# MultiDreamer3D: Multi-concept 3D Customization with Concept-Aware Diffusion Guidance

**Wooseok Song**[1], **Seunggyu Chang**[2†] and **Jaejun Yoo**[1†]
[1]Ulsan National Institute of Science and Technology (UNIST)
[2]NAVER Cloud
{wooseok.song, jaejun.yoo}@unist.ac.kr, seunggyu.chang@navercorp.com

## Abstract

While single-concept customization has been studied in 3D, multi-concept customization remains largely unexplored. To address this, we propose MultiDreamer3D that can generate coherent multi-concept 3D content in a divide-and-conquer manner. First, we generate 3D bounding boxes using an LLM-based layout controller. Next, a selective point cloud generator creates coarse point clouds for each concept. These point clouds are placed in the 3D bounding boxes and initialized into 3D Gaussian Splatting with concept labels, enabling precise identification of concept attributions in 2D projections. Finally, we refine 3D Gaussians via concept-aware interval score matching, guided by concept-aware diffusion. Our experimental results show that MultiDreamer3D not only ensures object presence and preserves the distinct identities of each concept but also successfully handles complex cases such as property change or interaction. To the best of our knowledge, we are the first to address the multi-concept customization in 3D.

## 1 Introduction

Recent advancements in text-to-3D methods [Poole *et al.*, 2022; Liang *et al.*, 2023] have significantly progressed the generation of 3D models [Mildenhall *et al.*, 2021; Kerbl *et al.*, 2023] from text prompts. The main idea is to optimize 3D models by distilling the score of text-to-image diffusion model [Rombach *et al.*, 2022; Gal *et al.*, 2022] using score distillation sampling (SDS). The SDS enables the generation of both general objects and personalized subjects or concepts, such as "*one's dog*" or "*unique sunglasses*" with personalized diffusion models [Ruiz *et al.*, 2023; Gal *et al.*, 2022]. However, the existing literature predominantly focuses on customizing a single-concept 3D model, thereby constraining its application in more diverse and complex scenarios.

In this study, we tackle multi-concept text-to-3D customization, aiming to produce a 3D model that includes multiple user-defined concepts. For example, consider the 3D model generated from the text prompt: "*A* `C0` *dog is wearing* `C1` *sunglasses*." where `C0` and `C1` represent user-specific concepts such as their "*one's dog*" or "*unique sunglasses*". Achieving high-quality multi-concept 3D models entails overcoming two main challenges: object missing and concept-mixing problems, as illustrated in Figure 1. First, current text-to-3D methods [Poole *et al.*, 2022; Liang *et al.*, 2023] struggle to generate 3D content that accurately represents multiple objects described in a given textual description. This issue arises primarily due to the limitations inherent in text-to-image diffusion models [Rombach *et al.*, 2022; Saharia *et al.*, 2022], which not only face challenges in generating multiple objects in 2D but also often suffer from poor layout context, leading to missing or incorrectly positioned objects. Second, naively adapting multi-concept 2D diffusion model [McMahan *et al.*, 2017; Gu *et al.*, 2024] to optimize 3D model using SDS struggles with the concept-mixing problem, where distinct concept identities are blended or lost. This issue arises from two main factors: the inherent instability of SDS, and the difficulty in managing multiple concepts within a single 2D diffusion model. When these two components are combined, the resulting 3D model often fails to accurately

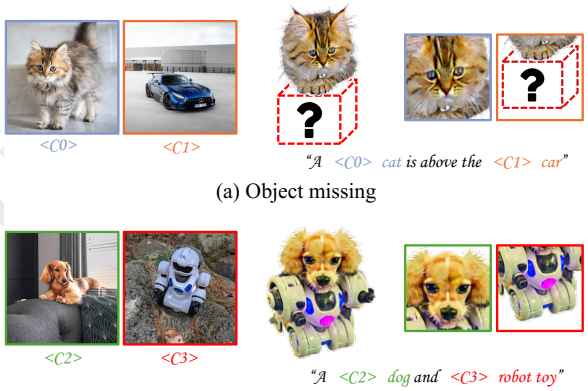



Figure 1: Challenges in multi-concept 3D customization. The 3D content is produced using multi-concept 2D diffusion models using the SDS-based method [Liang *et al.*, 2023]. (a) The "`C1 car`" is missing which leads to poor layout context. (b) The dog's head is combined with a robot toy's body, which we call a concept-mixing problem.

† Corresponding authors.

