

Supplementary Materials of ImageBind3D

Anonymous Authors

A DETAILS OF OUR NETWORK

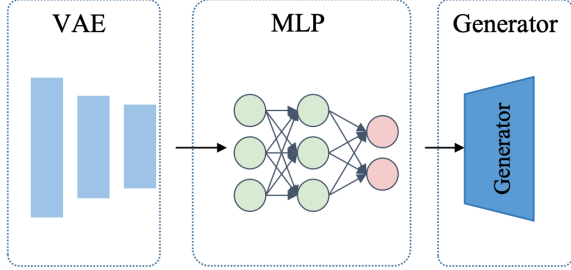


Figure 1: The inversion algorithm employs a three-part architecture consisting of a VAE encoder for extracting image features, an MLP (Multi-Layer Perceptron) network for mapping features to the latent space, and a decoder for reconstructing the 3D shape representation. This design enables the alignment of images and 3D shapes within the latent space.

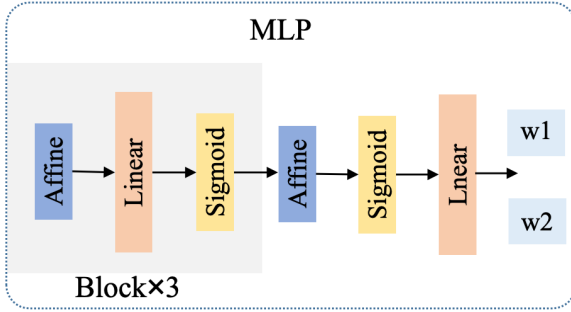


Figure 2: The MLP architecture comprises four stacked blocks, each consisting of an affine module, a linear mapping layer, and an activation layer.

The ImageBind3D framework facilitates controlled 3D generation via a two-tiered approach, supported by guidance from multiple conditions. The initial stage involves an inversion network, the workings of which are detailed first. Subsequently, the focus shifts to the multimodal diffusion model that defines the second stage.

A.1 Inversion network

The core goal of the inversion stage is to achieve alignment between image data and 3D shapes within a shared latent space. As illustrated in Figure 1, our inversion architecture comprises three key components: a VAE encoder, a feature mapper, and a decoder. Deviating from the traditional VAE structure of encoder-decoder pairs, we utilize the encoder to map image data into 512-dimensional features. Subsequently, these features are processed by an MLP structure, yielding latent codes w_1 and w_2 with dimensions of 512×31 . As shown in Figure 2, the MLP design adopts a shallow ResMLP architecture, consisting of four stacked

blocks. Each block incorporates an affine layer, a linear layer, and a Sigmoid activation function. The computation performed by the feature processing unit can be formally defined as follows:

$$w = \text{Linear}(\text{Sigmoid}(\text{Affine}(f))) \quad (1)$$

where f represents the image features extracted by the VAE, while w denotes the latent code. The input latent codes w_1 and w_2 possess dimensions of 512×22 and 512×9 , respectively, representing geometric and appearance codes. Finally, w_1 and w_2 are fed into a pre-trained GET3D model serving as the generator, with its parameters frozen during training.

A.2 multi-modal diffusion model

As shown in Figure 3, our proposed multi-modal diffusion model follows a similar design to Stable Diffusion, incorporating three key components: a denoiser, a conditioning augmentation module, and a decoder. The denoiser employs a U-Net architecture with skip connections to effectively capture multi-scale information. It takes a concatenated vector as input, consisting of the latent codes w_1 and w_2 , along with the feature vector extracted by the VAE encoder. To integrate textual and visual conditioning signals, we introduce a decoupled attention mechanism within the conditioning augmentation module. Visual features are extracted using both CLIP image encoder and VAE encoder and subsequently fused using Adaptive Instance Normalization (AdaIN). It can be defined as follow:

$$\hat{Q}_s = \text{AdaIN}(Q_s, Q_g), \quad (2)$$

$$\hat{K}_s = \text{AdaIN}(K_s, K_g), \quad (3)$$

$$\text{AdaIN}(x, y) = \sigma(y) \left(\frac{x - \mu(x)}{\sigma(x)} \right) + \mu(y), \quad (4)$$

where x, y present CLIP and VAE feature, μ, σ present the mean and standard deviation of features. We concatenate K_n and \hat{K}_m , as well as V_n and V_m , respectively, to obtain K_{nm} and V_{nm} . These fused visual features, together with textual features, undergo separate dot-product attention and cross-attention calculations. The resulting attention maps are then concatenated and injected into the U-Net to guide the denoising process. Finally, we utilize a pre-trained and frozen GET3D generator as the decoder to synthesize 3D shapes from the denoised latent representations.

