Concept-as-Tree: Synthetic Data is All You Need for VLM Personalization

Anonymous CVPR submission

Paper ID 10

Abstract

Vision-Language Models (VLMs) have demonstrated exceptional performance in various multi-modal tasks. Recently, there has been an increasing interest in improving the personalization capabilities of VLMs. To better integrate user-provided concepts into VLMs, many methods use positive and negative samples to fine-tune these models. However, the scarcity of user-provided positive samples and the low quality of retrieved negative samples pose challenges for fine-tuning. To reveal the relationship between sample and model performance, we systematically investigate the impact of positive and negative samples (easy and hard) and their diversity on VLM personalization tasks. Based on the detailed analysis, we introduce Concept-as-Tree (CaT), which represents a concept as a tree structure, thereby enabling the data generation of positive and negative samples with varying difficulty and diversity for VLM personalization. With a well-designed data filtering strategy, our CaT framework can ensure the quality of generated data, constituting a powerful pipeline. We perform thorough experiments with various VLM personalization baselines to assess the effectiveness of the pipeline, alleviating the lack of positive samples and the low quality of negative samples. Our results demonstrate that CaT equipped with the proposed data filter significantly enhances the personalization capabilities of VLMs across the MyVLM, Yo'LLaVA, and MC-LLaVA datasets. To our knowledge, this work is the first controllable synthetic data pipeline for VLM personalization. The code will be released.

1. Introduction

Vision-Language Models (VLMs) [6, 30, 33] have produced impressive results across various tasks, showcasing their potential as AI assistants [11, 21, 59]. Despite their success, VLMs still have difficulty generating personalized concepts in their responses, such as generating concept identifiers like ⟨Bob⟩ or ⟨Lina⟩. Several recent studies [2, 3, 19, 41] have explored the personalization of VLMs to address this challenge and seamlessly incorporate VLMs into everyday life. Among them, Yo'LLaVA [41] adopts a test-time fine-tuning strategy that can inject user-related knowledge into VLMs while preserving the models' existing knowledge as much as possible.

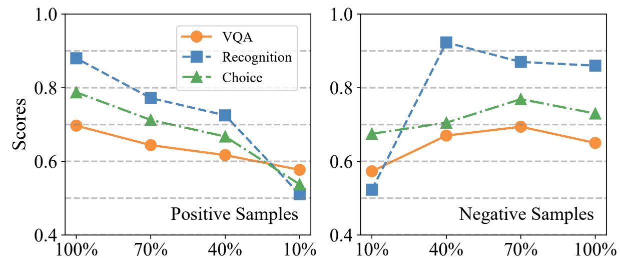

Figure 1. **The performance of Yo'LLaVA varies with the number of positive and negative samples.** (Left) Various tasks show a noticeable decrease in performance when the number of positive samples is limited. (Right) As the number of retrieved negative samples increases, the performance of various tasks does not improve consistently. This might be due to uncertainty in the image retrieval process, leading to low-quality negative samples.

Although test-time fine-tuning has demonstrated its effectiveness, its success largely hinges on the availability of both positive and negative samples during the fine-tuning process. In real-world applications [8, 50, 55], positive samples available for personalizing models are often limited. For instance, users typically provide only 1 to 3 concept images rather than more than 10 for convenience, significantly limiting the model's personalization performance. Furthermore, acquiring negative samples often depends on image retrieval [40], which is a cumbersome and unpredictable process, leading to difficulties in ensuring the quality of these negative samples. We performed preliminary experiments and the results in Fig. 1 support these findings. Although synthetic data can solve the issues mentioned, three major challenges still need to be addressed:

1. The specific roles of positive and negative samples in VLM personalization tasks are still not well understood. For instance, is the use of easy negative samples always necessary?
2. The effect of sample diversity on personalization tasks requires more in-depth investigation. Furthermore, it is important to clarify how the diversity of negative samples affects the performance of the model.

CVPR #10

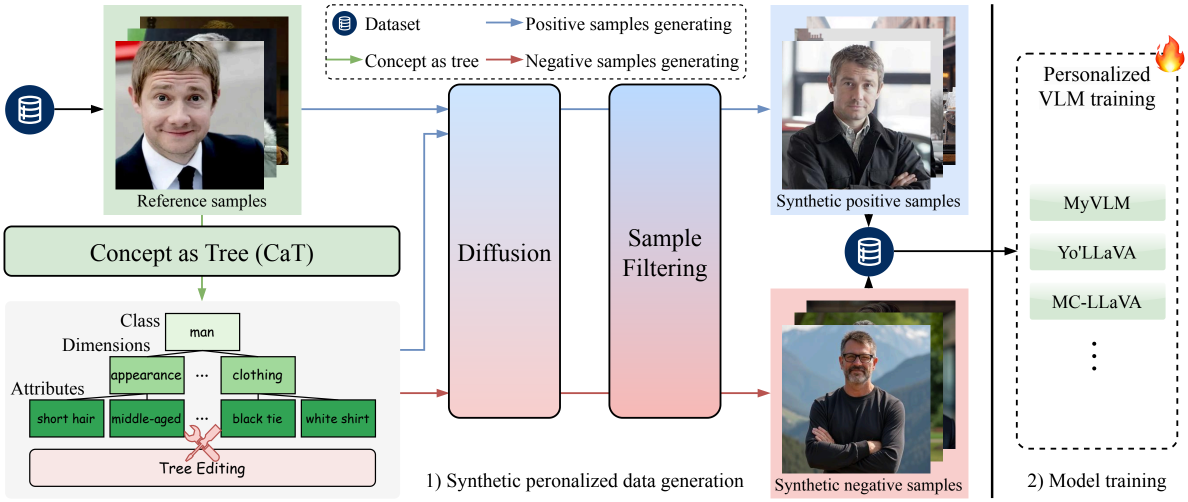

Figure 2. **Overview of unified and controllable data synthesis pipeline and personalized model training.** After a systematic analysis of positive and negative samples (Sec. 3.1), we utilize LLM and VLM to construct the concept tree and edit it to achieve controllable generation (Sec. 3.2). We then propose a well-designed data selection module and a new metric named PCS score to ensure the quality of synthetic data (Sec. 3.3). The ultimate high-quality data can be used to enhance test-time finetuning methods.

3. How can we ensure the quality of generated samples? This is closely linked to the model's performance.

To address the challenges mentioned above, we first conducted a series of observational experiments. Our findings revealed that positive samples generally enhance model performance, easy negative samples improve recognition and hard negative samples boost visual question answering capabilities. We further analyzed easy and hard negative samples with varying diversity and identified specific data diversity requirements for each sample type. These experiments are essential because they provide significant insights into understanding the roles of different data types and their diversity requirements, ultimately leading to enhanced model performance across various VLM personalization tasks.

Based on these findings, we propose **C**oncept **a**s **T**ree (CaT), a unified and controllable synthetic data framework for VLM personalization. Specifically, we define a three-layer tree as a representation of concept information in which the root node represents the concept category and the leaf nodes represent its attributes. By leveraging the power of Large Language Models and VLMs, we can automate the construction of trees for each concept. To synthesize positive samples, we use the root node as a prompt in the diffusion model, which is fine-tuned with concept images provided by the user. To create easy negative samples, we modify the root node information and generate a tree for controllable data generation. To create hard negative samples, we modify the leaf nodes of the concept-correlated tree in order to generate the prompt. Controlling diversity is equivalent to varying the number and types of editing operations performed on the tree during each generation.

A comprehensive data synthesis process must ensure the quality of generated data. Therefore, we propose a well-designed and easily implementable method for data filtering. We state that the information in an image includes two types: concept-specific features that are unique to each concept and concept-agnostic features that possess general characteristics [36]. Therefore, we apply the same perturbation to the synthesized images and calculate the distance between the synthesized images and the user-provided images both before and after the perturbation. The difference between these two distances can be defined as a Perturbation-based Concept-Specific (PCS) score, which serves as a key metric for assessing sample quality. We set filtering thresholds based on PCS score to ensure high data quality when filtering various types of generated samples. Through the aforementioned methods, we achieved a unified and controllable pipeline of data synthesis, as shown in Fig. 2.

To summarize, our work contributes in multiple aspects:

- We systematically study the impact of positive and negative samples and their diversity on VLM personalization.
- We propose a comprehensive synthetic data pipeline, which consists of CaT and a data filtering strategy.
- We utilize generated data to conduct extensive experiments on the MyVLM, Yo'LLaVA and MC-LLaVA datasets, where equipped with the proposed pipeline, all fine-tuning-based methods show significant improvements in various VLM personalization tasks, achieving state-of-the-art results across all datasets.

2. Related work

Personalization for Vision Language Models. Vision-Language Models(VLMs) [26, 27, 34] have exhibited remarkable capabilities across diverse domains, such as data

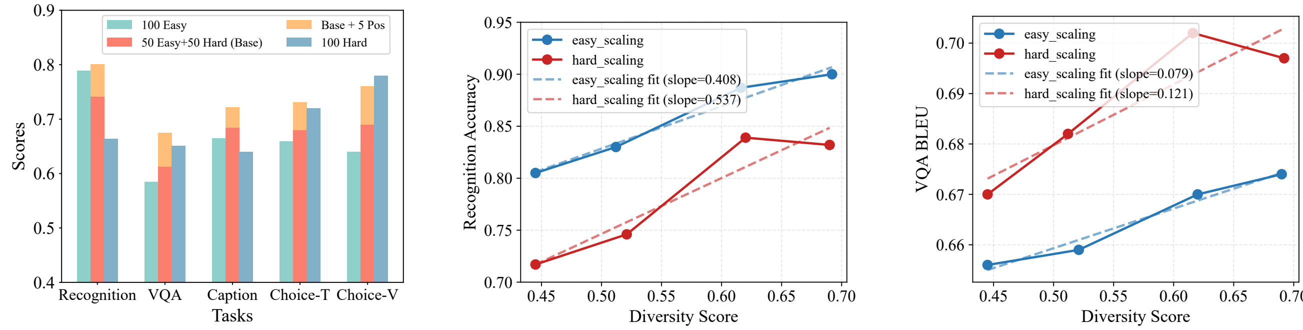

Figure 3. **Explore the role of positive and negative samples in personalization and their demand for diversity.** (Left) To better investigate sample effectiveness, we conduct experiments on half the number of positive and negative sample contrasts to the original setting. Adding positive samples brings a general raising among the tasks while easy and hard negative samples have their different improvement preferences. (Middle&Right) Keeping the number of samples unchanged, we vary the diversity of easy and hard negative samples. The performance improvements brought by them vary across different tasks. However, hard negative samples are more sensitive to diversity. Moreover, excessive diversity may lead to a decline in task performance. The diversity score is calculated by clustering the retrieved negative samples using K-means and measuring the distance of each data point to the cluster centroids.

mining [39], fine-grained understanding [32], and visual question answering [7]. To seamlessly integrate these models into daily life, there has been growing interest in VLM personalization [2, 4, 41, 44]. This task is first introduced by MyVLM [2], employing an additional module strategy for improved injection of concept information into VLMs. YoLLaVA [41] and MC-LLaVA [4] utilized efficient end-to-end fine-tuning methods to tackle challenges in both single and multi-concept scenarios. The performance of these fine-tuning methods is highly dependent on the positive and negative samples. However, users often employ a limited number of concept images (i.e., 1 to 3 images), and the challenge of acquiring negative samples complicates fine-tuning approaches. In this study, we investigate the impact of positive and negative samples and their diversity on VLM personalization tasks. And we propose a complete data synthesis pipeline, thus overcoming challenges in training VLM models with limited data availability.