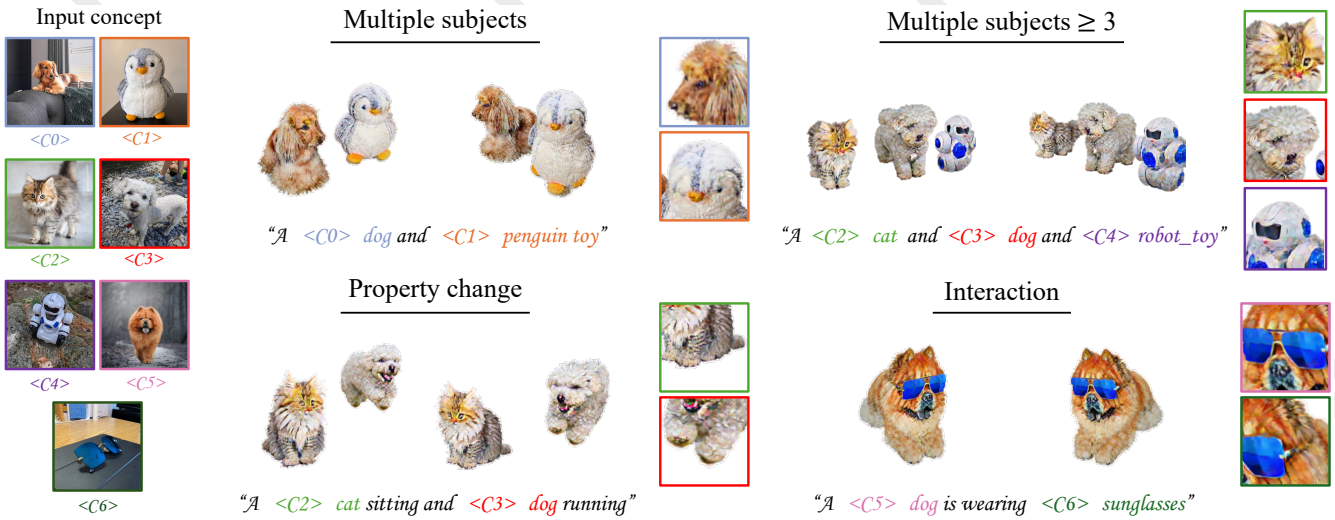



Figure 2: Multi-concept 3D customization with MultiDreamer3D. MultiDreamer3D can generate 3D content incorporating multiple input concepts in three cases: 1) multiple subjects, 2) property change, and 3) interaction.

preserve and distinguish between the multiple user-defined concepts.

To address these challenges, we introduce MultiDreamer3D, a method designed to preserve the individual identities of each concept within a coherent layout context in 3D. The MultiDreamer3D operates in two main stages, utilizing two primary modules: the 3D Layout Generator (LG) and Concept-aware Diffusion Guidance (CDG). In the first stage, LG addresses the object missing by incorporating a large language model (LLM) [Achiam *et al.*, 2023] based 3D layout controller and a selective concept point cloud generator. Specifically, we obtain 3D bounding boxes by querying text prompts to the 3D layout controller, ensuring the presence of objects and coherent layout context. Subsequently, the selective concept point cloud generator generates individual coarse point clouds for each concept, referred to as concept point clouds, and positions them within the 3D bounding boxes. In the second stage, CDG addresses the concept-mixing problem by updating the 3D Gaussian with the concept-aware diffusion score. Specifically, 3D Gaussians are initialized with the concept point clouds and explicit concept labels, and updated with the proposed concept-aware interval score matching (CISM) loss. This approach ensures that each concept maintains its distinct identity without blending or loss during 3D model optimization. As illustrated in Figure 2, our method can generate 3D models with multiple concepts. To demonstrate the effectiveness of MultiDreamer3D, we construct and evaluate three cases of multi-concept 3D content generation: 1) multiple subjects, 2) property change, and 3) interaction. These cases illustrate how MultiDreamer3D effectively maintains the distinct identities of multiple concepts while ensuring object presence and a coherent layout, even in cases involving complex interactions within a 3D space. Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to address multi-concept 3D customization.
- We introduce a 3D Layout Generator (LG) that generates 3D bounding boxes and individual concept point clouds, addressing the object-missing problem.
- We propose Concept-aware Diffusion Guidance (CDG) that updates 3D Gaussians based on concept-aware diffusion score, addressing the concept-mixing problem.
- Our experimental results demonstrate the effectiveness of our method, showcasing its ability to maintain distinct concept identities of multiple concepts within a coherent layout context in 3D.