A.3 Details of experiment setup

Our training process began with the inversion algorithm, aiming to establish a correspondence between image data and 3D shapes. For this purpose, we used synthetic data generated by the GET3D model, and then, stored both the intermediate latent codes and the four rendered views of each generated shape. A 2D discriminator network was then employed to evaluate the quality of each individual view. By aggregating the scores across all four views, we identified and removed low-quality generated samples from the dataset. It costs 15 hours for 3D inversion training. The training of the multi-modal 3D diffusion model was then initiated, focusing

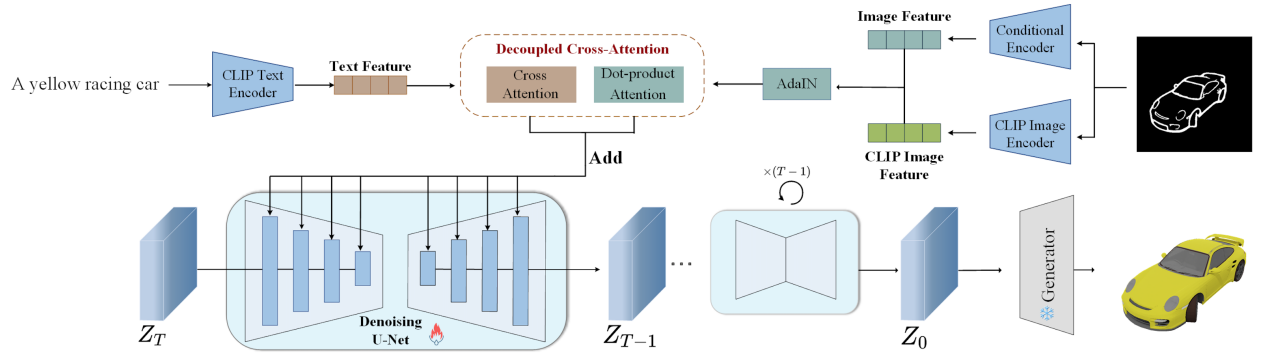


Figure 3: We present a 3D multi-modal diffusion model consisting of a denoiser for latent representation refinement, a decoder for 3D shape synthesis, and a decoupled attention module for incorporating textual and visual conditioning.

first on the text-guided 3D generation task. This stage of training was performed on a system equipped with two NVIDIA 3090 Ti GPUs, utilizing a batch size of 64 and a learning rate of $1e-4$ to optimize the model parameters. Finally, we trained the 3D multi-modal diffusion model using a decoupled attention mechanism. The training sessions lasted 15 and 6 hours for the respective stages.

B EXPERIMENT AND DISCUSSION

In order to provide a comprehensive evaluation of our proposed algorithm's performance, we present an extended analysis of experimental results in this section. Additionally, we conduct ablation studies to assess the contribution of individual components to the overall effectiveness of the model.

B.1 More results

Text-guided 3D generation. To further illustrate the capabilities of our model, we provides a wider range of text-guided 3D generation examples. These examples include multiple viewpoints of the generated shapes, along with visualizations of their mesh structures and hairlines, offering a more detailed understanding of the model's output. Figure 4 showcases a collection of car models generated using text-based descriptions as input. The results exhibit a high degree of fidelity to the provided text prompts, evident in the consistency observed across different viewpoints. Moreover, the generated views maintain coherence with each other, indicating the model's ability to capture the 3D structure of the described objects. We expand the evaluation of our model's generation capabilities in Figure 5 by showcasing a broader spectrum of 3D shapes spanning diverse object classes and stylistic attributes.

The multi-conditional control capabilities. Our multi-conditional control capability is further evidenced by the presented visualization results. In Figure 6, we demonstrate the control over generation achieved through the combination of text and diverse visual conditions. Figure 7 presents 3D generation results controlled by both text and multiple visual conditions, highlighting the model's ability to integrate complex inputs. Experimental results demonstrate that incorporating more visual cues as constraints during 3D generation leads to the creation of 3D objects that better align with desired specifications.

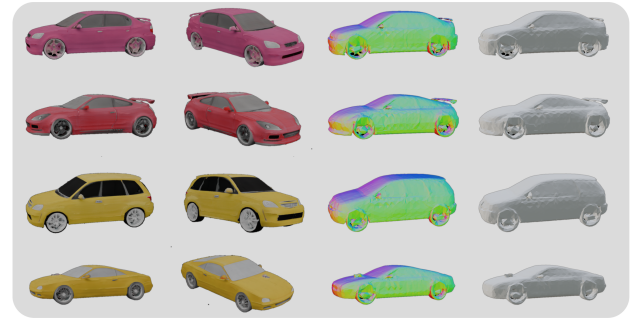


Figure 4: We present the generated 3D objects based on the description "a pink/yellow car". The generation results exhibit both diversity and view consistency.



