field based on the point cloud.

Unlike the previous 3D shape representation, where the neural fields only characterize the geometric information of each 3D shape and capture the shape distribution by regularizing the shape latent with KL-divergence via VAE [11, 27, 65] or VQ-VAE [34, 64], we investigate a novel 3D shape representation that further brings the semantic information into the neural fields and designs a Shape-Image-Text-Aligned Variational Auto-Encoder (SITA-VAE). Specifically, it uses a perceiver-based transformer [56, 22] to encode each 3D shape into the shape embeddings and utilizes a contrastive learning loss to align the 3D shape embeddings with pre-aligned CLIP [43] image/text feature space from large-scale image-text pairs. After that, a transformer-based neural implicit decoder reconstructs the shape of latent embeddings to a neural occupancy or SDF field with a high-quality 3D shape. With the help of the aligned 3D shape, image, and text space which closes the domain gap between the shape latent space and the image/text space, we propose an Aligned Shape Latent Diffusion Model (ASLDM) with a UNet-like skip connection-based transformer architecture [47, 4] to learn a better probabilistic mapping from the image or text to the aligned shape latent space and thereby generate a higher-quality and more diverse 3D shape with more semantic consistency conforming to the conditional image or text inputs.

To summarize, we explore bringing the semantic information into 3D shape representation via aligning the 3D shape, 2D image, and text into a compatible space. The encoded shape latent embeddings could also be decoded to a neural field that preserves high-quality details of a 3D shape. Based on the powerful aligned 3D shape, image, and text space, we propose an aligned shape latent diffusion model to generate a higher-quality 3D shape with more diversity when given the image or text input. We perform extensive experiments on a standard 3D shape generation benchmark, ShapeNet [10], and

a further collected 3D Cartoon Monster dataset with geometric details to validate the effectiveness of our proposed method. All codes will be publicly available.

2 Related Work

2.1 Neural 3D Shape Representation

Neural Fields have dominated the research of recent 3D shape representation, which predict the occupancy [32, 40], Sign Distance Function (SDF), density [38, 51], or feature vectors [9] of each 3D coordinate in the field via a neural network to preserve the high-fidelity of a specific 3D shape in a topology-free way. However, the vanilla neural field can only model a single 3D shape and cannot be generalized to other shapes. To this end, the researchers usually take additional latent codes, such as a global latent [38], a regular latent grid [40, 12], a set of latent points [63, 64], or latent embeddings [65, 24], which describe a particular shape along with each 3D coordinate to make the neural fields generalizable to other 3D shapes or scenes. Though current neural fields' 3D representation can characterize the low-level shape geometry information and preserve the high-fidelity shape details, bringing the high-level semantic information into the neural fields is still a relatively poorly studied problem. However, semantic neural fields are significant to downstream tasks, such as conditional 3D shape generations and 3D perception [21, 52].

2.2 Conditional 3D Shape Generation

Optimization-based approaches which employ a text-image matching loss function to optimize a 3D representation of the neural radiance field (NeRF). Dreamfields and AvatarCLIP [23, 20] adopt a pre-trained CLIP [43] model to measure the similarity between the rendering image and input text as the matching objective. On the other hand, DreamFusion [41] and Magic3D [28] utilize a powerful pre-trained diffusion-based text-to-image model as the optimization guidance and produce more complex and view-consistent results. However, per-scene optimization-based methods suffer from a low success rate and a long optimization time in hours to generate a high-quality 3D shape. However, they only require a pre-trained CLIP or text-to-image model and do not require any 3D data.

Optimization-free methods are an alternative approach to conditional 3D shape generation that leverages paired texts/3D shapes or images/3D shapes to directly learn a conditional generative model from the text or image to the 3D shape representations. CLIP-Forge [50] employs an invertible normalizing flow model to learn a distribution transformation from the CLIP image/text embedding to the shape embedding. AutoSDF [34], ShapeFormer [62], and 3DILG [64] explore an auto-regressive model to learn a marginal distribution of the 3D shapes conditioned on images or texts and then sample a regular grid latent or irregular point latent shape embeddings from the conditions. In recent years, diffusion-based generative models have achieved tremendous success in text-to-image, video, and human motion generation. Several contemporaneous works, including SDFusion [11], Diffusion-SDF [27, 13], 3D-LDM [36], 3DShape2VecSet [65], and Shap-E [24], propose to learn a probabilistic mapping from the textual or visual inputs to the shape latent embeddings via a diffusion model. Since these approaches learn the prior information of the 3D shape data, they could improve the yield rate of high-quality shape generation. Moreover, there is no long-time optimization process, and the inference time is orders of magnitude faster than the optimization-based approaches. However, directly learning a conditional generative model to sample the 3D shape from the conditions might produce low-quality with less-diverse results due to the significant distribution gap between the shape space and the image/text space.

2.3 Contrastive Learning in 3D

Contrastive Language-Image Pre-training (CLIP) [43] has emerged as a fundamental model in 2D visual recognition tasks and cross-modal image synthesis by building the representation connection between vision and language within an aligned space. Recent works have extended the multi-modal contrastive learning paradigm to 3D. CrossPoint [1] learns the 3D-2D alignment to enhance the 3D point cloud understanding. PointCLIP [68] takes full advantage of the CLIP model pre-trained on large-scale image-text pairs and performs alignment between CLIP-encoded point cloud and 3D category texts to generalize the ability of 3D zero-shot and few-shot classification. ULIP [61] and CLIP-goes-3D [15] further learn a unified and aligned representation of images, texts, and 3D point

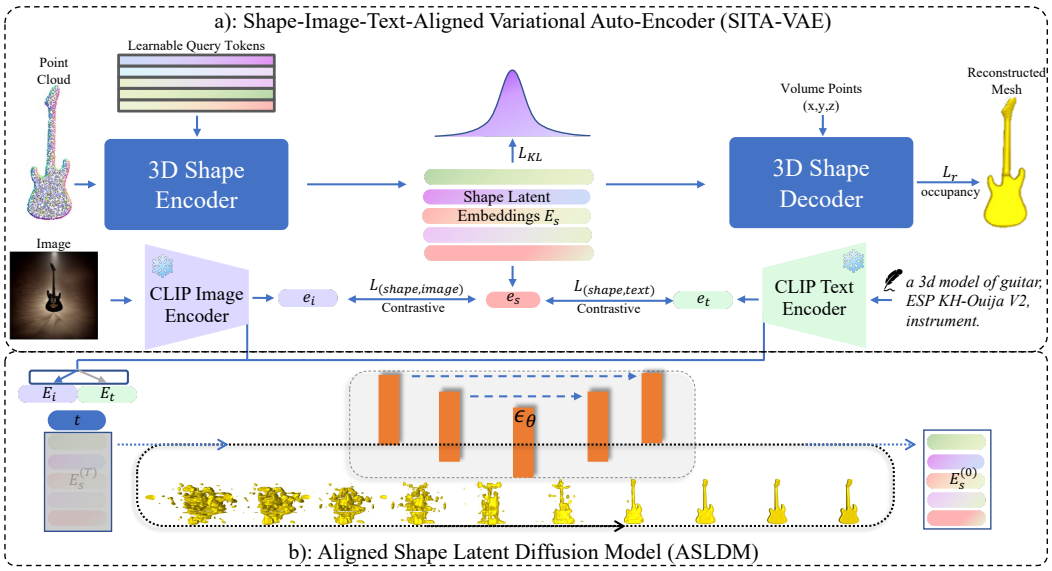


Figure 2: **Alignment-before-generation pipeline**. Our method contains two models: the Shape-Image-Text-Aligned Variational Auto-Encoder (SITA-VAE) and the Aligned Shape Latent Diffusion Model (ASLDM). The SITA-VAE consists of four modules: an image encoder, a text encoder, a 3D shape encoder, and a 3D shape decoder. Encoders encode inputs pair into an aligned space, and the 3D shape decoder reconstructs 3D shapes given embeddings from the aligned space. The ASLDM maps the image or text condition to the aligned shape latent space for sampling a high-quality 3D shape embedding, which latterly reconstructed to high-fidelity 3D shapes by the 3D shape decoder.

clouds by pre-training with object triplets from the three modalities to improve 3D understanding. While most of these works focus on 3D recognition tasks, establishing the connection between 3D recognition and generation tasks remains an under-explored problem.