## 2 Background

### 2.1 3D Gaussian Splatting

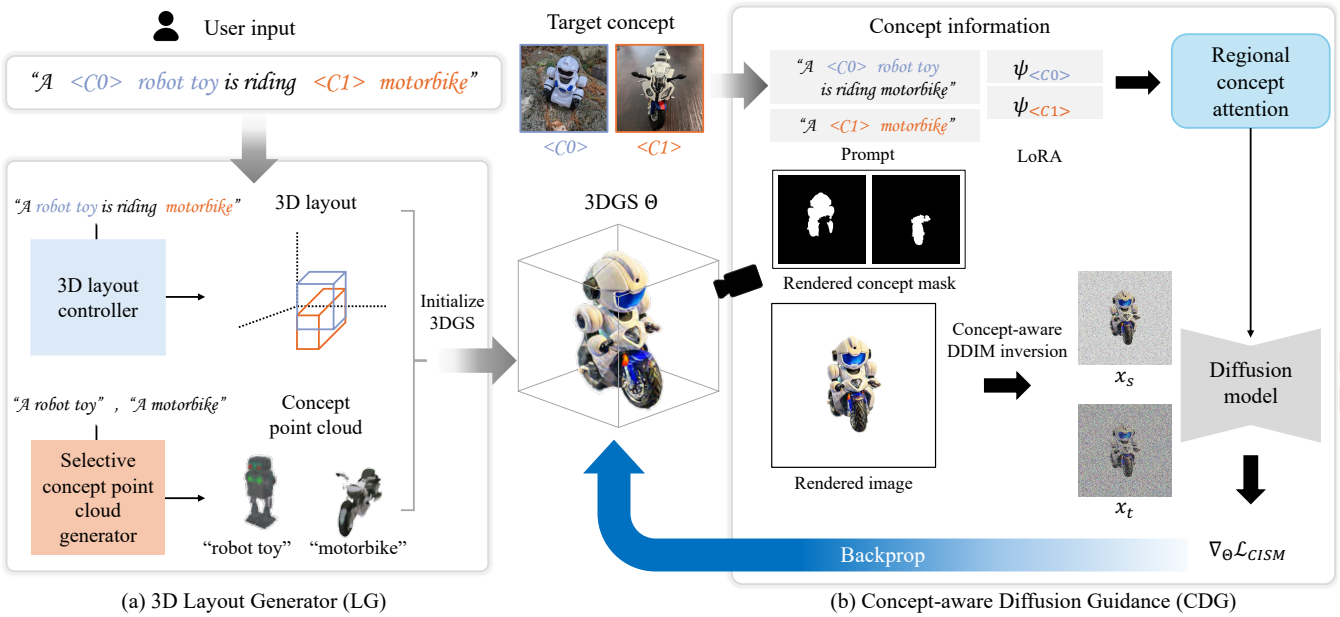
3D Gaussian Splatting (3DGS) [Kerbl *et al.*, 2023] has emerged as a leading explicit 3D representation for novel view synthesis. 3DGS is composed of updatable anisotropic 3D Gaussians denoted as $\Theta = \{\mu, \Sigma, \sigma, \mathbf{c}\}$. Here, $\mu \in \mathbb{R}^3$ represents the position, $\Sigma \in \mathbb{R}^{3\times3}$ is the 3D covariance, $\sigma \in \mathbb{R}$ denotes the opacity, and $\mathbf{c} \in \mathbb{R}^s$ represents the color, where $s$ indicates the degree of spherical harmonics (SH). The 3D Gaussian is formulated as follows:

$$G(\mathbf{x}) = e^{-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}}. \tag{1}$$

3DGS uses a neural point-based rendering technique for pixel color computation, which involves blending $\mathcal{N}$-ordered overlapping points:

$$C = \sum_{i\in\mathcal{N}} \mathbf{c}_i\alpha_i \prod_{j=1}^{i-1}(1-\alpha_j). \tag{2}$$

Here, $\mathbf{c}_i$ refers to the per-point color, and $\alpha_i$ is computed based on the per-point opacity $\sigma_i$ and the 2D projection of the 3D covariance $\Sigma$.




(a) 3D Layout Generator (LG) (b) Concept-aware Diffusion Guidance (CDG)

Figure 3: Overall pipeline of MultiDreamer3D. (a) The 3D layout controller produces 3D bounding boxes given text descriptions. Subsequently, the selective concept point cloud generator outputs coarse concept point clouds and positions within the 3D bounding boxes. (b) The images and concept masks are rendered from 3D Gaussian Splatting (3DGS) Θ and updated with concept-aware interval score matching (CISM) loss, facilitated by regional concept attention (RCA).

### 2.2 Lifting 2D Diffusion Model to 3D

Score distillation sampling (SDS) [Poole *et al.*, 2022] has become a promising method for text-to-3D generation. This technique cleverly adapts the text-to-image diffusion model to optimize 3D models, such as NeRF [Mildenhall *et al.*, 2021] or 3DGS [Kerbl *et al.*, 2023]. Recently, LucidDreamer [Liang *et al.*, 2023] proposed Interval Score Matching (ISM), which aims to improve 3D generation quality by updating Θ with multi-step noise prediction. The process begins with the prediction of noise $\epsilon_\phi(\mathbf{x}_s, \emptyset, s)$ at the diffusion timestep $s = t - \delta_T$. Here, $\delta_T$ indicates the step size of the Denoising Diffusion Implicit Model (DDIM) [Song *et al.*, 2020] inversion, and $\emptyset$ denotes null text prompt. Following this, $\mathbf{x}_t$ is derived through the DDIM inversion process. The gradient of ISM is calculated as follows:

$$\nabla_\Theta \mathcal{L}_{ISM}(\phi, \mathbf{x}) = \mathbb{E}_{t,\epsilon}\Big[w(t)\underbrace{(\epsilon_\phi(\mathbf{x}_t; y, t) - \epsilon_\phi(\mathbf{x}_s; \emptyset, s))}_{\text{ISM update direction}}\frac{\partial \mathbf{x}}{\partial \Theta}\Big]. \quad (3)$$

These methods enable the effective transfer of textual descriptions into precise 3D geometries without 3D supervision.