Learning from Synthetic Data. The use of synthetic data [18, 24, 38] has been extensively studied across various computer vision tasks, such as classification [49], segmentation [56], and object detection [15]. With the rapid consumption of existing data resources driven by the development of Large Language Models (LLMs) and VLMs, researchers aim for models to achieve further improvements using synthetic data [23, 29, 37, 42]. Synthetic corpora are generated for instruction fine-tuning of large models. At the same time, large models leverage recaptioned datasets to enhance quality and improve CLIP performance [14]. To better support the learning of VLMs, direct generation of text-image pairs is pursued [31]. However, existing generation methods only consider class information [52]. These methods rely on simplistic prompt templates, attributes, and class information, either manually defined or LLM-generated [12, 53]. To address the issue of insufficient data in VLM personalization scenarios, we supplement the dataset by synthesizing concept-centric positive and negative sample data. We adopt a unified framework to control the positivity and diversity of the generated samples.

3. Method

We present a comprehensive data synthesis pipeline to address the challenge of personalizing Vision-Language Models (VLM). This pipeline views the concept as a tree structure and generates both positive and negative samples through tree operations. It also encompasses a well-designed filtering strategy to select the generated samples. The overall design of our pipeline is illustrated in Fig. 2. Specifically, in Sec. 3.1, we systematically study the impact of positive and negative samples on VLM personalization tasks. Building on this, we introduce the CaT framework and use tree operations to generate synthetic positive and negative samples in Sec. 3.2. Furthermore, to ensure the quality of the synthetic samples, we propose a filtering strategy based on the PCS score in Sec. 3.3.

3.1. Impact of Positive and Negative Samples

In this section, we examine the effects of positive and negative samples and their diversity on the personalization of VLM. Based on the positive samples provided by the MC-LLaVA dataset and negative samples retrieved online, we conducted a series of experiments with the Yo'LLaVA method. As illustrated in Fig. 3 left, we found that an increase in positive samples consistently improves outcomes across all tasks. While easy negative samples primarily enhance recognition abilities, hard negative samples improve conversational capabilities. Interestingly, the captioning task fundamentally reflects conversational abilities. However, since the evaluation method assesses the presence of a concept identifier related to recognition capabilities, the training data that includes both easy and hard negative sam-

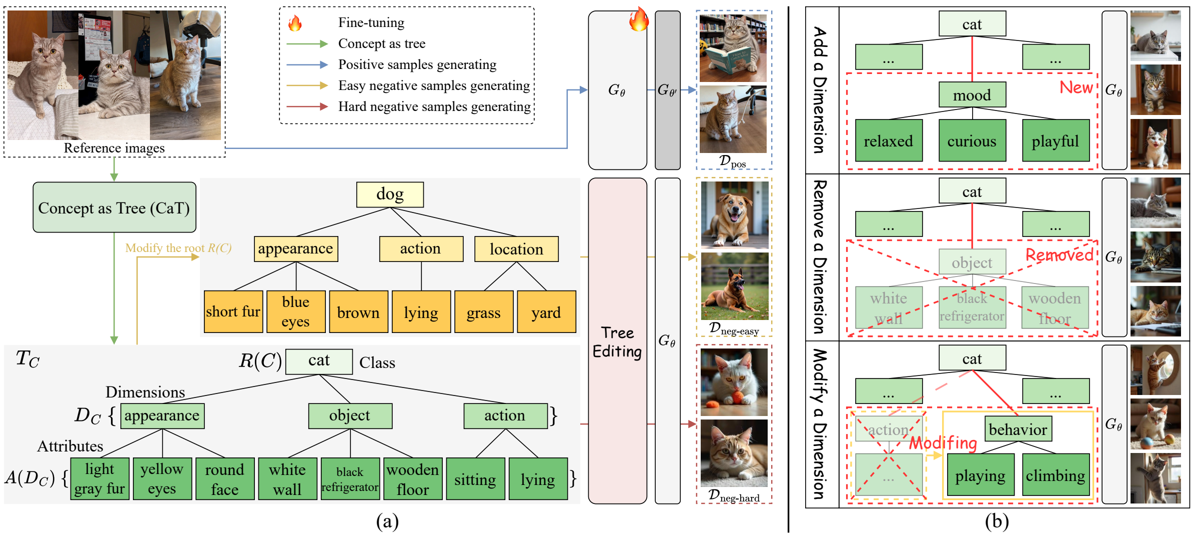

Figure 4. **Using concept trees to synthesize positive and negative samples and three editing operations for constructing diverse concept trees.** (a) We first utilize the CaT framework and reference images to obtain the concept tree T_C. Then, we employ a fine-tuned diffusion model along with the information from the root node $R(C)$ to synthesize positive samples. For easy negative samples, we modify the class of the root node $R(C)$ and obtain a new concept tree for synthesis. Finally, we apply three editing operations to the original concept tree to obtain multiple trees for synthesizing hard negative samples with diversity. The model used for synthesizing negative samples does not require fine-tuning. (b) We visualize the three mentioned tree editing operations in detail. Compared with the reference image, adding a dimension of "mood" actually leads to different emotions in cats. After removing the "object" dimension, the environment around the cats becomes simpler. By modifying a dimension, the behavior of cats in the composite image becomes more diverse.

ples yields the best performance. Furthermore, we find that easy negative samples are essential. Using only hard negative samples can help the model memorize the visual features of a concept and perform well in dialogue tasks. However, its ability to recognize the concept identifier significantly declines. Easy negative samples are irreplaceable, as they are diverse and widely sourced, providing valuable context for the model's learning process. Recognition tasks are often the initial step for a VLM acting as a personal assistant, representing an essential capability that is crucial.

Based on the above mentioned observation, we conducted additional experiments that examined different levels of diversity for negative samples. Here, diversity is calculated by clustering a set of negative samples and calculating the distance from the cluster center to each sample. From Fig. 3 middle and right, keeping the number of positive samples unchanged, it is evident that the gain curves for easy and hard negative samples differ across various tasks, indicating that their diversity requirements are distinct. Low diversity of both easy and hard negative samples can simultaneously restrict the model performance trained on them. When the diversity of hard negative samples is excessively high, it often introduces noise during training, resulting in a decline in model performance. These experiments clarify the role of diversity and motivate us to manage the diversity when creating negative samples. Therefore, we propose representing the concept as tree to facilitate the generation of negative samples with various difficulties and diversities.

3.2. Concept as Tree Framework

To integrate new concepts into a pre-trained VLM using synthetic data, we propose the Concept-as-Tree (CaT) framework, which transforms a concept into a structured tree for controllable data generation. Given a user-provided dataset $\mathcal{D}_{\text{user}}$, which contains the positive image samples $I_{\text{pos}} \in \mathcal{D}_{\text{user}}$, our goal is to construct a fine-tuning dataset:

$$\mathcal{D}_{\text{train}} = \mathcal{D}_{\text{user}} \cup \mathcal{D}_{\text{pos}} \cup \mathcal{D}_{\text{neg-easy}} \cup \mathcal{D}_{\text{neg-hard}} \quad (1)$$

where $\mathcal{D}_{\text{pos}}$ are synthetic positive samples generated based on I_{pos}, while $\mathcal{D}_{\text{neg-easy}}$ and $\mathcal{D}_{\text{neg-hard}}$ are negative samples obtained via concept tree modifications.

Concept Tree Representation and Construction. To enable controllable data generation, inspired by SS-DLLM [39], we represent each concept C as a hierarchical tree $T_C = (R(C), D_C, A(D_C))$. As shown in Fig. 4(a), $R(C)$ serves as the root node, encapsulating the high-level category of the concept *(e.g., "cat", "dog", etc.)*, while $D_C = \{D_1, D_2, \ldots, D_m\}$ defines a set of attribute dimensions that distinguish various aspects *(e.g., "appearance", "behavior", "location", etc.)*. Each dimension D_i is associated with a set of attributes $A(D_i)$, collectively forming $A(D_C) = \{A(D_1), A(D_2), \ldots, A(D_m)\}$ *(for instance, the "behavior" dimension might include attributes like "sitting", "lying" and "climbing", etc.)*.

To automatically construct T_C, we adopt a three-step process. First, a pre-trained VLM generates textual de-

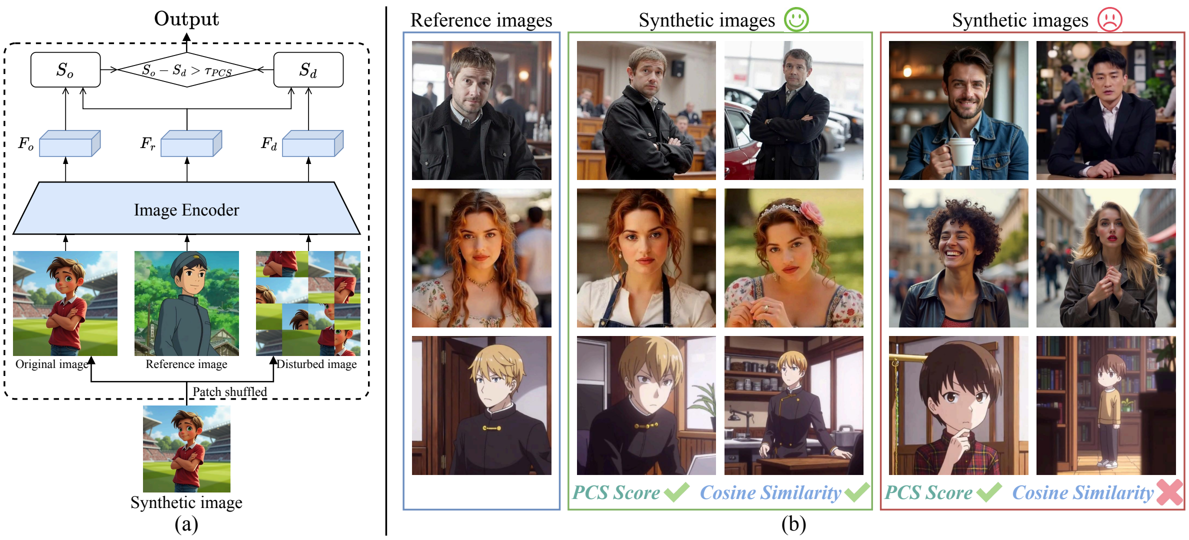

Figure 5. **Filtering method based on the PCS score and the visualization of the selection of high-quality images.** (a) We perform patch shuffling on the synthetic images and extract the features F_o, F_d and F_r from the original images, disturbed images and reference images, respectively. Subsequently, we calculate the image similarities S_o and S_d, then select high-quality images with high PCS score. (b) The reference images are marked with blue box, images containing more CS information are highlighted with green box and those containing more CA information are indicated with red box. Using image cosine similarity alone cannot distinguish between these two types of synthetic images, but our proposed PCS score can effectively filter out the images on the right.

scriptions from I_{pos}, capturing key visual features (Image Description). Next, we aggregate these descriptions at the batch level to extract meaningful dimensions D_C and corresponding attributes $A(D_C)$ (Batch Summarization). Finally, to improve accuracy and interpretability, a self-refine mechanism iteratively adjusts T_C based on feedback from a multi-round voting mechanism (Self-Refine). For a given concept image, if it cannot be assigned to the same attribute multiple times, this suggests redundancy or unsuitable attributes. Adjustments enhance the orthogonality and completeness of the attributes. Detailed prompts and visualizations of the concept tree can be found in the Appendix.