3 Our Approach

The direct learning of a probabilistic mapping from images or texts to 3D shapes is prone to produce inconsistent results due to the significant distribution gap between the 3D shapes and the 2D images and texts. To address this issue, we propose an alignment-before-generation solution for cross-modal 3D shape generation, as illustrated in Figure 2. Our approach involves two models: the Shape-Image-Text-Aligned Variational Auto-Encoder (SITA-VAE)(Section 3.1) and the Aligned Shape Latent Diffusion Model (ASLDM) (Section 3.2). The former model learns an alignment among the 3D shapes, images, and texts via contrastive learning and then reconstructs the shape embeddings back to the neural field. The latter model is based on the aligned space and is designed to learn a better conditional generative model from the images or texts to shape latent embeddings. By adopting this alignment-before-generation approach, we aim to overcome the challenges posed by the distribution gap and produce more consistent and high-quality results in cross-modal 3D shape generation.

3.1 Shape-Image-Text Aligned Variational Auto-Encoder

Our SITA-VAE contains four components, a pre-trained and fixed CLIP image encoder $\mathcal{E}_i$ and CLIP text encoder $\mathcal{E}_t$, a trainable 3D shape encoder $\mathcal{E}_s$ and neural field decoder $\mathcal{D}_s$. The CLIP image encoder and text encoder take 2D images $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and tokenized texts $\mathbf{T} \in \mathbb{R}^{L_t\times d_t}$ as input, and generate image tokens $\mathbf{E}_i \in \mathbb{R}^{(1+L_i)\times d}$ and text tokens $\mathbf{E}_t \in \mathbb{R}^{L_t\times d}$, where $(1 + L_i)$ and L_t are the sequence length of image tokens $\mathbf{E}_i$ and text tokens $\mathbf{E}_t$. We take advantage of the pre-trained image encoder and text encoder from CLIP. These two encoders are trained on large-scale image-text pairs and robust enough to capture a well-aligned vision-language space, which will enrich the semantics of the 3D shape representation after multi-modal alignment via contrastive learning.

3D shape encoder aims to extract powerful feature representations to effectively characterize each 3D shape. To achieve this, we first sample point clouds $\mathbf{P} \in \mathbb{R}^{N\times(3+C)}$ from the surface of 3D shapes, where N represents the number of points, and C denotes additional point features such as normal or color. Next, we use a linear layer to project the concatenation of the Fourier positional encoded point clouds $\mathbf{P}$ to the 3D shape encoder input $\mathbf{X} \in \mathbb{R}^{N\times d}$. Drawing inspiration from previous transformer-based architectures for point cloud understanding [22], we build our 3D shape encoder on a perceiver-based transformer. Specifically, we use a cross-attention layer to inject the 3D shape information from the input $\mathbf{X}$ into a series of learnable query tokens $\mathbf{Q} \in \mathbb{R}^{(1+L_s)\times d}$, where $1 + L_s$ is the length of query tokens $\mathbf{Q}$, consisting of one global head token $\mathbf{Q}_g \in \mathbb{R}^{1\times d}$ with high-level semantics and L_s local tokens $\mathbf{Q}_l \in \mathbb{R}^{L\times d}$ containing low-level geometric structure information. Then, several self-attention blocks are used to iteratively improve the feature representation and obtain the final shape embeddings, $\mathbf{E}_s \in \mathbb{R}^{(1+L_s)\times d}$.

Alignment among 3D shapes, images, and texts plays a crucial role in SITA-VAE and the conditional generative models. Since the 3D data is the order of magnitudes smaller than the images and texts data, to learn a better-aligned shape among 3D shapes, images, and texts, we enforce the 3D shape encoder close to a pre-aligned vision-language space which is pre-trained on a large-scale image-text pair with rich image and text representations by leveraging the contrastive learning strategy. Consider an input pair of 3D shapes $\mathbf{X}$, images $\mathbf{I}$ and tokenized texts $\mathbf{T}$. The triplet encoders generate the corresponding shape embedding $\mathbf{e}_s$, image embedding $\mathbf{e}_i$ and text-embedding $\mathbf{e}_t$ by projecting the extracted shape tokens $\mathbf{E}_s$, image tokens $\mathbf{E}_i$ and text tokens $\mathbf{E}_t$ as three vectors with the same dimension, which is expressed as: $\mathbf{e}_s = \mathcal{F}_s(\mathbf{E}_s), \mathbf{e}_i = \mathcal{F}_i(\mathbf{E}_i)$, and $\mathbf{e}_t = \mathcal{F}_t(\mathbf{E}_t)$, where $\mathcal{F}_s$ is a learnable shape embedding projector, while image embedding projector $\mathcal{F}_i$ and text embedding projector $\mathcal{F}_t$ are pre-trained and frozen during training and inference. The contrastive loss is:

$$\begin{aligned}
\mathcal{L}_{(shape,image)} &= -\frac{1}{2}\sum_{(j,k)}(\log\frac{\exp(\mathbf{e}_s^j\mathbf{e}_i^k)}{\sum_l \exp(\mathbf{e}_s^j\mathbf{e}_i^l)} + \log\frac{\exp(\mathbf{e}_s^j\mathbf{e}_i^k)}{\sum_l \exp(\mathbf{e}_s^l\mathbf{e}_i^k)}),\\
\mathcal{L}_{(shape,text)} &= -\frac{1}{2}\sum_{(j,k)}(\log\frac{\exp(\mathbf{e}_s^j\mathbf{e}_t^k)}{\sum_l \exp(\mathbf{e}_s^j\mathbf{e}_t^l)} + \log\frac{\exp(\mathbf{e}_s^j\mathbf{e}_t^k)}{\sum_l \exp(\mathbf{e}_s^l\mathbf{e}_t^k)}),
\end{aligned} \tag{1}$$

where (j, k) indicates the positive pair in training batches, and since we utilize pre-trained encoders from CLIP, the model is free from constraint $\mathcal{L}_{(image,text)}$.

3D shape decoder, $\mathcal{D}_s$, takes the shape embeddings $\mathbf{E}_s$ as inputs to reconstruct the 3D neural field in a high quality. We use the KL divergence loss $\mathcal{L}_{KL}$ to facilitate the generative process to maintain the latent space as a continuous distribution. Besides, we leverage a projection layer to compress the latent from dimension d to lower dimensions d_0 for a compact representation. Then, another projection layer is used to transform the sampled latent from dimension d_0 back to high dimension d for reconstructing neural fields of 3D shapes. Like the encoder, our decoder model also builds on a transformer with the cross-attention mechanism. Given a query 3D point $\mathbf{x} \in \mathbb{R}^3$ in the field and its corresponding shape latent embeddings $\mathbf{E}_s$, the decoder computes cross attention iterative for predicting the occupancy of the query point $\mathcal{O}(x)$. The training loss expresses as:

$$\mathcal{L}_r = \mathbb{E}_{x\in\mathbb{R}^3}[\text{BCE}(\mathcal{D}(\mathbf{x}|\mathbf{E_s}), \mathcal{O}(\mathbf{x}))], \tag{2}$$

where BCE is binary cross-entropy loss, and the total loss for training Shape-Image-Text Aligned Variational Auto-Encoder (SITA) is written as:

$$\mathcal{L}_{SITA} = \lambda_c(\mathcal{L}_{(shape,image)} + \mathcal{L}_{(shape,text)}) + \mathcal{L}_r + \lambda_{KL}\mathcal{L}_{KL}. \tag{3}$$

3.2 Aligned Shape Latent Diffusion Model

After training the SITA-VAE, we obtain an alignment space among 3D shapes, images, and texts, as well as a 3D shape encoder and decoder that compress the 3D shape into low-dimensional shape latent embeddings and reconstruct shape latent embeddings to a neural field with high quality. Building on the success of the Latent Diffusion Model (LDM) [46] in the text-to-image generation, which strikes a balance between computational overhead and generation quality, we propose a shape latent diffusion model on the aligned space to learn a better probabilistic mapping from 2D images or texts to 3D shape latent embeddings. By leveraging the alignment space and the shape latent diffusion model, we can generate high-quality 3D shapes that better conform to the visual or textural conditional inputs.