## 3 Method

The overall pipeline of our method is illustrated in Figure 3. Our method consists of two stages, utilizing two primary modules: 1) 3D Layout Generator (LG) and 2) Concept-aware Diffusion Guidance (CDG). In the first stage, the LG generates 3D bounding boxes with a 3D layout controller to specify individual concept objects, considering the layout context. Subsequently, the LG generates and selects point clouds for each concept, termed concept point clouds, with a selective concept point cloud generator that acquires their coarse geometry. These concept point clouds are then positioned within their respective 3D bounding boxes. In the second stage, we initialize a 3DGS with the concept point clouds and assign concept labels to identify the concepts of each 3D Gaussian. The 3D Gaussians are then updated using CDG, specifically through a concept-aware interval score matching (CISM) loss that incorporates regional concept attention (RCA), designed to preserve the distinct identities of the concepts throughout the process.

### 3.1 3D Layout Generator

**3D Layout Controller.** To produce multi-concept 3D content of high quality, it is essential to ensure both the presence of objects and layout context based on textual descriptions. To address this, we propose a 3D layout controller that leverages Large Language Models [Achiam *et al.*, 2023], which generates 3D bounding boxes for individual concepts based on text prompts. We create examples for three cases (multiple subjects, property change, and interaction) to serve as samples for in-context learning. The 3D layout controller then uses in-context examples with instruction to output the parameter of 3D bounding boxes $Bbox_i = [X_i, Y_i, Z_i, W_i, D_i, H_i]$ for each concept in global coordinate system. Then, we derive scale $s_i$ and translation $t_i$ to position $i$-th concept objects into 3D bounding boxes:

$$s_i = \min(\frac{W_i}{W}, \frac{H_i}{H}), t_i = \Big[X_i + \frac{W_i}{2}, Y_i + \frac{D_i}{2}, Z_i + \frac{H_i}{2}\Big]. \quad (4)$$

$W$ and $H$ denote the maximum width and height, $(X_i, Y_i, Z_i)$ are the coordinates of the lowest left corner, and $(W_i, D_i, H_i)$ represent the width, depth, and height of the bounding box for the $i$-th concept.

**Selective Concept Point Cloud Generator.** The goal of concept point cloud generation is to acquire the coarse geometry of individual concepts. To achieve this, we employ Shap-E [Jun and Nichol, 2023] to generate initial concept point clouds. These prompts can include either simple concept class tokens or brief descriptions, such as "*a dog*" or "*a jumping dog*". Shap-E then generates implicit neural representation (INR) weights corresponding to these text prompts. Following this, vertices of the voxel grid are queried through the INR to obtain colors and signed distance function values, which are subsequently used to construct the concept point clouds. In our method, we manually input text prompt to Shap-E for each concept point clouds.

However, Shap-E often generates point clouds with distorted geometry. To mitigate this issue, we introduce a point cloud selector that utilizes a vision language model (VLM) [Achiam *et al.*, 2023] to ensure reliable 3D geometry. Our selection module begins by generating multiple candidate point clouds from a single text prompt using Shap-E. The point cloud selector then evaluates these candidates by analyzing renderings from fixed viewpoints, selecting the point cloud that best matches the text prompt. The selected point cloud $\mathbf{pcd}_i$ is then positioned within the 3D bounding box in global coordinate system:

$$\mathbf{pcd}_{global} = s_i \times \mathbf{pcd}_i + t_i. \tag{5}$$

Here, $s_i$ and $t_i$ denote scale and translation of $i$-th concept.

### 3.2 Concept-aware Diffusion Guidance

**3DGS Initialization with Concept Labeling.** After concept point clouds are generated and positioned, they are initialized into 3D Gaussians. However, initializing 3D Gaussian without embedding concept information cannot give precise feedback for individual concepts. To address this, we propose concept labeling by incorporating a $k$-class one-hot encoded concept label $\mathbf{m} \in \mathbb{R}^k$ into each 3D Gaussian, represented as $\Theta_i = \{\mu_i, \Sigma_i, \sigma_i, \mathbf{c}_i, \mathbf{m}_i\}$. This setup enables the rendering of a 2D binary concept mask $\mathcal{M} \in \mathcal{R}^{k\times h\times w}$, facilitating precise concept-specific feedback for each Gaussian. The rendering process of concept rendering $M$ follows:

$$M = \sum_{i\in\mathcal{N}} \mathbf{m}_i \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j). \tag{6}$$

Here, $\mathbf{m}$ denotes the concept label. The concept rendering $M_k \in \mathbb{R}^{1\times h\times w}$ represents the contribution of the $k$-th concept to the projected 2D pixel within the range $[0, 1]$. However, this includes low-concept contributions that are noisy. To minimize such noisy contributions, we apply a threshold factor $\tau$ to the concept rendering $M$, producing a binary concept mask $\mathcal{M} \in \mathbb{R}^{k\times h\times w}$.