Personalized Data Generation. After constructing T_C, we proceed to generate synthetic samples (see Fig. 4(a)). We define an image generation model G_θ for controllable data synthesis and a transformation function $F(R(C), D_C)$ to systematically modify either the tree root or its dimensions.

- **Positive Samples:** Generated by fine-tuning G_θ on user-provided images I_{pos}, conditioned on $R(C)$:

$$\mathcal{D}_{\text{pos}} = G_\theta(I_{\text{pos}}, R(C)) \tag{2}$$

- **Negative Samples:** Constructed by editing T_C to introduce controlled variations.
  - **Easy Negative Samples:** Generated by replacing $R(C)$ with another category $R(C')$ and editing dimensions to D'_C, forming a modified tree $T_{C'} = F(R(C'), D'_C)$. Synthetic images are produced as $\mathcal{D}_{\text{neg-easy}} = G_\theta(T_{C'})$.
  - **Hard Negative Samples:** Retaining $R(C)$ while modifying dimensions of T_C to form $T'_C = F(R(C), D'_C)$. The generated samples resemble C visually but differ semantically, represented as $\mathcal{D}_{\text{neg-hard}} = G_\theta(T'_C)$.

Tree Editing Operations. We define three fundamental tree editing operations $F(R(C), D'_C)$ (see Fig. 4(b)), which modify T_C to systematically generate negative samples:

$$D'_C = \begin{cases} D_C \cup \{D_{\text{new}}\}, & \text{(Add a Dimension)} \\ D_C \setminus \{D_{\text{remove}}\}, & \text{(Remove a Dimension)} \\ (D_C \setminus \{D_{\text{old}}\}) \cup \{D_{\text{new}}\}, & \text{(Modify a Dimension)} \end{cases} \tag{3}$$

where D'_C is the updated dimension set, D_{new} is an added dimension, and D_{remove} is a removed one. The modification operation replaces D_{old} with D_{new}. The different forms of perturbations and their application frequencies significantly affect the diversity of the generated samples, which will be discussed in the ablation experiments.

3.3. Sample Filtering with PCS Score

To ensure the quality of synthesized samples, we implement a filtering process for all generated samples. We classify the information contained in images into two types: concept-specific (CS) and concept-agnostic (CA). The former refers to features that are distinctive to the concept, such as appearance or type, while the latter encompasses other elements present in the image, such as background or weather conditions. We consider synthetic data predominantly containing CS information to be regarded as high-quality samples. Conversely, samples with relatively more CA information may be categorized as low-quality samples.

Inspired by this, we first mix patches from the reference and original images, then calculate the similarity between the altered images and the reference images using the CLIP model [43]. The difference between the similarities before and after perturbation is defined as the Perturbation-based Concept-Specific (PCS) score. Typically, the PCS score of an image should be quite high because mixing patches alters the appearance or shape of objects within the image, thus disrupting the original CS information. This disruption causes CLIP to struggle to understand the conceptual relationship between the synthetic image and the reference image, leading to a notable reduction in the similarity of the altered image. However, if the similarity of an image relies not on CS information but on CA information, CLIP might perceive uncommon or incorrect object relationships with the reference image. In such cases, the similarity before and after perturbation changes little, meaning the PCS score will be small. Therefore, we claim that the PCS score can be used to determine the information components in the image. A high PCS score indicates that the image carries more CS information, while a low PCS score suggests that the image contains more CA information.

As illustrated in Fig. 5(a), let the reference image be $I_{\text{reference}} \in \mathcal{D}_{\text{user}}$, the synthetic image before perturbation be I_{original} and the image after perturbation be $I_{\text{disturbed}}$. We extract the visual features F_r, F_o and F_d of the corresponding images using CLIP [43] image encoder f_θ:

$$F_I = f_\theta(I), \tag{4}$$

and then calculate the cosine similarity S_o and S_d, respectively. The cosine similarity $S_x \in [-1, 1]$ is calculated as:

$$S_x = \frac{F_1 F_2^\top}{\| F_1 \| \cdot \| F_2 \|}, \tag{5}$$

The difference between S_o and S_d is the PCS score. Based on this score, we can select high-quality synthetic data, i.e., $G(I) = \{I \mid \text{PCS}(I) > \tau_{PCS}\}$, where $G(I)$ represents high-quality samples and τ_{PCS} is the threshold for containing more CS information. As shown in Fig. 5(b), when only using the filtering method based on cosine similarity, both the images in the green and red boxes achieve high similarity scores. However, the concept in the red box lacks specific visual features of the reference image's concept, such as hair, facial structure and body shape. Instead, the background and other CA information in these images have a greater impact on the similarity. However, through the newly proposed perturbation-based filtering method, we can effectively identify and filter out images with more CA information using PCS score, while retaining high-quality images with more CS information for better personalization.

4. Experiments

4.1. Experimental Setup

Implement Details. Regarding the experiments conducted on all datasets, the quantity of positive and negative samples for all the baselines follows their original settings. The training data used for Baseline (syn) is entirely generated by CaT. To ensure the consistency of the positive sample count, Baseline (Real+Syn) is articulated as comprising 1 to 3 original concept images and several synthesized positive samples. Baseline (Real+Syn) (Plus) indicates that an equal number of synthesized positive samples have been added to the previous positive samples. All negative sample quantities in the above experiments are strictly controlled for consistency. When the tree dimension corresponding to the concept is greater than or equal to 5, we do not use Add. When it is less than or equal to 3, we do not use Remove. For other instances, we utilize the three editing operations evenly for experiments. With regard to the synthesized images, we utilize GPT-4o to generate the instruction text pairs required for their training. Details of baseline, datasets and training hyperparameters can be found in the Appendix.

4.2. Personalized Capability from Synthetic Data

The ability to recognize concepts is fundamental to personalized VLM. As illustrated in Tab. 1, it is evident that after training with extra synthesized data, the model's recognition ability shows an average improvement of 4.1% on the MC-LLaVA dataset, 2.9% on the Yo'LLaVA dataset, and 1.7% on MyVLM dataset. However, training solely on synthesized data does not yield consistent improvements, which may be attributed to a distribution shift between synthesized data and the real data used for testing. Within the MC-LLaVA framework, despite the use of purely synthesized data, performance still surpasses the baseline, which is attributed to the efficient utilization of visual information relevant to concepts in the MC-LLaVA methods. Including original real images of concepts in the training data can mitigate this phenomenon, resulting in a modest improvement.

We further evaluate the capability of our method to address questions related to personalized concepts. The introduction of synthesized data yields an increase of up to 3.8% in choice-based tasks, 4.2% in Visual Question Answering (VQA) tasks, and 5.3% in captioning tasks, achieving competitive performance with GPT-4o in certain scenarios. Training with purely synthesized data aligns with the trends observed in recognition tasks. Notably, on text-related tasks such as Choice-T and VQA, it consistently enhances model performance. This improvement stems from our targeted modifications in negative sample generation, where specific features of positive samples are altered. By reducing the model's reliance on these adjusted features, it learns to focus more on the concept's intrinsic attributes. As a result, the model achieves higher accuracy in concept-related multiple-choice tasks and generates more precise descriptions of the concept's defining characteristics.

In contrast, RAP-MLLM processes only a single image for each concept. As a result, the synthesized data fails

Dataset	MC-LLaVA Evaluation					Yo'LLaVA Evaluation			MyVLM Evaluation	
Method/Task	**Rec Acc**	**Choice-V Acc**	**Choice-T Acc**	**VQA BLEU**	**Caption Recall**	**Rec Acc**	**Choice-V Acc**	**Choice-T Acc**	**Rec Acc**	**Caption Recall**
GPT-4o [1]+Prompt	0.746	0.888	0.712	0.728	0.836	0.856	0.932	0.897	0.891	0.969
MyVLM [2](Real)	0.795	0.779	-	0.640	0.714	0.911	0.897	-	0.938	0.921
MyVLM(Syn)	0.788	0.772	-	0.653	0.711	0.908	0.895	-	0.935	0.918
MyVLM(Real+Syn)	0.799	0.783	-	0.662	0.722	0.916	0.907	-	0.945	0.926
MyVLM(Real+Syn)(P)	0.827	0.805	-	0.685	0.740	0.945	0.916	-	0.960	0.947
	+3.2%	+2.6%	-	+4.5%	+2.6%	+3.4%	+1.9%	-	+2.2%	+2.6%
Yo'LLaVA [41](Real)	0.841	0.801	0.703	0.643	0.701	0.924	0.929	0.883	0.964	0.931
Yo'LLaVA(Syn)	0.840	0.805	0.705	0.654	0.699	0.921	0.925	0.885	0.962	0.927
Yo'LLaVA(Real+Syn)	0.851	0.817	0.711	0.662	0.714	0.928	0.933	0.891	0.969	0.938
Yo'LLaVA(Real+Syn)(P)	0.885	0.845	0.738	0.682	0.754	0.946	0.943	0.900	0.981	0.950
	+4.4%	+4.4%	+3.5%	+3.9%	+5.3%	+2.2%	+1.4%	+1.7%	+1.7%	+1.9%
MC-LLaVA [4](Real)	0.917	0.890	0.727	0.684	0.750	0.947	0.941	0.893	0.975	0.959
MC-LLaVA(Syn)	0.920	0.892	0.731	0.695	0.755	0.951	0.945	0.896	0.978	0.963
MC-LLaVA(Real+Syn)	0.928	0.890	0.739	0.704	0.768	0.958	0.947	0.901	0.984	0.968
MC-LLaVA(Real+Syn)(P)	0.963	0.928	0.760	0.726	0.803	0.977	0.953	0.910	0.987	0.971
	+4.6%	+3.8%	+3.3%	+4.2%	+5.3%	+3.0%	+1.2%	+1.7%	+1.2%	+2.2%
RAP-MLLM [19]	0.747	0.832	0.709	0.424	0.711	0.845	0.917	0.874	0.870	0.937

Table 1. **Results of synthetic data used in different personalization methods.** All methods achieve improvements across three datasets when using synthetic data. When trained with a combination of original positive data and all synthetic data, all methods can even achieve results comparable to GPT-4o. Noting, the Choice-T is a pure text task, but MyVLM requires concept images to load the corresponding embeddings during testing time, which makes it unable to perform the Choice-T task. The green numbers represent the improvement of the new SOTA compared to original baseline. Real = Original Data; Syn = Synthetic Data; P = Plus.

to contribute effectively to its learning, further limiting its comprehension of the concept. While RAP-MLLM may initially outperform models like MyVLM and Yo'LLaVA in vision-based multiple-choice tasks, the synthesized data enables these models to surpass RAP-MLLM by developing a deeper understanding of concept-specific features.