Our Aligned Shape Latent Diffusion Model (ASLDM) builds on a UNet-like transformer [47, 56, 4], aim to fit a distribution of the shape latent embeddings, accompanied by an auto-encoder for encoding data samples into the latent space and reconstructing the data samples given the sampled latent. By learning in the latent space, the latent diffusion model is computationally efficient, and leveraging such a compact representation enables the model to fit the target distribution faster. Specifically, the model ϵ_θ focuses on generating shape latent embeddings $\mathbf{E}_s$ conditioned on $\mathbf{C}$, which is represented by the CLIP image or text encoder. Following LDM [46], the objective is

$$\mathcal{L} = \mathbb{E}_{\mathbf{E}_s, \epsilon\sim\mathcal{N}(0,1), t}[\|\epsilon - \epsilon_\theta(\mathbf{E}_s^{(t)}, \mathbf{C}, t)\|_2^2], \tag{4}$$

where t is uniformaly samppled from $\{1, ..., T\}$ and $\mathbf{E}_s^{(t)}$ is a noisy version of $\mathbf{E}_s^{(0)}$. During inference, sampling a Gaussian noise, the model gradually denoises the signal until reaching $\mathbf{E}_s^{(0)}$. Followed with classifier-free guidance (CFG) [19], we train our conditional latent diffusion model with classifier-free guidance. In the training phase, the condition $\mathbf{C}$ randomly converts to an empty set $\emptyset$ with a fixed probability 10%. Then, we perform the sampling with the linear combination of conditional and unconditional samples:

$$\epsilon_\theta(\mathbf{E}_s^{(t)}, \mathbf{C}, t) = \epsilon_\theta(\mathbf{E}_s^{(t)}, \emptyset, t) + \lambda(\epsilon_\theta(\mathbf{E}_s^{(t)}, \mathbf{C}, t) - \epsilon_\theta(\mathbf{E}_s^{(t)}, \emptyset, t)), \tag{5}$$

where λ is the guidance scale for trading off the sampling fidelity and diversity.

4 Experiments

To validate the effectiveness of our proposed framework, we conducted extensive experiments. In this section, we provide implementation details of our model in Section A. We also describe the data preparation process, including comparisons with baselines and metrics used in our evaluation, in Section 4.2. Of particular importance, we present quantitative comparison results to validate our model's generation ability. Additionally, we provide visual comparison results to illustrate the quality of the generative outputs in Section 4.3. Also, we conduct ablation studies in Section 4.4 to validate the effectiveness of training the generative model in the aligned space, the effectiveness of pre-trained vision-language models (VLM) on the SITA-VAE and the impact of learnable query embeddings.

4.1 Implementations

We implement our Shape-Image-Text-Aligned Variational Auto-Encoder (SITA-VAE) based on perceiver-based transformer architecture [22], where the 3D shape encoder consists of 1 cross-attention block and eight self-attention blocks. At the same time, the neural field decoder has 16 sefl-attention blocks with a final cross-attention block for the implicit neural field modeling. All attention modules are the transformer [56] style with multi-head attention mechanism (with 12 heads and 64 dimensions of each head), Layer Normalization (Pre-Norm) [2], Feed-Forward Network (with 3072 dimensions) [56] and GELU activation [16]. The learnable query embeddings are $\mathbf{E} \in \mathbb{R}^{513\times768}$ with one head-class token for multi-modal contrastive learning and left 512 shape tokens with a linear projection layer to the VAE space $\in \mathbb{R}^{512\times64}$ for the 3D shape reconstruction. Moreover, we employ pre-train encoders in the CLIP (*ViT-L-14*) [43] as our visual encoder and text encoder and freeze them during training and sampling. Besides, our aligned shape latent diffusion model (ASLDM) builds on a UNet-like transformer [47, 56, 4] consisting of 13 self-attention blocks with skip-connection by default. It contains 12 heads with 64 dimensions for each, and 3076 dimensions in the Feed-Forward Network. Both models use an AdamW-based gradient decent optimizer [29] with a 1e-4 learning rate. Our framework is implemented with PyTorch [39], and we both train the SITA-VAE and ASLDM models with 8 Tesla V100 GPUs for around 5 days. We use DDIM sampling scheduler [54] with 50 steps, which generates a high-quality 3D mesh within 10 seconds.

4.2 Datasets and Evaluation Metrics

Dataset. We use a standard benchmark, ShapeNet [10], to evaluate our model, which provides about 50K manufactured meshes in 55 categories. Each mesh has a category tag and corresponding texts, like fine-grained categories or brief descriptions given by the creator. We follow the train/val/test protocol with 3DILG [64]. We further collect 811 Cartoon Monster 3D shapes with detailed structures, with 615 shapes for training, 71 shapes for validation, and 125 for testing, to evaluate the models' ability to generate complex 3D shapes. To prepare the triplet data (3D shape, image, text), we first

augment the provided texts in two ways. First, we string the shape tag and corresponding description in the format "a 3D model of (*shape tag*), in the style of (*description*)" or "a 3D model of (*shape tag*), (*description*)". Then, inspired by ULIP [61], we also leverage multiple templates containing 65 predefined phrases to provide more text information during training. As for the image data, we render each mesh under four camera poses, augmenting and improving the rendering diversity via the depth-condition-based ControlNet [67].

Metrics. We use the Intersection of Union (IoU) to reflect the accuracy of reconstructions. Then, we propose two new metrics for evaluating 3D shape generation methods. The first is a shape-image score (SI-S). We use a 3D shape encoder and image encoder to extract corresponding shape embedding and image embedding and compute the Cosine Similarity of these two modalities. Another is a shape-text score (ST-S), which computes the similarity between the generated 3D shape and the conditional text input in the aligned shape embedding and text embedding space. Both metrics evaluate the similarity between results and their corresponding conditions. Moreover, we use both the pre-trained ULIP [61] and our SITA to compute SI-S and ST-S, in terms of SI-S (ULIP), ST-S (ULIP), SI-S (SITA) and ST-S (SITA), respectively. Besides, we follow the metrics of P-IS and P-FID as introduced in Point-E [37] and use a pre-trained PointNet++ [42] to compute the point cloud analogous Inception Score [49] and FID [17] to evaluate the diversity and quality of the generated 3D shapes.

4.3 Experimental Comparision

Baselines. In the representation stage, we compare our method with Occ [32], ConvOcc [40], IF-Net [12], 3DILG [64], and 3DS2V [65] on reconstruction tasks to valid the ability of the model to recover a neural field given shape embeddings on the ShapeNet dataset [10]. For the conditional generation stage, we choose the baselines of two recent powerful 3D generation methods, 3DILG and 3DS2V. We first finetune their shape representation module on a mixture dataset of the ShapeNet and the 3D Cartoon Monster. Then we both retrain the text and image conditional generative models of 3DILG and 3DS2V with all the same protocols as ours.

	Overall	Selected	Table	Chair	Airplane	Car	Rifle	Lamp
OccNet [32]	0.825	0.81	0.823	0.803	0.835	0.911	0.755	0.735
ConvOccNet [40]	0.888	0.873	0.847	0.856	0.881	0.921	0.871	0.859
IF-Net [12]	0.934	0.924	0.901	0.927	0.937	0.952	0.914	0.914
3DILG [64]	0.950	0.948	0.963	0.95	0.952	0.961	0.938	0.926
3DS2V [65]	0.955	0.955	**0.965**	0.957	0.962	0.966	0.947	0.931
Ours	**0.966**	**0.964**	**0.965**	**0.966**	**0.966**	**0.969**	**0.967**	**0.95**

Table 1: **Numerical results for reconstruction comparison on IoU(↑, a larger value is better)**. The results show that our model has the best performance in 55 overall categories. The results of selected categories further prove that our model could reconstruct each category faithfully.