**Regional Concept Attention.** Updating the concept 3D Gaussians with concept-specific feedback is essential to prevent concept mixing. To achieve this, we introduce the Regional Concept Attention (RCA) module as shown in Figure 4. The RCA modulates the cross-attention map in a text-to-image diffusion model [Rombach *et al.*, 2022] by incorporating individual concept information. This enables unified noise prediction while preserving the distinct identities of each concept. The noise prediction process is as follows.

First, we observe that text prompts containing multiple concepts often lead to concept-mixing problems, as illustrated in Figure 1 (b). To address this issue, we decompose the text prompts into individual concept prompts. For example, when generating a 3D model from a text prompt *"A `C0` robot toy is riding `C1` motorbike"*, we break it down into the following concept prompts:

$$p_0 = \text{"A } \texttt{C0} \text{ robot toy is riding motorbike"},$$
$$p_1 = \text{"A } \texttt{C1} \text{ motorbike"},$$
$$p_{bg} = \text{"A robot toy is riding motorbike"}.$$

The input text prompts can either be decomposed manually by the user or automatically using LLM [Achiam *et al.*, 2023]. (For more details, see Suppl. 2.5.)

Next, we modulate the cross-attention layer with the RCA. The RCA inputs concept masks $\mathcal{M}$, concept LoRAs $\psi$, and concept prompts $p$ and outputs an aggregated concept-specific attention feature. The concept-specific query vector is computed:

$$Q_i = W^q \cdot (\mathcal{M}_i \cdot F), Q_{bg} = W^q \cdot (\mathcal{M}_{bg} \cdot F). \tag{7}$$

Here, $W^q$ denotes the query projection matrix and $F$ denotes the input image feature. $\mathcal{M}_i$ denotes the $i$-th concept mask, while $\mathcal{M}_{bg} = (\mathcal{M}_0 \cup \mathcal{M}_1 ... \cup \mathcal{M}_k)^c$ represents the background mask. This process ensures isolated concept query vectors are used for attention computation. Subsequently, concept-specific keys and values are computed:

$$K_i = (W^k + \lambda \cdot \psi_i^k) \cdot p_i, V_i = (W^v + \lambda \cdot \psi_i^v) \cdot p_i \tag{8}$$

$$K_{bg} = W^k \cdot p_{bg}, V_{bg} = W^v \cdot p_{bg}. \tag{9}$$

Here, $W^k$ and $W^v$ denote the key and value projection matrices. The $\psi_i$ and $p_i$ represent the $i$-th concept LoRA and the concept prompt, while $\lambda$ is the LoRA scale. This ensures that individual concept information is encoded into keys and values. Then, concept-specific attention features are computed:

$$A_i = \text{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{d}}\right) \cdot V_i. \tag{10}$$

Here, $A_i$ denotes the concept-specific attention feature. Finally, we aggregate concept-specific attention features:

$$\hat{A}(\Psi, P, \mathcal{M}) = \mathcal{M}_{bg} \cdot A_{bg} + \sum_{i=1}^{k} \mathcal{M}_i \cdot A_i. \tag{11}$$

Here, $\hat{A}$ represents the aggregated attention feature, and $\Psi$, $P$ denote the set of concept LoRAs and text prompts. The $k$ denotes the number of concepts. The noise prediction with our RCA module is represented as $\epsilon_\phi(\mathbf{x}_t, t, \hat{A}(\Psi, P, \mathcal{M}))$.

**Concept-aware Interval Score Matching.** We introduce concept-aware interval score matching (CISM), a method designed to optimize each concept's 3D Gaussians using concept-aware diffusion scores. The process begins by rendering a novel view image $\mathbf{x}$ and a concept mask $\mathcal{M}$ from the 3D Gaussian $\Theta$. Let $\mathbf{x}_t$ and $\mathbf{x}_s$ denote latents at timesteps $t$ and $s$, where $s = t - \delta_T$, that are derived through DDIM inversion [Song *et al.*, 2020] with null text prompts (i.e. "