Overall, the results obtained from training with purely synthesized data are comparable to those from purely real data, demonstrating the quality assurance of our synthesized data. In scenarios with limited real data, training in conjunction with our synthesized data achieves results that surpass the baseline, preserving the personalization capabilities of VLMs under data scarcity. The stable improvements stem from our thorough understanding of concept information and the highly controllable CaT framework used during data synthesis. Moreover, augmenting real data with additional synthesized data significantly surpasses all baselines, reflecting the efficacy of synthesized data.

4.3. Ablations and Analysis

Tree Editing Operations. We investigate the impact of tree editing operations on hard negative samples within the MC-LLaVA dataset. Index A represents the Yo'LLaVA baseline without any tree editing operation. The diversity score is defined for a set of data; although there may be a distribution shift between retrieved and generated data, their diversity can still be compared. In Tab. 2, the comparison between A, B, E, and G clearly demonstrates that different tree editing operations lead to variations in the diversity of generated data, resulting in varying task performance. Both Add and Modify operations can enhance diversity to some extent, with Add yielding a greater increase than Modify. In contrast, Remove decreases the diversity of generated results due to a reduction in combinable attributes. In accordance with Sec. 3.1, high-diversity negative samples lead to a more pronounced improvement in tasks related to conversation. However, excessive diversity can introduce noise, thus degrading the performance of the model. For the results of B, C, D, E, F, G and H, we observe that an increase in the number of editing operations per instance amplifies their impact on diversity and results. Therefore, our CaT

Methods	Index	TEO		Diversity	Dataset: MC-LLaVA		
		Category	Times		Rec	VQA	Caption
Yo'LLaVA +CaT	A	None	0	0.497	0.841	0.643	0.701
	B	Add	1	0.563	0.852	0.655	0.723
	C	Add	2	0.642	0.866	0.674	0.736
	D	Add	3	0.708	0.834	0.654	0.717
	E	Remove	1	0.451	0.827	0.621	0.687
	F	Remove	2	0.379	0.801	0.605	0.663
	G	Modify	1	0.542	0.847	0.652	0.718
	H	Modify	2	0.613	0.859	0.671	0.734

Table 2. **Ablation study on tree editing operations.** The type and number of operations can both alter the diversity of synthesized data. Excessive or insufficient diversity negatively impacts model performance. TEO = Tree editing operations.

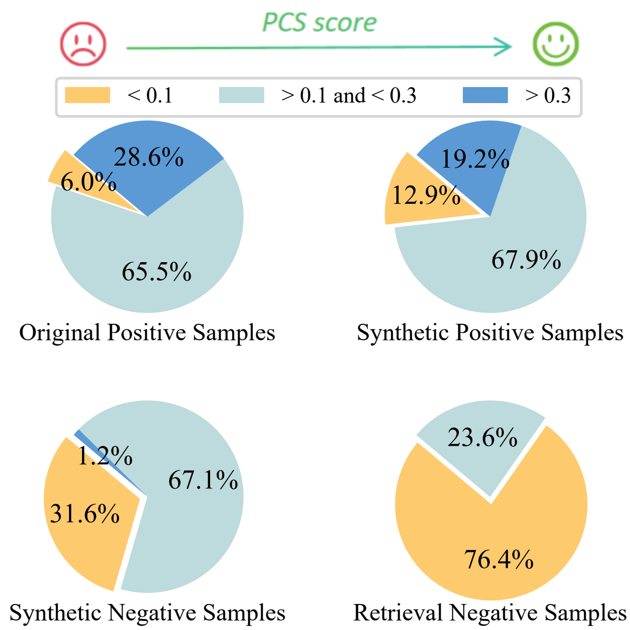

Figure 6. **A significant difference in PCS score between these datasets.** Higher PCS score reflects a greater quantity of CS information, correlating with improved synthetic image quality.

framework can provide precise control over this process, ensuring that optimal diversity leads to stable improvements.

Quality of Synthesized Data. As discussed in Sec. 3.3, a higher PCS score indicates that an image contains more CS information, representing a higher-quality personalized sample. From the results shown in Fig. 6, we observe that the distributions of PCS scores for the original positive samples and the synthetic positive samples are quite similar, demonstrating that our synthetic positive samples can effectively serve the same purpose as the original positive samples. Additionally, we find that the proportion of low PCS scores in the original retrieved negative samples is significantly higher than that in our synthetic negative samples, indicating that the quality of synthetic negative samples surpasses that of retrieved negative samples. Based on these two comparisons, we demonstrate that our CaT framework effectively synthesizes high-quality samples, enabling efficient personalization for vision language models.

Hard Negative Sample Generation. We compare different methods for synthesizing hard negative samples, and the results are presented in Fig. 7. Our CaT framework ensures the diversity of synthetic samples, leading to significant improvements across all tasks. In contrast, the random synthesis of hard negative samples lacks control over the diversity of the results. The baseline using these data may overfit to a narrow range of negative samples, leading to an inability to achieve improvements across all tasks.

Data Component. We analyze the key components of the generated datasets. The original training data consists of provided positive samples and retrieved negative samples. Then we examine the role of synthetic data in these three components by first adding synthetic positive samples and then progressively replacing the retrieved negative samples with synthetic negative samples. At each step, we advance incrementally while keeping the other components unchanged. As shown in Tab. 3, the original Yo'LLaVA already possesses a certain level of personalization capability. However, as samples are progressively transformed into synthesized data, performance on several tasks continues to improve, demonstrating the effectiveness of our CaT approach. The enhancements brought by incorporating easy and hard negative samples vary across different tasks, which is consistent with our observations in Sec. 3.1.

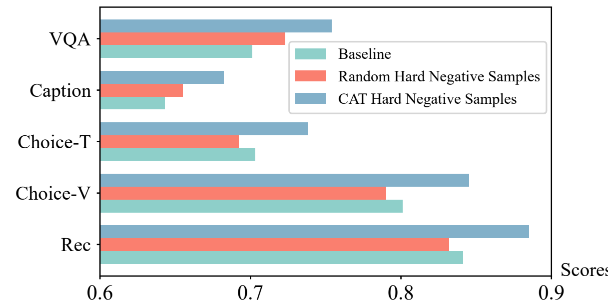

Figure 7. **A comparison of synthesis methods for hard negative samples.** Randomly generated prompts, lacking detailed visual descriptions, tend to produce a large number of similar images. Our CaT framework ensures the diversity of synthetic samples, thereby enhancing the model's personalization capabilities.

Module/Task	Rec	Choice-V	Choice-T	VQA	Caption
Yo'LLaVA	0.841	0.801	0.703	0.643	0.701
w/ Syn Positive	0.858 (+0.017)	0.816 (+0.015)	0.716 (+0.013)	0.656 (+0.013)	0.721 (+0.020)
w/ Syn Easy Neg	0.873 (+0.015)	0.824 (+0.008)	0.721 (+0.005)	0.667 (+0.011)	0.727 (+0.006)
w/ Syn Hard Neg	0.885 (+0.012)	0.845 (+0.021)	0.738 (+0.017)	0.680 (+0.013)	0.754 (+0.027)

Table 3. **Ablation study on data component.**

5. Conclusion

In this work, we present a new roadmap to enhance the personalization capability of Vision-Language Models (VLMs) via leveraging synthetic positive and negative images from advanced generative models. Based on our solid and systematic observation experiment, we propose a unified and controllable data synthesis pipeline, which consists of Concept as Tree framework and a well-designed data selection module. Remarkably, our pipeline requires only 1 to 3 user-provided images to work, synthesizing high-quality samples that enable VLMs to perform comparably to models trained on real images in data-scarce scenarios. This pipeline addresses the critical challenge of insufficient user-provided concept data, offering a controllable, interpretable, and data-efficient solution for VLM personalization tasks, paving the way for VLMs to become more accessible and practical human assistants.

References

[1] 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. 7, 1

[2] Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. Myvlm: Personalizing vlms for user-specific queries. In *European Conference on Computer Vision*, pages 73–91. Springer, 2025. 1, 3, 7

[3] Ruichuan An, Sihan Yang, Ming Lu, Kai Zeng, Yulin Luo, Ying Chen, Jiajun Cao, Hao Liang, Qi She, Shanghang Zhang, and Wentao Zhang. Mc-llava: Multi-concept personalized vision-language model, 2024. 1

[4] Ruichuan An, Sihan Yang, Ming Lu, Kai Zeng, Yulin Luo, Ying Chen, Jiajun Cao, Hao Liang, Qi She, Shanghang Zhang, et al. Mc-llava: Multi-concept personalized vision-language model. *arXiv preprint arXiv:2411.11706*, 2024. 3, 7, 1

[5] Joseph Bamidele Awotunde, Chinmay Chakraborty, and Abidemi Emmanuel Adeniyi. Intrusion detection in industrial internet of things network-based on deep learning model with rule-based feature selection. *Wireless communications and mobile computing*, 2021(1):7154587, 2021. 2

[6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1

[7] Jiajun Cao, Yuan Zhang, Tao Huang, Ming Lu, Qizhe Zhang, Ruichuan An, Ningning Ma, and Shanghang Zhang. Move-kd: Knowledge distillation for vlms with mixture of visual encoders. *arXiv preprint arXiv:2501.01709*, 2025. 3

[8] Nitesh V Chawla and Darcy A Davis. Bringing big data to personalized healthcare: a patient-centered framework. *Journal of general internal medicine*, 28:660–665, 2013. 1

[9] Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. *arXiv preprint arXiv:2402.12501*, 2024. 2

[10] Hongyuan Dong, Zijian Kang, Weijie Yin, Xiao Liang, Chao Feng, and Jiao Ran. Scalable vision language model training via high quality data curation. *arXiv preprint arXiv:2501.05952*, 2025. 2