	Image-Conditioned				Text-Conditioned			
	SI-S (ULIP)↑	SI-S (SITA)↑	P-FID↓	P-IS↑	ST-S (ULIP)↑	ST-S (SITA)↑	P-FID↓	P-IS↑
3DILG	9.134	11.703	4.592	12.247	10.293	6.878	10.283	12.921
3DS2V	13.289	15.156	2.921	12.92	12.934	9.833	5.704	13.149
Ours	**13.818**	**15.206**	**1.586**	**13.233**	**16.647**	**13.128**	**2.075**	**13.558**

Table 2: **Numerical results for conditional generation comparison**. The results show that our model achieves the best generative performance. The SI-S and ST-S indicate that our model generates high-fidelity results by well-mapping the condition information to its related 3D shapes. Moreover, P-FID reflects that our model generates the most realistic 3D shapes, and P-IS indicates that the generated samples are diverse. ↑ means a larger value is better, and ↓ otherwise.

Numerical Comparison. We report the numerical results in Table 1 and Table 2. Table 1 shows that our model achieves the best reconstruction performance on 55 overall categories. Results of the selected category further proves that our model could faithfully reconstruct 3D shapes in each of 55 categories. Table 2 reports the numerical results for conditional 3D shape generation. Our model achieves the best on all the SI-S and ST-S, indicating that it could map the information from the image or text to its corresponding 3D shape information for generating high-fidelity results. Moreover, the P-FID proves that our model could produce high-quality shape-tokens for generating realistic

3D shapes, and P-IS indicates the diversity of the samples. Specifically, the four left columns show that our model surpasses the baselines on image-conditioned generation, proving that our model can better map visual information to 3D shapes. The four right columns validate the generative quality of text-conditioned generation. Since natural language, compared to the 2D image, usually provides limited and abstract information, and thus learning a model to map text information to the 3D shape is challenging. However, benefiting from training on the aligned latent space, our model significantly improves text-conditioned generation, shown in the right of Table 2, which reflects that our model well-maps natural language information to 3D shapes and generates diverse and high-quality results.

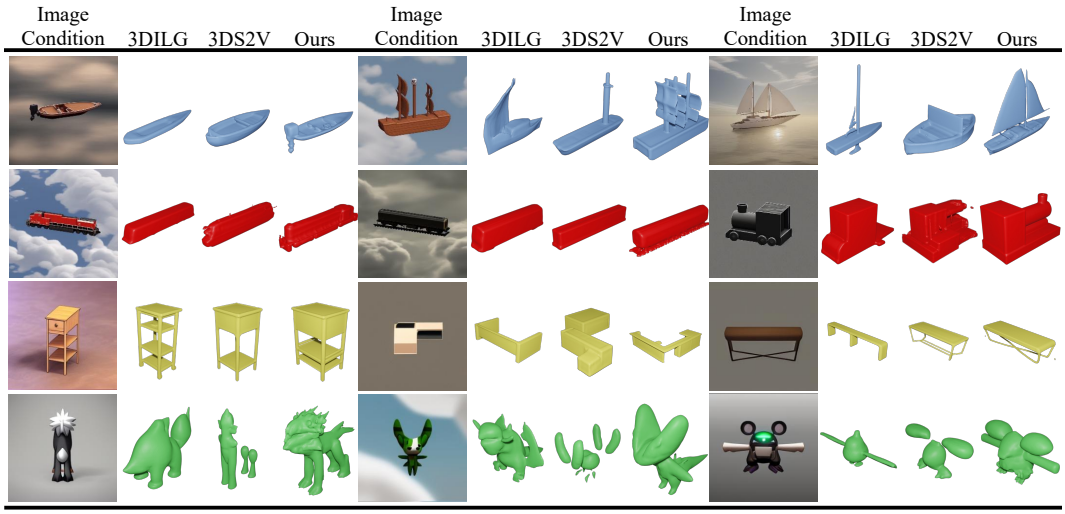
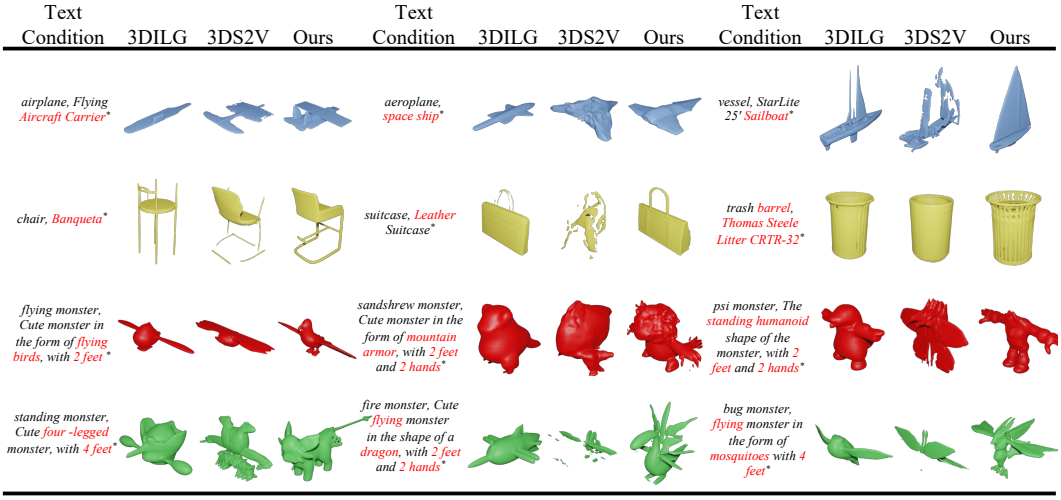


Figure 3: **Visual results for image-conditioned generation comparison**. The figure shows that 3DILG [64] generates over-smooth surfaces and lacks details of shapes, whereas 3DS2V [65] generates few details with noisy and discontinuous surfaces of shapes. In contrast to baselines, our method produces smooth surfaces and portrays shape details. Please zoom in for more visual details.



*denotes "a 3D model of …"

Figure 4: **Visual results for text-conditioned generation comparison**. In the first two rows, we test the model with abstract texts, and the result shows that only our model could generate a 3D shape that conforms to the target text with a smooth surface and fine details. The last two rows show the result given texts containing detailed descriptions, which further shows that our model could capture the global conditional information and the local information for generating high-fidelity 3D shapes. Keywords are highlighted in red; please zoom in for more visual details.

Visual Comparison. The visual comparisons of the image- and text-conditional 3D shape generations illustrates in Figure 8 and Figure 9. Figure 8 shows that 3DILG [64] pays more attention to the global

shape in the auto-regressive generation process, where its results lack depictions of details of 3D shapes. While 3DS2V [65] generates more details of 3D shapes and discontinuous surfaces and noisy results. Besides, both methods are unstable to generate a complete shape when the given conditions maps to a complex object, fine machine, or rare monster. Figure 9 shows the visual comparison of text-conditional generation. In the upper-half rows, we show the results given simple and abstract concepts, while in the lower-half rows, we show the results given detailed texts like descriptions for deterministic parts of the target shape. Similar to the observation above, 3DILG [64] generates an over-smooth shape surface with fewer details, and 3DS2V [65] produces fewer details on the discontinuous object surface. Therefore, only our model produces correct shapes that conform to the given concepts or detailed descriptions with delicate details on smooth surfaces.

4.4 Ablation Studies and Analysis

We ablation study our model from three perspectives, the effectiveness of training generative model in the aligned space, the effectiveness of vision-language models (VLMs) on the SITA-VAE, and the impact of learnable query embeddings.

The effectiveness of training generative model in the aligned space. We perform a visual comparison for ablation study the effectiveness of training the generative model in the aligned space, as illustrated in the Figure 5. The uppers are sampled from the generative model that trains in the aligned space, while the lowers are sampled from the generative model that trains without aligned space. It proves that the uppers conform to the given text and the lower does not, which indicates that training the generative model in the aligned space leads to high-fidelity samples.