Figure 5: More diverse results in text-guided generation.

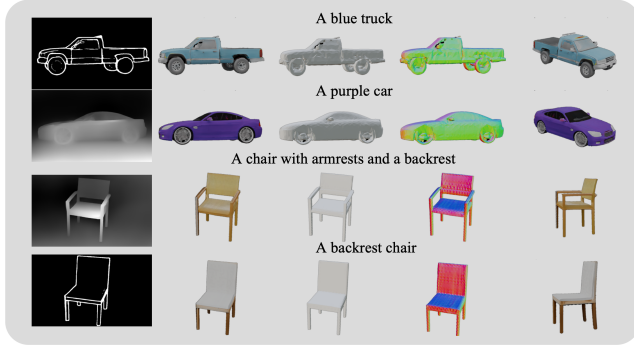


Figure 6: We present 3D generation results achieved using text and diverse visual cues as guiding inputs.

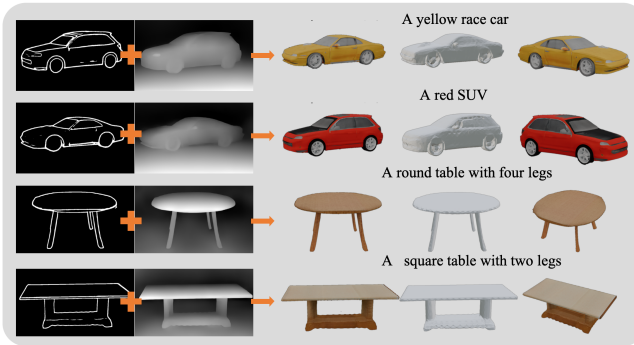


Figure 7: We showcase 3D object generation results achieved through the utilization of multiple conditioning inputs.

B.2 Ablation study

With the aim of validating the efficacy of our proposed ImageBind3D algorithm, we performed three sets of ablation studies. These studies targeted the inversion strategy, the pseudo-labeling module, and the decoupled attention module, respectively. Further analysis showcasing the comparative results of these ablations across various viewpoints and geometries is available in the supplementary materials. **3D Inversion Module.** We conducted ablation-1 by removing the additional L_{clip} and L_2 losses to assess the effectiveness of our inversion strategy. The results presented in Figure 8 demonstrate that our inversion algorithm leads to generation outputs with greater detail.

Pseudo Label Module. For ablation-2, we directly removed the pseudo-label module and utilized a diffusion model to control the image directly for 3D generation. As shown in Figure 8, our pseudo-label module effectively bridges the gap between modalities.

Decoupled Attention Module. In ablation-3 and 4, we compared our decoupled attention module with alternative fusion strategies. The results, presented in Figure 9, indicate that our decoupled attention module enhances information fusion and improves the level of detail in the generated 3D objects.

B.3 Discussion

The field of 3D generation seeks to achieve two primary objectives: generating high-quality results and enabling controllability.

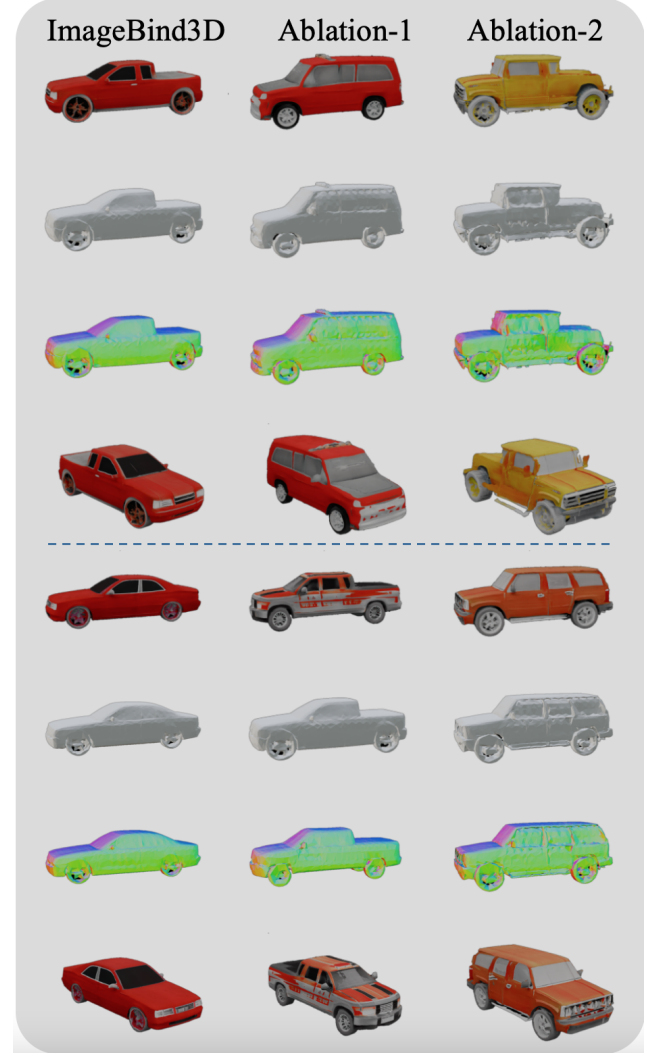


Figure 8: We perform ablation studies in our method, with text prompts "a red car".