[11] Nicolai Dorka, Janusz Marecki, and Ammar Anwar. Training a vision language model as smartphone assistant. *arXiv preprint arXiv:2404.08755*, 2024. 1

[12] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. *Advances in neural information processing systems*, 36:79024–79034, 2023. 3

[13] Marco Fabris, Riccardo Ceccato, and Andrea Zanella. Efficient sensors selection for traffic flow monitoring: An overview of model-based techniques leveraging network observability. *Sensors*, 25(5):1416, 2025. 2

[14] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575, 2023. 3

[15] Chengjian Feng, Yujie Zhong, Zequn Jie, Weidi Xie, and Lin Ma. Instagen: Enhancing object detection by training on synthetic dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14121–14130, 2024. 3

[16] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023. 2

[17] Sachin Goyal, Pratyush Maini, Zachary C Lipton, Aditi Raghunathan, and J Zico Kolter. Scaling laws for data filtering–data curation cannot be compute agnostic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22702–22711, 2024. 2

[18] Xu Guo and Yiqiang Chen. Generative ai for synthetic data generation: Methods, challenges and the future. *arXiv preprint arXiv:2403.04190*, 2024. 3

[19] Haoran Hao, Jiaming Han, Changsheng Li, Yu-Feng Li, and Xiangyu Yue. Remember, retrieve and generate: Understanding infinite visual concepts as your personalized assistant, 2024. 1, 7

[20] Xiao He, Mingrui Zhu, Dongxin Chen, Nannan Wang, and Xinbo Gao. Diff-privacy: Diffusion-based face privacy protection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[21] Mengze Hong, Chen Jason Zhang, Chaotao Chen, Rongzhong Lian, and Di Jiang. Dialogue language model with large-scale persona data engineering, 2025. 1

[22] Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. Who's in and who's out? a case study of multimodal clip-filtering in datacomp. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–17, 2024. 2

[23] Renhong Huang, Jiarong Xu, Xin Jiang, Ruichuan An, and Yang Yang. Can modifying data address graph domain adaptation?, 2024. 3

[24] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*, 2022. 3

[25] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1, 2

[26] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2

[27] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 2

[28] Guoliang Li, Xuanhe Zhou, and Xinyang Zhao. Llm for data management. *Proceedings of the VLDB Endowment*, 17(12): 4213–4216, 2024. 2

[29] Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, et al. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. *arXiv preprint arXiv:2402.13064*, 2024. 3

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[31] Hao Liang, Linzhuang Sun, Jingxuan Wei, Xijie Huang, Linkun Sun, Bihui Yu, Conghui He, and Wentao Zhang. Synth-empathy: Towards high-quality synthetic empathy data. *arXiv preprint arXiv:2407.21669*, 2024. 3

[32] Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2403.20271*, 2024. 3

[33] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 1

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2, 1

[35] Jiang Liu, Chun Pong Lau, and Rama Chellappa. Diff-protect: Generate adversarial examples with diffusion models for facial privacy protection. *arXiv preprint arXiv:2305.13625*, 2023. 2

[36] Jiaming Liu, Senqiao Yang, Peidong Jia, Renrui Zhang, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. Vida: Homeostatic visual domain adapter for continual test time adaptation. *arXiv preprint arXiv:2306.04344*, 2023. 2

[37] Zheng Liu, Hao Liang, Xijie Huang, Wentao Xiong, Qinhan Yu, Linzhuang Sun, Chong Chen, Conghui He, Bin Cui, and Wentao Zhang. Synthvlm: High-efficiency and high-quality synthetic data for vision language models. *arXiv preprint arXiv:2407.20756*, 2024. 3

[38] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*, 2024. 3

[39] Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. Llm as dataset analyst: Subpopulation structure discovery with large language model. *arXiv preprint arXiv:2405.02363*, 2024. 3, 4

[40] Hansa Meghwani. Enhancing retrieval performance: An ensemble approach for hard negative mining. *arXiv preprint arXiv:2411.02404*, 2024. 1

[41] Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'llava: Your personalized language and vision assistant. *arXiv preprint arXiv:2406.09400*, 2024. 1, 3, 7

[42] Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize diagnose and optimize: Towards fine-grained vision-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13279–13288, 2024. 3

[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. 6, 1

[44] Hamed Rahimi, Adil Bahaj, Mouad Abrini, Mahdi Khoramshahi, Mounir Ghogho, and Mohamed Chetouani. User-vlm 360: Personalized vision language models with user-aware tuning for social human-robot interactions. *arXiv preprint arXiv:2502.10636*, 2025. 3

[45] Fakhitah Ridzuan and Wan Mohd Nazmee Wan Zainon. A review on data quality dimensions for big data. *Procedia Computer Science*, 234:341–348, 2024. 2

[46] 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. 1

[47] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 2

[48] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. *arXiv preprint arXiv:2308.00755*, 2023. 2

[49] Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 769–778, 2023. 3

[50] Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. Bob: Bert over bert for training persona-based dialogue models from limited personalized data. *arXiv preprint arXiv:2106.06169*, 2021. 1

[51] Xingzhe Su, Wenwen Qiang, Zeen Song, Hang Gao, Fengge Wu, and Changwen Zheng. Manifold-guided sampling in diffusion models for unbiased image generation. *arXiv preprint arXiv:2307.08199*, 9, 2023. 2

[52] Shijian Wang, Linxin Song, Ryotaro Shimizu, Masayuki Goto, and Hanqian wu. Attributed synthetic data generation for zero-shot image classification. In *Synthetic Data for Computer Vision Workshop @ CVPR 2024*, 2024. 3

[53] Shijian Wang, Linxin Song, Jieyu Zhang, Ryotaro Shimizu, Ao Luo, Li Yao, Cunjian Chen, Julian McAuley, and Hanqian Wu. Template matters: Understanding the role of instruction templates in multimodal language model evaluation and training. *arXiv preprint arXiv:2412.08307*, 2024. 3

[54] Weizhi Wang, Khalil Mrini, Linjie Yang, Sateesh Kumar, Yu Tian, Xifeng Yan, and Heng Wang. Finetuned multimodal language models are high-quality image-text data filters. *arXiv preprint arXiv:2403.02677*, 2024. 2

[55] Charles Welch, Chenxi Gu, Jonathan K Kummerfeld, Veronica Perez-Rosas, and Rada Mihalcea. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1742–1752, 2022. 1

[56] Lihe Yang, Xiaogang Xu, Bingyi Kang, Yinghuan Shi, and Hengshuang Zhao. Freemask: Synthetic images with dense annotations make stronger segmentation models. *Advances in Neural Information Processing Systems*, 36:18659–18675, 2023. 3

[57] Chao Yi, Yuhang He, De-Chuan Zhan, and Han-Jia Ye. Bridge the modality and capability gaps in vision-language model selection. *Advances in Neural Information Processing Systems*, 37:34429–34452, 2025. 2

[58] Xingyi You, Xiaohu Zhao, Yue Wang, and Weiqing Sun. Generation of face privacy-protected images based on the diffusion model. *Entropy*, 26(6):479, 2024. 2

[59] Zhixin Zhang, Yiyuan Zhang, Xiaohan Ding, and Xiangyu Yue. Vision search assistant: Empower vision-language models as multimodal search engines. *arXiv preprint arXiv:2410.21220*, 2024. 1

Concept-as-Tree: Synthetic Data is All You Need for VLM Personalization

Supplementary Material

6. More Implementation Details

Dataset. We utilize datasets from Yo'LLaVA [41] and MyVLM [2]. Yo'LLaVA consists of 40 categories of objects, buildings and people, with 4 to 10 images available per category for training and validation. In contrast, MyVLM encompasses 29 object categories, each containing 7 to 17 images, with 4 images designated for training and the remainder for validation. Additionally, we incorporate the Single concept portion of MC-LLaVA [4] dataset, which is a challenging dataset meticulously constructed from movies, featuring textual descriptions generated with the assistance of GPT-4. MC-LLaVA includes 50 scenarios that encompass various characters and objects, amounting to a total of 118 concepts.

Model Settings. For all LLMs and VLMs involved in the CaT framework, we utilized GPT-4o [1]. After obtaining the concept tree, we synthesize positive and negative samples using the DreamBooth [46] fine-tuned FLUX.1-dev model [25]. The positive samples required for fine-tuning are derived from 1 to 3 original concept images. Finally, we select high-quality images using the CLIP ViT-L/14 visual encoder [43] with our proposed filtering method based on the PCS score. Regarding the personalization of VLMs, we test four methods: MyVLM [2], Yo'LLaVA [41], MC-LLaVA [4] and RAP-MLLM [19]. All configurations adhered to the original papers, and we use LLaVA-1.5-13B [34] as the VLM backbone for all experiments.

Hyperparatemers. The hyperparameters for fine-tuning the FLUX model can be referenced from DreamBooth [46]. In the filtering module, the patch size used in patch shuffle is set to 14, consistent with the CLIP visual encoder we employ. Additionally, based on the distribution of PCS scores from the four datasets, we set the PCS score thresholds for synthesized positive and hard negative samples to 0.3 and 0.1, respectively. For easy negative samples, which do not require CS information, we merely employed conventional text-to-image similarity [43] filtering to ensure that the images matched the prompts.

7. Visualization of Concept Trees

We provide detailed visualizations of Concept Trees, covering three categories: humans in Figs. 9 to 16, pets in Figs. 17 to 20, and objects in Figs. 21 to 24. For each concept, we first obtain its concept tree using the CaT. Subsequently, we modify the original tree through three types of tree editing operations: add, modify, and remove. After each operation, a new concept tree is generated, which is then used to synthesize diverse personalized images.

8. CaT Prompt

We provide all the prompts used in the CaT framework. Table 5 includes the three steps for concept tree generation: obtaining the reference image description, performing batch summarization to generate the initial concept tree, and then refining it through self-refinement to obtain the final, well-developed concept tree. Table 6 presents the initialization prompts for the simple negative sample concept tree. Finally, Table 7 provides the three editing operations for the tree—addition, removal, and modification—as well as the prompts used for synthesizing text for image generation based on the concept tree.

9. The Use of PCS Score Filter

We compare our proposed filtering strategy based on the PCS score with the cosine similarity-based filtering strategy that relies on image features. The results presented in Tab. 4 show that the performance of synthetic data without any filtering is even worse than the original Yo'LLaVA [41] baseline. This is because the images synthesized by the generative model may contain low-quality samples that are blurry, unrealistic, or unnatural, which can interfere with the model's training process. After applying the cosine similarity-based filtering using image features, some low-quality images or those inconsistent with the style of the reference images are filtered out, leading to improved model performance. This demonstrates the necessity of a filtering strategy for synthetic data. However, relying solely on conventional image similarity filtering may fail to remove images that achieve high similarity scores based solely on background, style, or other CA information. Our proposed PCS score-based filtering strategy addresses this limitation by perturbing the synthetic images and calculating the difference in similarity scores before and after perturbation. This approach effectively eliminates the interference of CA information, ultimately selecting images with a higher proportion of CS information. This enables further improvement in the model's personalization performance in all tasks.