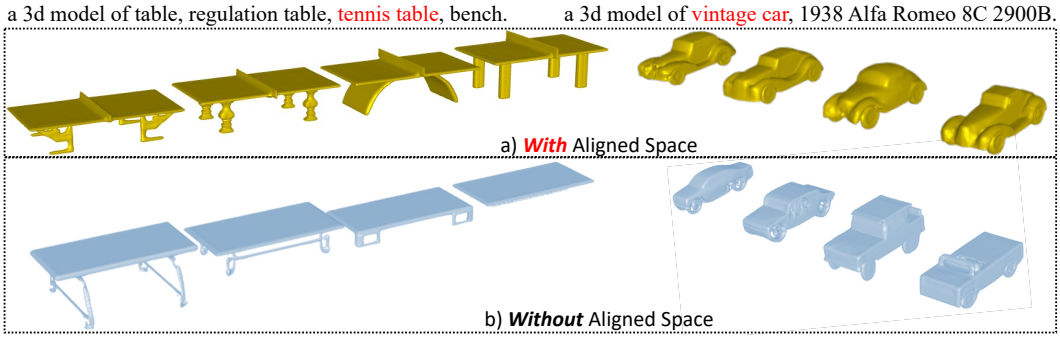


Figure 5: **Ablation study the effectiveness of training generative model in the aligned space**. This figure illustrates visual comparisons for ablation studies on the effectiveness of training the generative model in the aligned space. Compared with the lower samples based on the conditional texts, the upper samples are closer to the conditions semantically, which indicates the effectiveness of the training generative model in the aligned space.

The effectiveness of vision-language models. Except for the well-known vision-language model (VLM) CLIP [43], we introduce another vision-language model (VLM) SLIP [35] for training the SITA-VAE for a comprehensive comparison. First, we evaluate the impact of the vision-language model on SITA-VAE's reconstruction ability, and the results are shown in Figure 6. It shows that our model composed with CLIP achieves the best performance. Then, we evaluate the vision-language model's impact on the ability to align multi-modal space. We select standard and zero-shot classification tasks to reflect the impact of the vision-language model. Note that the classification is performed by a feature matching operation, where we provide multiple 3D shapes and phrases to the SITA-VAE; it returns the similarity between 3D shapes to each phrase as classification results, which indicates that the more the multi-modal space is aligned, leading the higher classification accuracy. The results show that our model composed with CLIP achieves the best performance.

The impact of the learnable query embeddings. We ablation study learnable query embeddings with the same experiments as the above, and the results show that using 512 learnable query embeddings leads to the best performance on reconstructions and classifications.

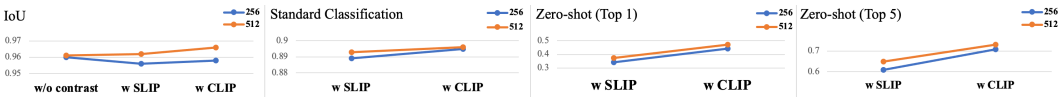


Figure 6: **Ablation study the effectiveness of vision-language models and the impact of learnable query embeddings**. This figure shows the ablation study on the effectiveness of the vision-language model and the impact of learnable query embeddings. According to the table, our model composed with CLIP and 512 learnable query embeddings achieves the best reconstruction and classification performance, indicating its ability to recover 3D shapes and align multi-modal space.

5 Disscusion and Conclusion

Though our method has achieved excellent results, it still has some limitations. First, our method needs the ground truth 3D shapes from training, while 3D data is usually an order of magnitude small than the 2D data. Learning the shape representation with a 3D shape-image-text aligned space from only 2D (multi-view) images via differentiable rendering is a promising direction. Furthermore, since we represent each 3D shape as an occupancy field, it needs to convert the 3D mesh into a watertight one, which will inevitably degrade the original quality of the 3D mesh.

In conclusion, we propose a novel framework for cross-modal 3D shape generation that involves aligning 3D shapes with 2D images and text. We introduce a new 3D shape representation that can reconstruct high-quality 3D shapes from latent embeddings and incorporate semantic information by aligning 3D shapes, 2D images, and text in a compatible space. This aligned space effectively closes the domain gap between the shape latent space and the image/text space, making it easier to learn a better probabilistic mapping from the image or text to the aligned shape latent space. As a result, our proposed method generates higher-quality and more diverse 3D shapes with greater semantic consistency that conform to the conditional image or text inputs.

References

[1] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9902–9912, June 2022.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[4] Fan Bao, Chongxuan Li, Yue Cao, and Jun Zhu. All are worth words: a vit backbone for score-based diffusion models. *arXiv preprint arXiv:2209.12152*, 2022.

[5] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: clothed 3d humans. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 344–359. Springer, 2020.

[6] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. *arXiv preprint arXiv:2304.08818*, 2023.

[7] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Generative and discriminative voxel modeling with convolutional neural networks. *arXiv preprint arXiv:1608.04236*, 2016.

[8] Ruojin Cai, Guandao Yang, Hadar Averbuch-Elor, Zekun Hao, Serge Belongie, Noah Snavely, and Bharath Hariharan. Learning gradient fields for shape generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[9] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.

[10] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

[11] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tuyakov, Alex Schwing, and Liangyan Gui. SDFusion: Multimodal 3d shape completion, reconstruction, and generation. *arXiv*, 2022.

[12] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020.

[13] Gene Chou, Yuval Bahat, and Felix Heide. Diffusionsdf: Conditional generative modeling of signed distance functions. *arXiv preprint arXiv:2211.13757*, 2022.

[14] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[15] Deepti Hegde, Jeya Maria Jose Valanarasu, and Vishal M Patel. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. *arXiv preprint arXiv:2303.11313*, 2023.

[16] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[20] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022.

[21] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[22] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[23] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. 2022.

[24] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

[25] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[26] Yichong Leng, Zehua Chen, Junliang Guo, Haohe Liu, Jiawei Chen, Xu Tan, Danilo Mandic, Lei He, Xiangyang Li, Tao Qin, et al. Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis. *Advances in Neural Information Processing Systems*, 35:23689–23700, 2022.

[27] Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. Diffusion-sdf: Text-to-shape via voxelized diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[28] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[30] Andrew Luo, Tianqin Li, Wen-Hao Zhang, and Tai Sing Lee. Surfgen: Adversarial 3d shape synthesis with explicit surface discriminators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16238–16248, 2021.

[31] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020.

[32] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019.

[33] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[34] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 306–315, 2022.

[35] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 529–544. Springer, 2022.

[36] Gimin Nam, Mariem Khlifi, Andrew Rodriguez, Alberto Tono, Linqi Zhou, and Paul Guerrero. 3d-ldm: Neural implicit 3d shape generation with latent diffusion models. *arXiv preprint arXiv:2212.00842*, 2022.

[37] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

[38] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

[39] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[40] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020.

[41] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022.

[42] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[43] 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*, pages 8748–8763. PMLR, 2021.

[44] 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.

[45] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[49] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[50] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022.

[51] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[52] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.

[53] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[55] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[57] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.

[58] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016.

[59] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[60] Chen Xin, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[61] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning unified representation of language, image and point cloud for 3d understanding. *arXiv preprint arXiv:2212.05171*, 2022.

[62] Xingguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6239–6249, 2022.

[63] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[64] Biao Zhang, Matthias Nießner, and Peter Wonka. 3dilg: Irregular latent grids for 3d generative modeling. In *NeurIPS*, 2022.

[65] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *arXiv preprint arXiv:2301.11445*, 2023.

[66] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[67] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[68] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8542–8552, 2022.

This appendix serves as a supplementary extension, enriching and expanding upon the core content presented in the main body. We first describe the training details of the shape-image-text aligned auto-encoder (SITA-VAE) and aligned shape latent diffusion model (ASLDM) in section A. In section B, we describe more details for the zero-shot classification experiments in Figure 5 in the main text. Furthermore, in section C, we provide the predefined phrases for augmenting the shape-image-text data pair. Benefiting from the alignment among 3D shapes, images, and texts via contrastive learning, our model can retrieve 3D shapes given a query image, and we show the visual result in section D. We also show more visual comparisons in section E. Moreover, we test our model with conditioning input from the internet and show results in section F. Note that HTML files in the zip file accompany all visual results in browsers with interactive 3D viewing.

A Training Details

Stage 1: SITA-VAE. The encoder takes $N = 4096$ point clouds with normal features as the inputs. Equation (3) is the training loss for SITA-VAE. We set λ_c as 0.1 and λ_{KL} as 0.001. For the reconstruction term L_r, we follow the training strategies with 3DILG [64], which first normalize all mesh into $[-1, 1]$, and then separately samples 1024 volumetric points and 1024 near-surface points with ground-truth inside/outside labels from the watertight mesh. The mini-batch size is 40, and we train this model around 200,000 steps.