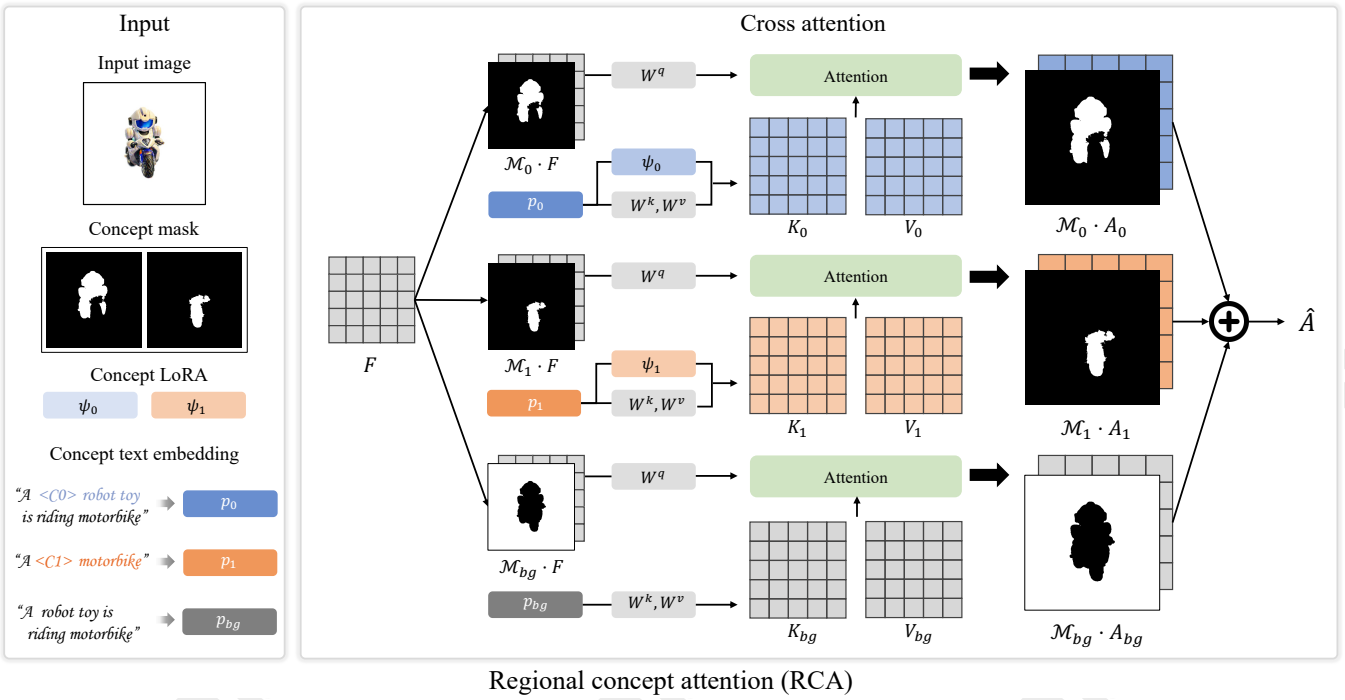



Figure 4: The Regional Concept Attention (RCA) modulates the cross-attention layer in the diffusion model. Individual concept query vectors are computed with image features and each concept masks. Subsequently, key and value vectors for each concept are derived using concept-specific LoRAs and prompts. Then concept-specific attention features are computed with each query, key, and value. The final cross-attention features are aggregated with masked concept-specific attention features.

"). However, DDIM inversion using a single weight diffusion model [Rombach *et al.*, 2022] lacks concept-specific knowledge, leading to suboptimal inversion results. To overcome this limitation, we introduce concept-aware DDIM inversion, which adapts the RCA module during the inversion process to incorporate multi-concept knowledge. The proposed concept-aware DDIM inversion is formulated as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\hat{\mathbf{x}}_0^s + \sqrt{1-\bar{\alpha}_t}\epsilon_\phi(\mathbf{x}_s, s, \hat{A}(\Psi, \emptyset, \mathcal{M})). \quad (12)$$

Here, $\hat{\mathbf{x}}_0^s = \frac{1}{\sqrt{\bar{\alpha}_s}}\mathbf{x}_s - \frac{\sqrt{1-\bar{\alpha}_s}}{\sqrt{\bar{\alpha}_s}}\epsilon_\phi(\mathbf{x}_s, s, \hat{A}(\Psi, \emptyset, \mathcal{M}))$, and $\emptyset$ and $\hat{A}(\cdot)$ denote null text prompts and aggregated concept features using the RCA module, respectively. Technically, the null text prompt is tokenized into a $<BOS>$ token followed by a sequence of $<EOS>$ tokens of maximum token length, which is encoded into a null text embedding via a text encoder. The null text embedding is then processed by the RCA module to produce an unconditional part of the concept-aware diffusion score. This diffusion score is subsequently used to predict $x_s \rightarrow x_t$ with Eq. (12). The CISM loss is then computed using the following equation:

$$\nabla_\Theta \mathcal{L}_{CISM} = \mathbb{E}_{t,\epsilon}[w(t)(\epsilon_\phi(\mathbf{x}_t; t, \hat{A}(\Psi, P, \mathcal{M})) - \epsilon_\phi(\mathbf{x}_s; s, \hat{A}(\Psi, \emptyset, \mathcal{M})))\frac{\partial \mathbf{x}}{\partial \Theta}]. \quad (13)$$

Here, $\Psi$ and $\mathcal{M}$ denote concept LoRA and masks, while $P$ and $\emptyset$ denote the set of concept prompts and null prompts, respectively. Using the CISM loss, we can effectively update the 3D Gaussian, ensuring individual concept identities.