Filter Strategy	Rec	Choice-V	Choice-T	VQA	Caption
None	0.835	0.793	0.691	0.633	0.697
Cosine Similarity	0.862	0.816	0.716	0.652	0.726
PCS Score	**0.885**	**0.845**	**0.738**	**0.682**	**0.754**

Table 4. **Ablation of different filtering methods.** The filtering method based solely on cosine similarity cannot exclude images lacking CS information. Our method can ensure high-quality samples, improving model performance.

10. Human Evaluation

In addition to the various experiments in the main text that demonstrate the diversity and effectiveness of our synthesized positive and hard negative samples, we conducted a site-by-site evaluation experiment comparing synthetic data with original data. The evaluators were college student volunteers recruited by us. Our evaluation objective was to demonstrate that the synthesized positive and hard negative samples have high quality to the original concept images.

The original data were sourced from MyVLM, Yo'llaVA, and MC-LLaVA. For each concept, we selected 1 to 3 positive samples as reference images. We then conducted site-by-site evaluation experiments for both positive and negative samples. For the original positive samples (excluding those selected as reference images) and synthesized positive samples, we randomly paired one from each group and compared them sequentially with all reference images of the corresponding concept. Evaluators were required to choose the image with higher similarity to the reference image and assign it one point. The winner of each comparison was determined based on the cumulative scores across all reference images. The same evaluation method was applied to the original retrieved complex negative samples and synthesized hard negative samples. Note that if the number of reference images was even, there was a possibility of a tie.

The results of the two sets of experiments are presented in Fig. 8. We find that the synthesized positive samples perform comparably to the original positive samples. Meanwhile, the synthesized hard negative samples achieve significantly higher scores than the retrieved hard negative samples. This result clearly demonstrates that our synthesized data can effectively support the personalization training process of vision language models.

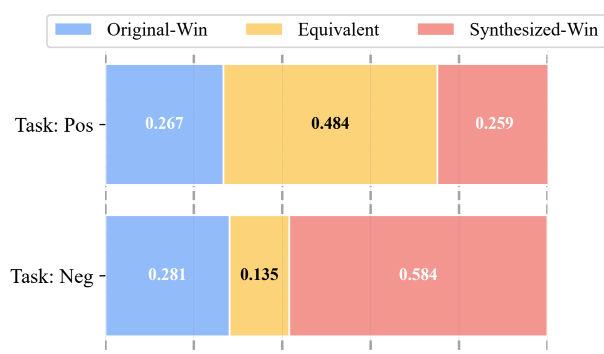

Figure 8. Human evaluation of similarity between original and synthesized images compared to reference images: The synthesized positive samples were found to be on par with the original positive samples, while the synthesized hard negative samples demonstrated significantly higher similarity compared to the originally retrieved hard negative samples.

11. Additionally Related Work

Data Quality and Selection. The development of LLMs and VLMs relies heavily on data [10, 17, 28]. The well-known principle is "garbage in, garbage out." High-quality data can significantly enhance model performance [45]. To ensure data quality, data selection and filtering are common practices [9, 57]. Primarily, there are rule-based [5] and model-based methods [13]. Although rule-based methods are simple and heuristic, they are effective. Methods based on LLMs are widely used in data selection. In multimodal scenarios, Clip-filter is one of the most commonly used methods and VLMs are also employed as data filters for self-selection and improvement [22, 54]. However, since the $\langle$sks$\rangle$ we need to evaluate is not in the vocabulary of the language model, retraining the VLM for each new word is cumbersome. Thus, we employ a simple yet effective method by perturbing synthetic images. We evaluate the concept information contained in the original image by calculating the similarity between the image and the reference image before and after perturbation, and apply a predetermined threshold for selection and filtering.

12. Limitation and Future Work

Although our method demonstrates effectiveness in generating concept-centric data, several limitations warrant consideration, along with potential directions for future work.

The first limitation stems from our framework's current inability to handle scenarios involving multiple concepts, primarily due to its dependence on FLUX [25] for image generation. Nevertheless, we believe the inherent scalability of our tree structure paradigm offers promising opportunities for extension through forest-based architectures. Additionally, Future work could explore the integration of enhanced diffusion models with stronger guidance mechanisms, such as layout or box conditions, which have shown potential in better capturing multiple concepts information.

A second limitation arises from potential biases inherited from the LAION-5B [47] dataset used for training FLUX. However, this issue is actively being addressed in ongoing research. Recent studies [16, 48, 51] have made significant progress in developing techniques to enhance fairness and reduce biases in generative models, providing valuable directions for future improvements in our framework.

Finally, the generated concept-centric images may bring data privacy and ethical concerns. This challenge has garnered increasing attention in the research community, with recent works [20, 35, 58] making notable advances in privacy-preserving generation techniques. We anticipate that continued research in GenAI safety will yield effective solutions to mitigate these concerns, enabling more responsible use of synthetic data generation methods.

Image Description.
Please identify the primary object class name in the image and describe the image in detail with the class name.

Batch Summarization.
Task: I want to classify and organize captions for some images
Requirement: I will provide a batch of captions and the main object they revolve around. Please describe the attributes related to the subject from multiple dimensions based on the subtitles. The final result is output as {tree_example}
Input: main object: {class_name}; captions: {captions}
Output:

Self-Refinement.
Task: I'm trying to classify the following captions into a classification criteria. However, it seems the current criteria fails to capture certain details that would help differentiate this caption.
Requirement: We are unable to classify these captions using the provided criteria due to one of the following reasons.
1. LLM Hallucination: If you believe the current criteria is reasonable, and the sample can be classified under one of them, the current failure may be due to LLM hallucination. If the majority of classifications are correct and only a small portion of the results appears highly unreasonable, it is likely due to hallucination. In this case, please do nothing. Answer format: {{"hallucination": []}}
2. Attribute Redundancy: If there are redundant attributes in the criteria, please identify and replace them with a single, unified keyword that represents all the redundant attributes. Answer format: {{"redundant": ["unified keyword"]}} (Only include a single keyword that replaces all the redundant or duplicated attributes).
3. Missing Attributes: If some important attributes are missing and need to be added to the criteria to accurately classify the caption, please suggest one attribute to add. Answer format: {{"missing": ["keyword"]}}
After obtaining the content in the curly braces (for example, "hallucination": [], "redundant": [...], "missing": [...]), please do the following processing according to the situation.
1. For hallucination, it means that the previous LLM judgment is wrong, and there is no need to modify the current tree;
2. For redundancy, please merge them and put the final attribute into a dimension you think is appropriate (if necessary, you can delete the redundant dimension);
3. For missing, please find the key information in the content of captions and extract the missing attributes, and add them to the appropriate dimension (if necessary, you can create a new dimension)
Input: caption: {captions}; current concept tree: {concept_tree}
Output:

Table 5. The three steps of concept tree synthesis.

Easy Negative Sample Concept Tree Generation.
Task: I want to modify the visual definition of a class to synthesize a new visual definition of the class
Requirement: I will input the visual definition tree of a class. Please modify its class name to obtain a new class, and generate some visual dimensions and corresponding attributes that match the new class according to the tree format. Finally, output the visual definition tree in the original tree format to return the new visual definition tree
Input: concept tree: {concept_tree}
Output:

Table 6. Modify the root node of the original concept tree to obtain an easy negative sample concept tree.

Tree Operation—Add.
Task: I want to modify the visual definition tree of a class
Requirement: I will input a visual definition tree for a class, please randomly add {num} of its dimensions to obtain a new visual definition tree, note that it conforms to the natural definition of this class. Only output the new visual definition tree.
Input: concept tree: {concept_tree}
Output:

Tree Operation—Remove.
Task: I want to modify the visual definition tree of a class
Requirement: I will input a visual definition tree for a class, please randomly delete {num} of its dimensions to obtain a new visual definition tree, note that it conforms to the natural definition of this class. Only output the new visual definition tree.
Input: concept tree: {concept_tree}
Output:

Tree Operation—Modify.
Task: I want to modify the visual definition tree of a class
Requirement: I will input a visual definition tree for a class, please randomly modify {num} of its dimensions to obtain a new visual definition tree, it means deleting some visual dimensions and adding new visual dimensions, not just modifying attributes. Only output the new visual definition tree.
Input: concept tree: {concept_tree}
Output:

Image Prompt Generation.
Task: I want to generate different prompts based on the visual definition of a category to prompt the diffusion model to synthesize images.
Requirement: I will provide a visual definition of a category, including the category name and its multiple visual dimensions and attributes. You can combine the attributes of these dimensions to obtain prompts that match the real scene, such as "a photo of attribute of dimension1, attribute of dimension2, ..., classname". Please generate at least 100 prompts, ensuring that the similarity of each prompt is 0, and only output prompts line by line.
Input: class category: {category}; concept tree: {concept_tree}
Output:

Table 7. Three editing operations of a concept tree and using a concept tree to synthesize diverse image prompts.

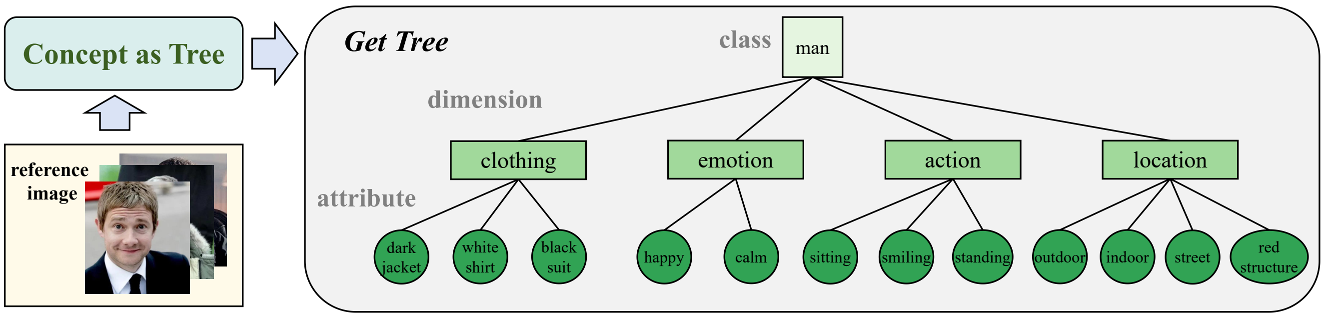

Figure 9. Using the CaT framework, we generate a concept tree for the "man" concept. The CaT framework can accurately extract the visual information associated with this concept, such as clothing and location.