Stage 2: ASLDM. We the training diffusion scheduler with LDM [46] whose training diffusion steps are 1000, $\beta \in [0.00085, 0.012]$ with scaled linear β scheduler. The mini-batch size is 64, and we train the model around 500,000 steps. In the inference phase, we follow with the classifier-free guidance (CFG) [19] as shown in Equation (5), and we set the guidance scale λ as 7.5.

B Details in zero-shot classification experiments

Dataset. We conduct zero-shot classification experiments on ModelNet40 [59], which provides 12311 synthetic 3D CAD models in 40 categories. The dataset splits into two parts for training and testing, respectively, where the training set contains 9843 models and the testing set contains 2468 models.

Settings. We first train our shape-image-text aligned variational auto-encoder (SITA-VAE) on shapent [10]. Then, we utilize the trained encoders of SITA-VAE for classification on the testing set of ModelNet40 directly. Specifically, for a query 3D shape, we compute the cosine similarity between the shape and each category, where the category reformulates by the phrase "a 3D model of {}". Besides, we report top-1 accuracy and top-5 accuracy, where top-1 accuracy indicates that the ground-truth category achieves the highest similarity, and top-5 accuracy indicates that the ground-truth category achieves similarity in the top 5.

C Template in building shape-image-text data pair

We list the phrase in the predefined template in Table 3. Except for the template introduced in previous work [14, 61], we add one more phrase, "a 3D model of {}" in the template, and while training the model, we replace "{}" with the tag of 3D shapes.

Phrases		
"a 3D model of {}.",	"a point cloud model of {}.",	"There is a {} in the scene.",
"There is the {} in the scene.",	"a photo of a {} in the scene.",	"a photo of the {} in the scene.",
"a photo of one {} in the scene.",	"itap of a {}.",	"itap of my {}.",
"itap of the {}.",	"a photo of a {}.",	"a photo of my {}.",
"a photo of the {}.",	"a photo of one {}.",	"a photo of many {}.",
"a good photo of a {}.",	"a good photo of the {}.",	"a bad photo of a {}.",
"a bad photo of the {}.",	"a photo of a nice {}.",	"a photo of the nice {}.",
"a photo of a cool {}.",	"a photo of the cool {}.",	"a photo of a weird {}.",
"a photo of the weird {}.",	"a photo of a small {}.",	"a photo of the small {}.",
"a photo of a large {}.",	"a photo of the large {}.",	"a photo of a clean {}.",
"a photo of the clean {}.",	"a photo of a dirty {}.",	"a photo of the dirty {}.",
"a bright photo of a {}.",	"a bright photo of the {}.",	"a dark photo of a {}.",
"a dark photo of the {}.",	"a photo of a hard to see {}.",	"a photo of the hard to see {}.",
"a low resolution photo of a {}.",	"a low resolution photo of the {}.",	"a cropped photo of a {}.",
"a cropped photo of the {}.",	"a close-up photo of a {}.",	"a close-up photo of the {}.",
"a jpeg corrupted photo of a {}.",	"a jpeg corrupted photo of the {}.",	"a blurry photo of a {}.",
"a blurry photo of the {}.",	"a pixelated photo of a {}.",	"a pixelated photo of the {}.",
"a black and white photo of the {}.",	"a black and white photo of a {}",	"a plastic {}.",
"the plastic {}.",	"a toy {}.",	"the toy {}.",
"a plushie {}.",	"the plushie {}.",	"a cartoon {}.",
"the cartoon {}.",	"an embroidered {}.",	"the embroidered {}.",
"a painting of the {}.",	"a painting of a {}."	

Table 3: Predefined templates for building shape-image-text pairs. Note that "{}" will be replaced by tags of the 3D shape during training.

D Visualization for image/shape retrieval

Benefiting from the alignment among 3D shapes, images, and texts via contrastive learning, our model can measure the similarity between 3D shapes and images. Therefore, our model could retrieve 3D shapes from the database given a query image. Specifically, given a query image, our model travels through the database and computes the similarity between the image and each 3D shape, where the similarity reflects the visual alignment between the image and the 3D shape. We show visual results in Figure 7, where the golden model is the 3D shape most similar to the query image.

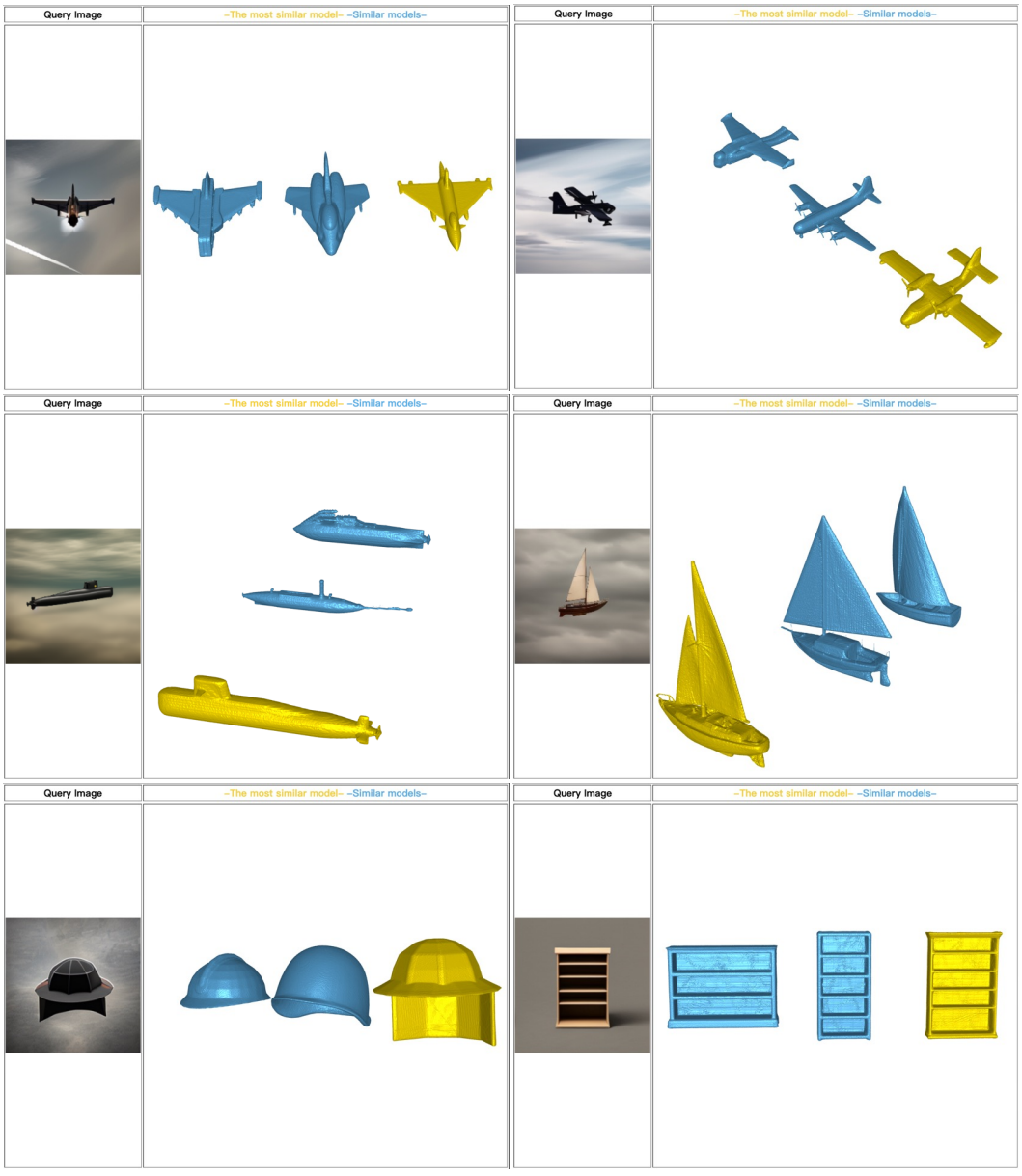


Figure 7: **3D shapes retrieval**. Given a query image, our model could retrieve similar 3D shapes from the database. Results show that the visual information is close, which proves our model could capture 3D shape information aligned with image information. (Please refer to the $'supp\_retrieve/*.html'$ files in the supplementary materials for the interactive 3D viewing visualization.)