# 4 Experiments

## 4.1 Datasets

We selectively choose real concept image data from the Custom Diffusion [Kumari *et al.*, 2023] and DreamBooth [Ruiz *et al.*, 2023] datasets, which contain 13 unique objects (three wearables and 10 unique objects). This selection is made to explore three specific cases: 1) multiple subjects, 2) property change, and 3) interaction. First, the multiple subjects case involves generating 3D models that incorporate several distinct objects simultaneously. Second, the property change case focuses on subjects with altered attributes, such as different poses (e.g., "*jumping*" or "*sitting*"). Third, the interaction case examines where multiple subjects interact in complex ways, such as one subject "*wearing*" another. These cases evaluate MultiDreamer3D's ability to both preserve concept identity and maintain the presence of objects while handling complex cases such as property changes or interactions. To comprehensively address these cases, we craft and utilize 47 text prompts specifically designed to cover these three cases.

## 4.2 Baseline Methods

In the absence of multi-concept customization method in 3D, we devise a series of baseline methods using existing 2D approaches. The most intuitive and straightforward baseline involves adapting multi-concept 2D diffusion model to train a single 3D model with interval score matching (ISM) [Liang *et al.*, 2023]. Here, we establish two baselines: 3DGS + ISM with FedAVG [McMahan *et al.*, 2017] and Mix-of-Show [Gu


Figure 5: Qualitative results. We compare our method with other baselines in three cases, multiple subjects, property change, and interaction. The red dashed line indicates the objects mentioned in the text prompt that are missing.

| Method | Text-align ↑ | Image-align ↑ |
|---|---|---|
| 3DGS + ISM with Mix-of-Show [Gu *et al.*, 2024] | 0.2024 | N/A |
| 3DGS + ISM with FedAVG [McMahan *et al.*, 2017] | 0.2396 | N/A |
| LG + ISM with Mix-of-Show [Gu *et al.*, 2024] | 0.2199 | 0.6081 |
| LG + ISM with FedAVG [McMahan *et al.*, 2017] | 0.2578 | 0.6338 |
| MultiDreamer3D (Ours) | **0.2732** | **0.6582** |

Table 1: Quantitative results. We assess the text-concept alignment with 3D models using CLIP scores. Here, ours is LG + CISM.

*et al.*, 2024]. For the 3DGS, we initialize the 3D Gaussian using a randomly generated sphere. In the FedAVG approach, multiple single-concept DB-LoRA weights [Ruiz *et al.*, 2023] are merged into a single LoRA weight using a weighted sum. Similarly, in the Mix-of-Show method, multiple ED-LoRA weights [Gu *et al.*, 2024] are merged using a gradient fusion technique. Both the single-concept DB-LoRA and ED-LoRA models are trained on 13 unique objects before applying these techniques for multi-concept training.

| Method | Text-align ↑ | Image-align ↑ |
|---|---|---|
| 3DGS + ISM with Mix-of-Show | 1.62 | 1.91 |
| 3DGS + ISM with FedAVG | 2.17 | 2.18 |
| MultiDreamer3D (Ours) | **4.72** | **4.66** |

Table 2: User study. Participants rate alignment on a 5-point Likert scale (1 indicating strong disagreement, 5 indicating strong agreement). Here, ours is LG + CISM.

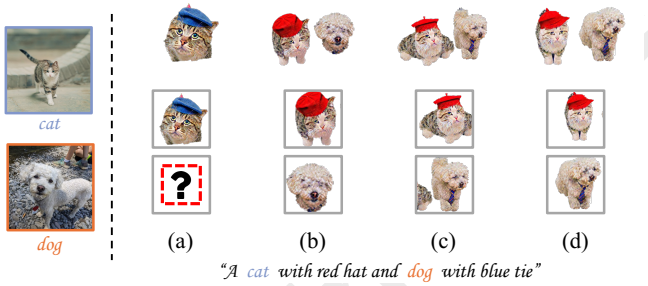



Figure 6: Ablation study. (a) generated with baseline (3DGS + FedAVG [McMahan *et al.*, 2017]), (b) with proposed 3D Layout Controller + CISM, (c) combined with Shap-E [Jun and Nichol, 2023], (d) combined with pointcloud selection.

| Components | Text-align ↑ | Image-align ↑ |
|---|---|---|
| Baseline (3DGS + FedAVG) | 0.2396 | N/A |
| (+) 3D Layout Controller + CISM | 0.2637 | 0.6011 |
| (+) Shape-E | 0.2720 | 0.6487 |
| (+) Pointcloud selection (ours) | **0.2732** | **0.6582** |

Table 3: Ablation study. For the ablation study, we used 3DGS + FedAVG [McMahan *et al.*, 2017] for the baseline.