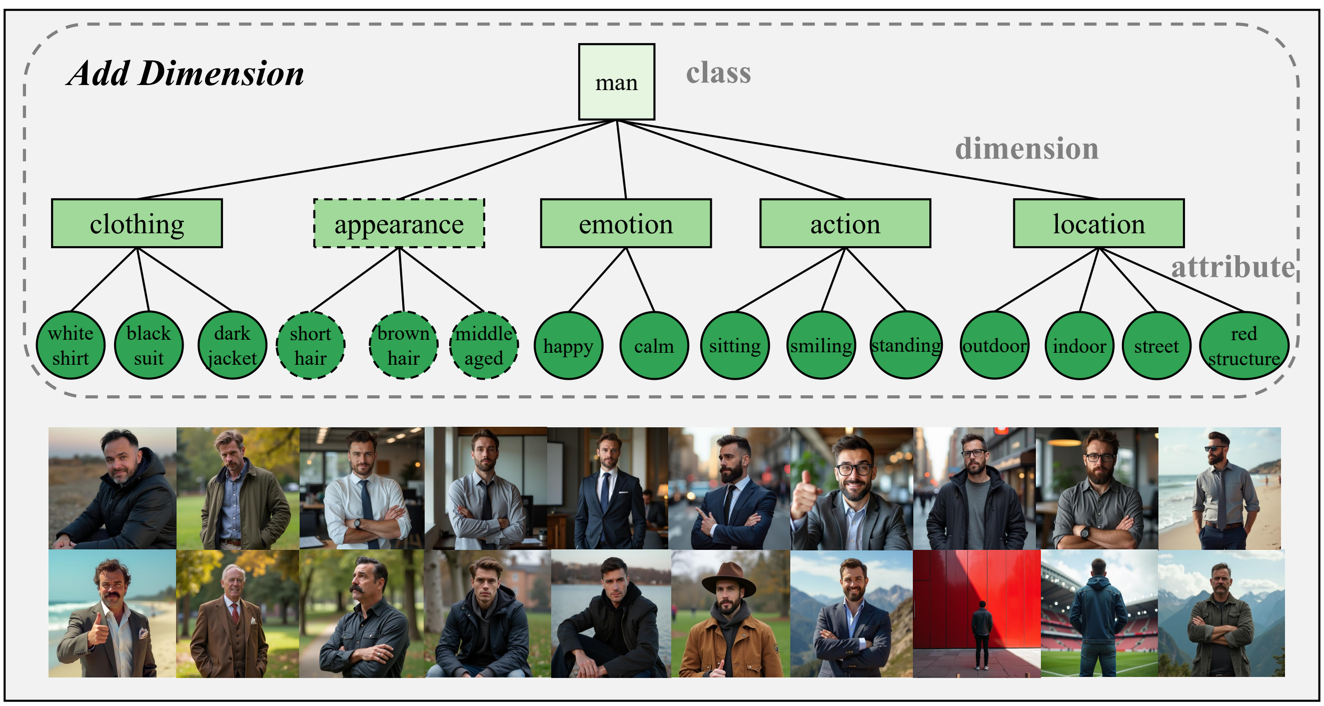

Figure 10. Adding an "appearance" dimension to the original concept tree results in significant changes in people's appearance, such as hair length and color, which are combined with the original attributes to create a variety of styles of images

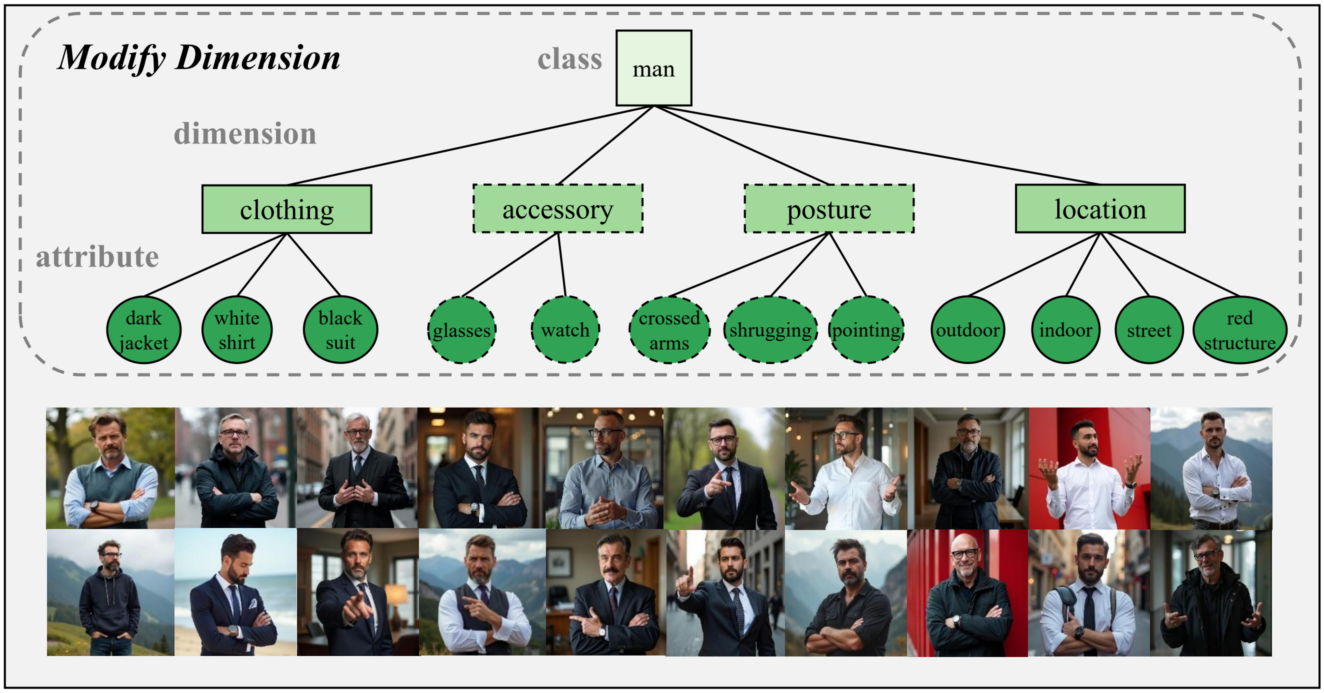

Figure 11. Change the “emotion” and “action” dimensions in the original concept tree to the “accessory” and “poster” dimensions, allowing for various combinations of male accessories, standing posture, and movements.

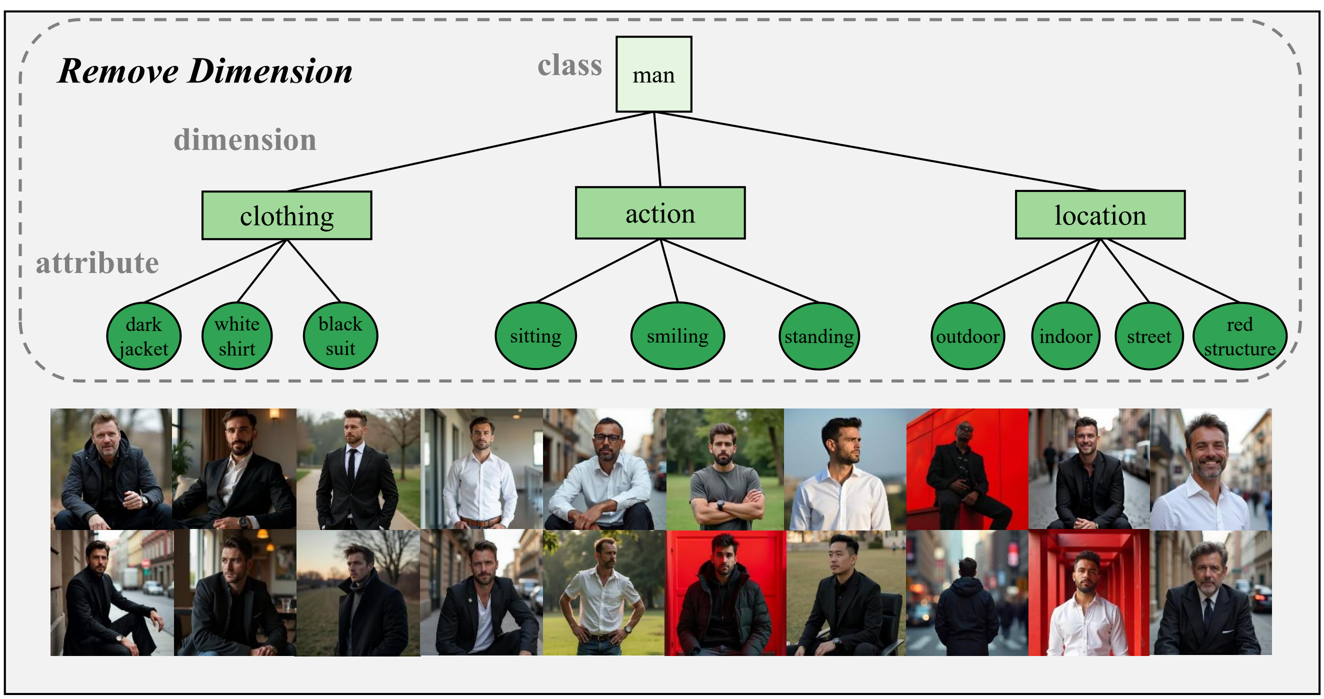

Figure 12. Removing the “emotion” dimension from the original concept tree, the previously happy, excited, and varied expressions almost uniformly turn into silence, which reduces the diversity of images.

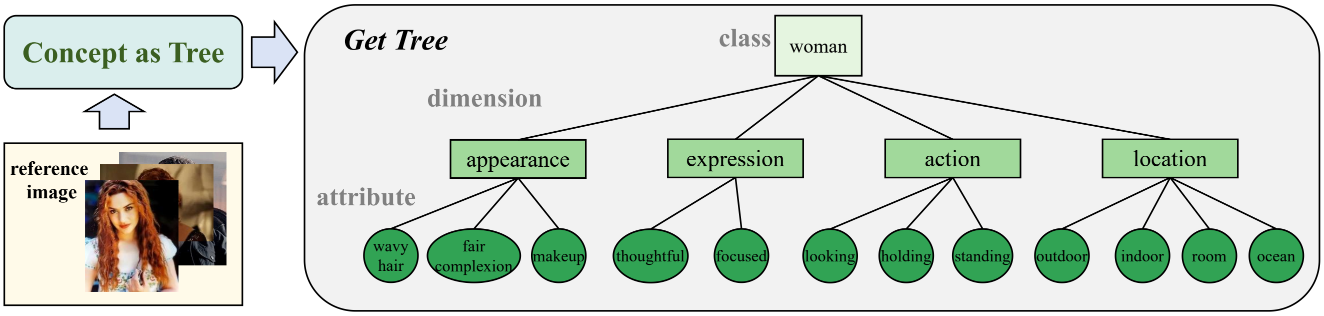

Figure 13. Using the CaT framework, we generate a concept tree for the "woman" concept. The CaT framework can accurately extract the visual information associated with this concept, such as expression and location.