E More visual comparison

Image-conditioned generation. We illustrate more image-conditioned 3D shape generation examples in Figure 8. Furthermore, the result proves that our model could capture details in the image and further generate 3D shapes faithfully. Since images only propose single-view information of 3D models, our model could also imagine plausible solutions for generating complete 3D shapes.

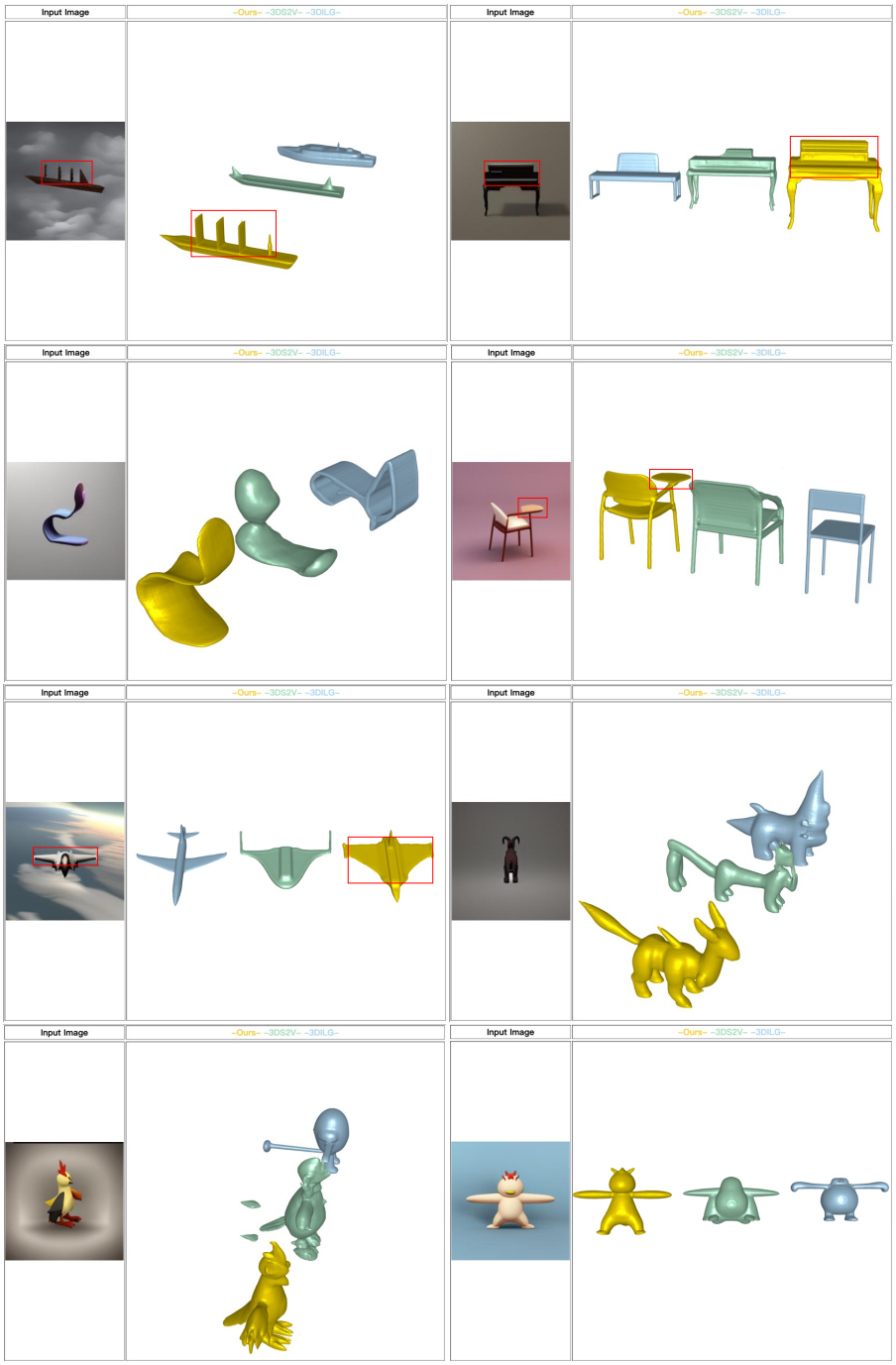


Figure 8: **Image-conditioned generation comparison:** Ours, 3DS2V [65], and 3DILG [64]. (Please refer to the $'supp\_image\_cond/*.html'$ files in the supplementary materials for the interactive 3D viewing visualization.)

Text-conditioned generation. We show more text-conditioned 3D shape generation results in Figure 9. According to the result, our model could understand the language correctly and map the keyword to corresponding parts in 3D shapes. The result further shows that training the model on the shape-image-text aligned space boosts the model's generative ability.

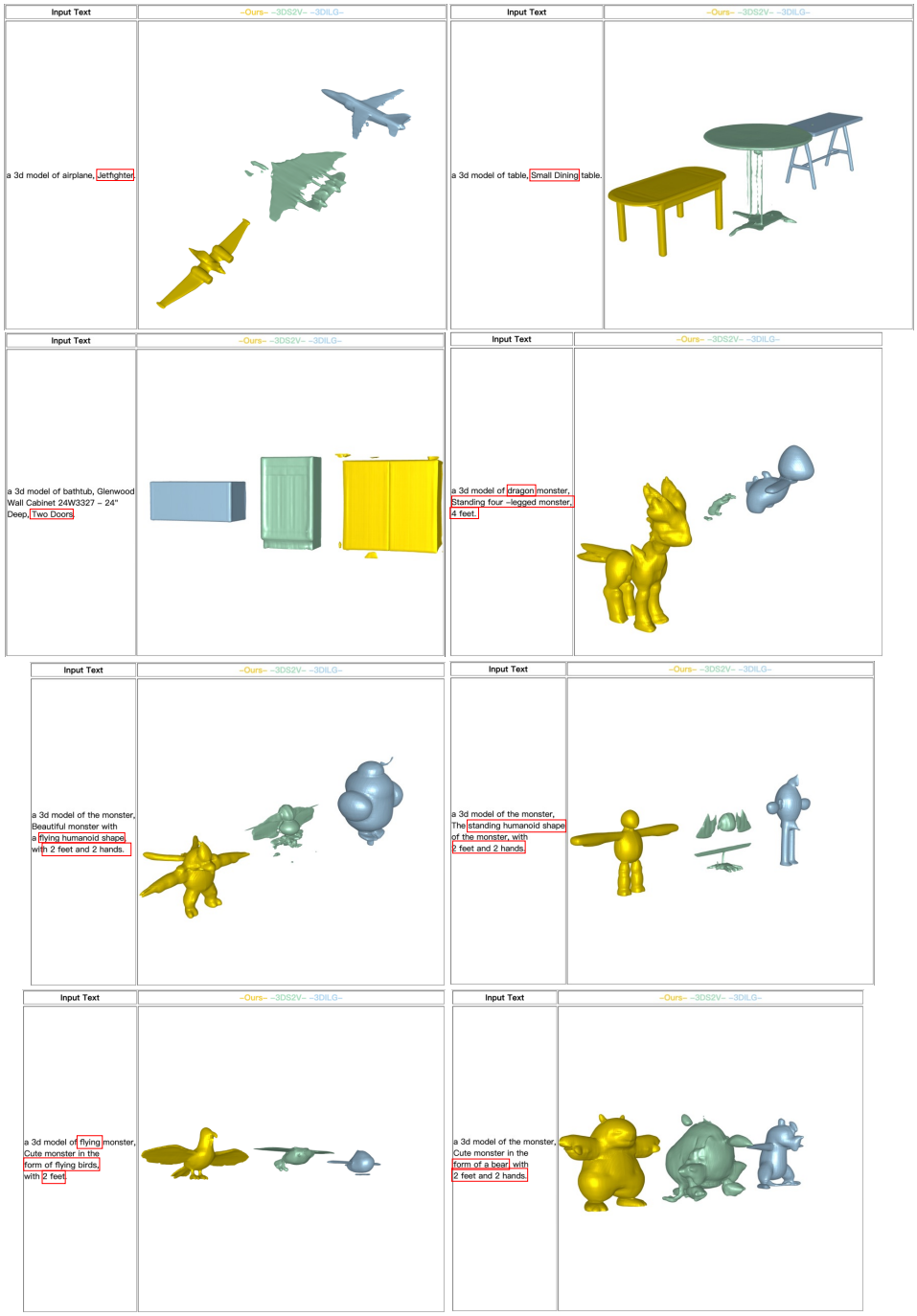


Figure 9: **Text-conditioned generation comparison:** Ours, 3DS2V [65], and 3DILG [64]. (Please refer to the $'supp\_text\_cond/*.html'$ files in the supplementary materials for the interactive 3D viewing visualization.)