### 4.3 Evaluation Metrics

We evaluate both text-3D and image-3D alignments with CLIP [Radford *et al.*, 2021]. For text-3D alignment, we render 30 evenly spaced views within an azimuth range of $[-45, 45]$ degrees to avoid occlusion and compute the average CLIP score between the text prompt and these renders. For image-3D alignment, we decompose each concept 3D Gaussians with our concept labeling, rendering each isolated concept from 120 views spanning $[-180, 180]$ degrees, which are compared to the corresponding real concept images.

### 4.4 Qualitative Results

In Figure 5, we compare our method with baselines. Both 3DGS + ISM with FedAVG and 3DGS + ISM with Mix-of-Show struggle to preserve individual concept identities, leading to concept mixing and/or object missing. In contrast, our method consistently maintains object presence and distinct concept identities. In multiple subjects, our method preserves concept identities and aligns with text prompts, avoiding the concept mixing seen in other methods. In property changes, our approach maintains concept integrity and enables pose variations, while others often miss objects or cannot achieve pose variations. In interaction, our method performs comparably to 3DGS + ISM (FedAVG) and better than 3DGS + ISM (Mix-of-Show), effectively capturing complex interactions.

### 4.5 Quantitative Results

In Table 1, we evaluate the image-3D and text-3D alignments of generated outputs. Our method achieves the highest text and image alignment scores, which indicates that our method faithfully reflects text descriptions into multi-concept 3D content while preserving the identities of individual concepts. For image alignment, since other baselines are initialized with a random sphere, isolating the concept 3D Gaussians for these baselines is not feasible. For fair comparison, we utilize our 3D Layout Generator (LG) module to initialize the 3DGS (third and fourth rows of Table 1). This comparison demonstrates that the RCA, followed by self-attention layers in diffusion model, effectively preserves concept identities by maintaining long-range dependencies and ensuring scene coherence across the entire image.

### 4.6 User Study

To demonstrate the effectiveness of our method, we conduct a user study with 32 participants. The study compares 10 3D samples, where participants evaluate three methods based on two criteria: 1) text alignment, assessing how well the 3D model reflects the text prompts, and 2) image alignment, measuring how accurately the 3D model represents real concept images. Participants rate each model on a 5-point Likert scale [Joshi *et al.*, 2015], where 1 signifies "strongly disagree" and 5 signifies "strongly agree". The results are presented in Table 2. Our method achieves the highest human preference for both text and image alignment across all baselines, demonstrating its ability to accurately reflect text prompts and real concept images.

### 4.7 Ablation Study

In Figure 6 and Table 3, we demonstrate the effectiveness of the components in our method. Figure 6 (a) shows the generation of the baseline model (3DGS + FedAVG), which suffers from object missing. Figure 6 (b) presents the generation using our 3D Layout Controller with CISM, which successfully maintains the presence of individual objects. Figure 6 (c) showcases improved geometry in the generated outputs but still suffer from distorted geometry. Figure 6 (d) highlights further enhanced results enabled by the selection module.

## 5 Conclusion

In this paper, we introduced MultiDreamer3D, a method for multi-concept 3D customization that effectively addresses the challenges of object missing and concept mixing. Our 3D Layout Generator facilitates the presence of concept objects and coherent layout context through the use of a 3D layout controller and selective concept point cloud generator. By initializing 3D Gaussian Splatting with explicit concept labeling, we enable clear concept identification. The subsequent update of the 3D Gaussians using Concept-aware Diffusion Guidance ensures the preservation of distinct identities of each concept. Our results showed that MultiDreamer3D is effective across various baselines.

## Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2022R1C1C100849612) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT): (No.2022-0-00959, No.RS-2022-II220959 (Part 2) Few-Shot Learning of Causal Inference in Vision and Language for Decision Making), (No.RS-2022-II220264, Comprehensive Video Understanding and Generation with Knowledge-based Deep Logic Neural Network), (No.RS-2020-II201336, Artificial Intelligence Graduate School Program (UNIST)), and Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City.

## References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Gal *et al.*, 2022] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[Gu *et al.*, 2024] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[Joshi *et al.*, 2015] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403, 2015.

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

[Kerbl *et al.*, 2023] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.

[Kumari *et al.*, 2023] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023.

[Liang *et al.*, 2023] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[Mildenhall *et al.*, 2021] 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.

[Poole *et al.*, 2022] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[Radford *et al.*, 2021] 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.

[Rombach *et al.*, 2022] 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.

[Ruiz *et al.*, 2023] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[Saharia *et al.*, 2022] 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.

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