While existing methods have demonstrated success in generating high-quality 3D objects, they often suffer from a lack of controllability, hindering users from obtaining desired outcomes based on specific input conditions. Our ImageBind3D addresses this limitation by introducing multi-conditional control capabilities to the 3D generation process, allowing for the combination of diverse input conditions. Experimental results demonstrate that our algorithm achieves both high-quality generation and a high degree of controllability. We acknowledge the current limitation of lacking fine-grained semantic information for detailed local editing, and we identify this as a key direction for future research.

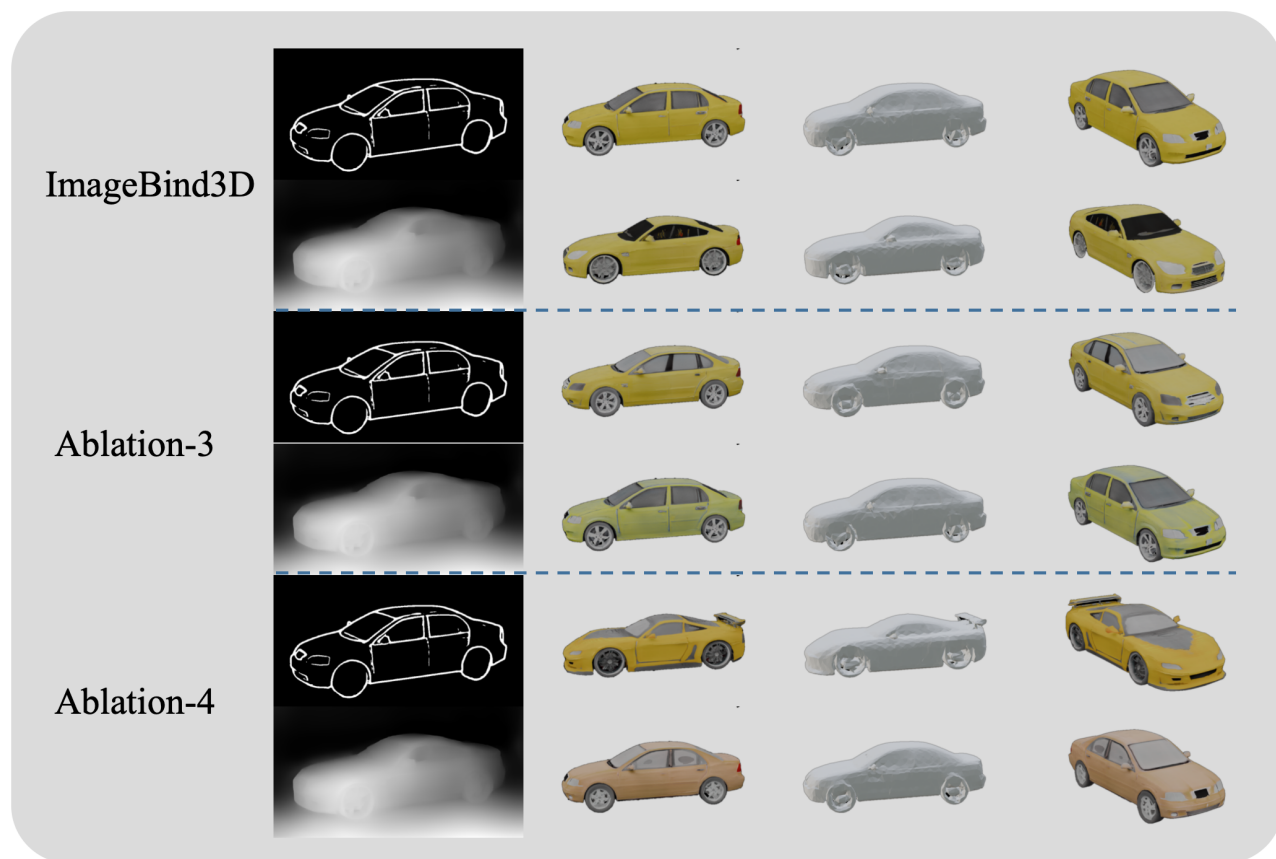


Figure 9: We show 3D results using various fusion mechanisms. All generated outputs are derived from identical text "a yellow car" and visual prompts.