F Test in the wild

We also test the model with data in the wild, including images from the internet and manually design text.

Conditional 3D shape generation on images from the Internet. We select some images from the Internet as conditions for the model. Results are shown in Figure 10. According to the generated 3D shapes, the model could map the visual information to 3D shapes, proving that our model could robustly handle some out-of-domain images.

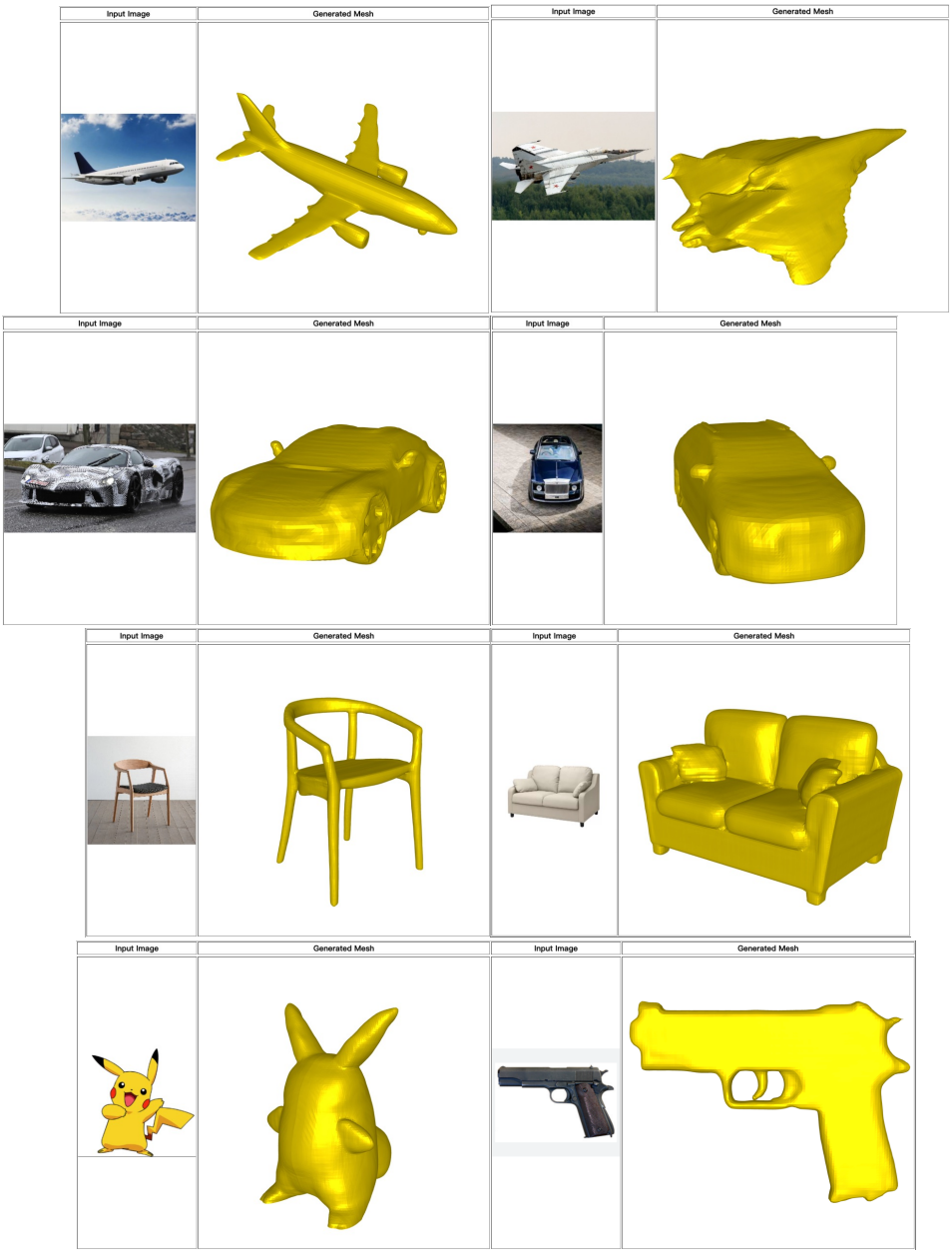


Figure 10: **Conditional 3D shape generation on images from the Internet**. (Please refer to the $'supp\_wild/image/*.html'$ files in the supplementary materials for the interactive 3D viewing visualization.)

Conditional 3D shape generation on manual input text. Moreover, we manually design input texts as conditions for the model, and the results are shown in Figure 11. The generated 3D shapes prove that our model could capture keyword information and produce results that conform to the text.

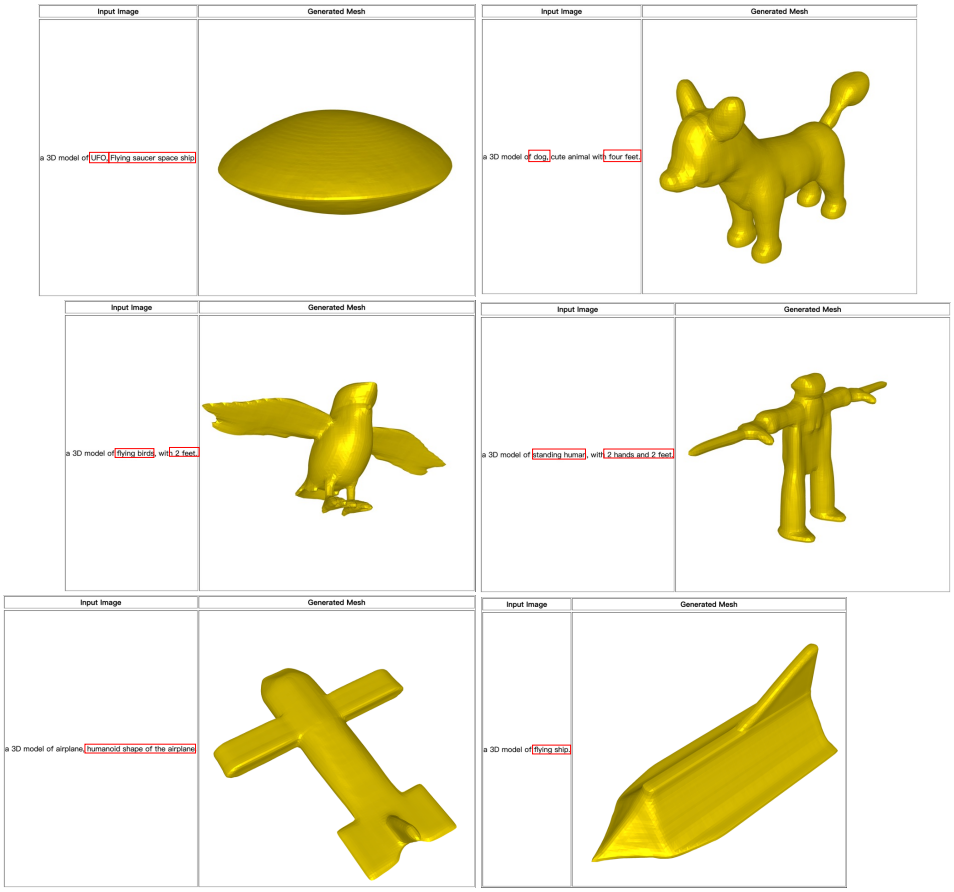


Figure 11: **Conditional 3D shape generation on manually design text**. (Please refer to the $'supp\_wild/text/*.html'$ files in the supplementary materials for the interactive 3D viewing visualization.)