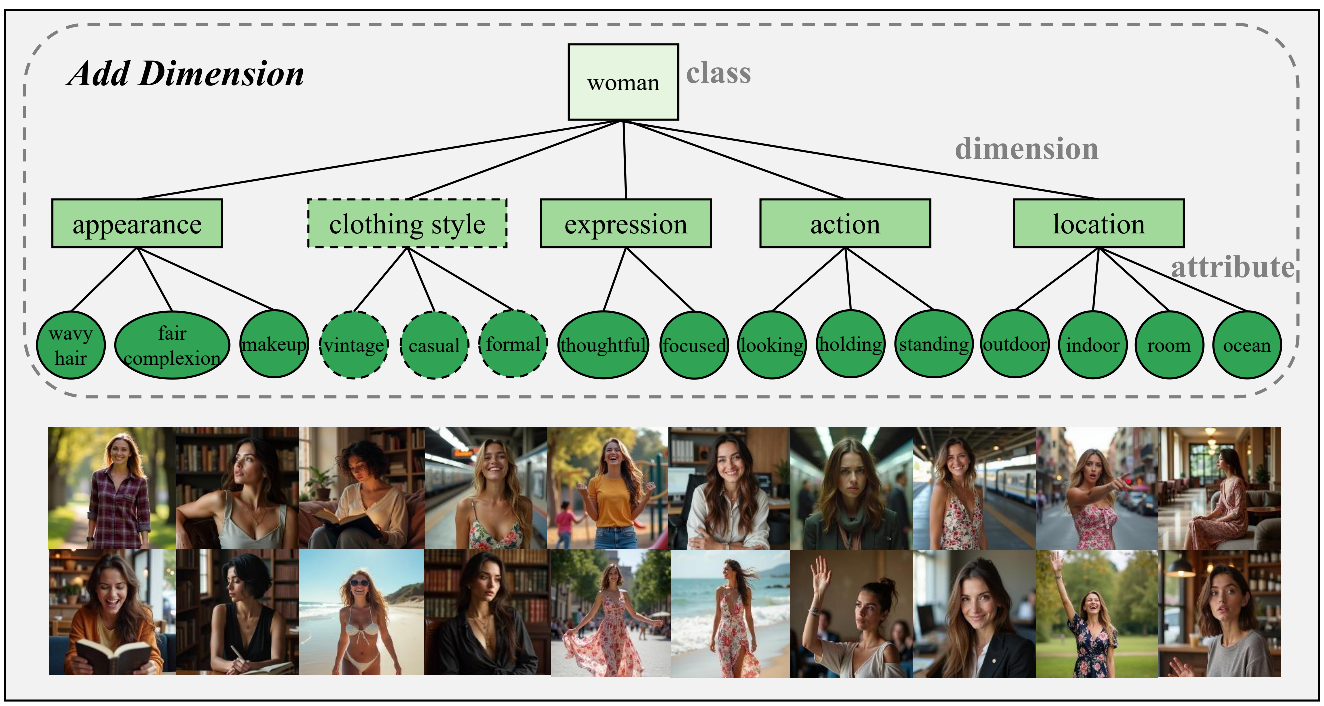

Figure 14. Adding a "clothing style" dimension to the original concept tree results in a significant change in woman's attire, bringing multiple styles to the images.

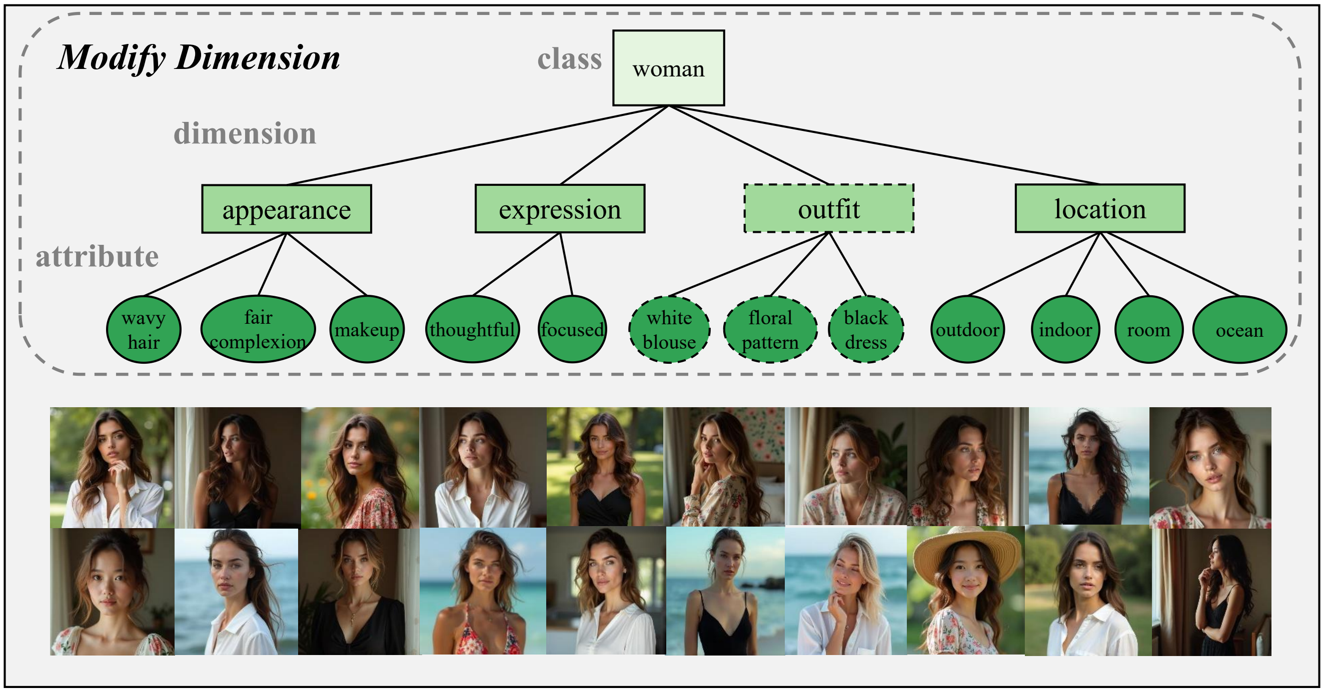

Figure 15. Change the “action” dimension in the original concept tree to the “outfit” dimension, so that woman’s clothing is no longer just one or two types of reference pictures

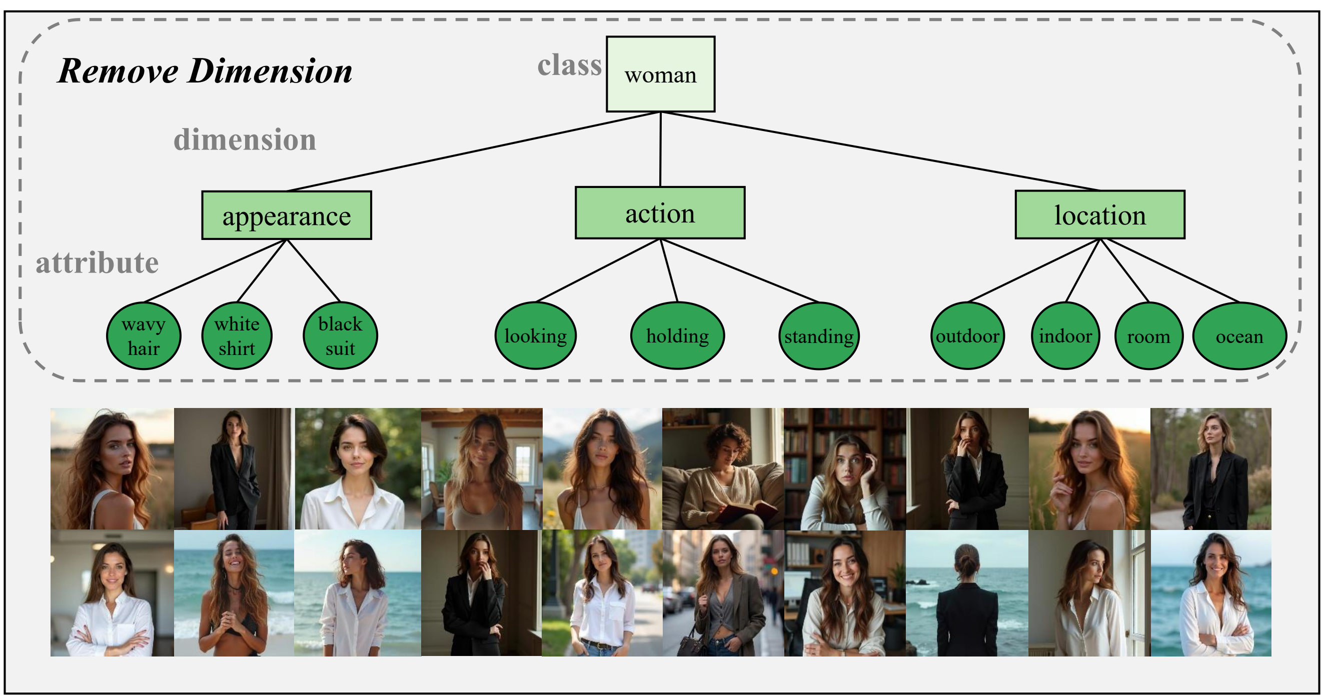

Figure 16. Removing the “expression” dimension from the original concept tree leaves almost only a calm expression on the woman’s face, which reduces the diversity of images

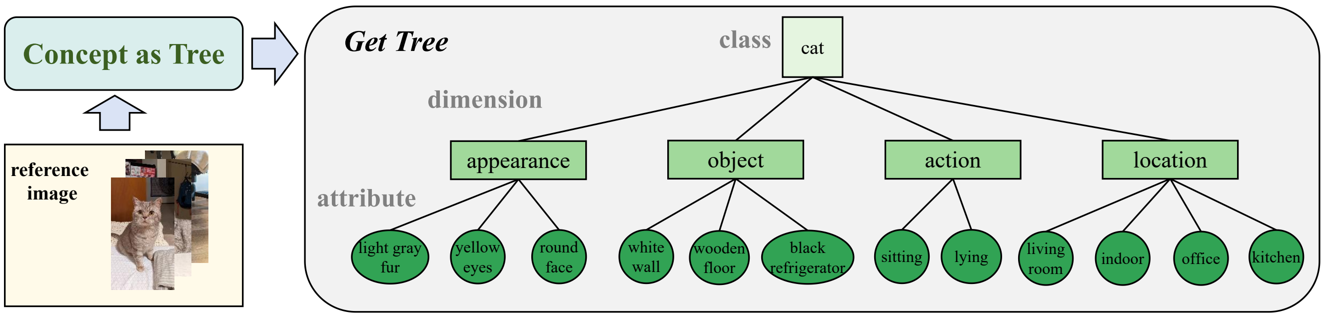

Figure 17. Using the CaT framework, we generate a concept tree for the "cat" concept. The CaT framework can accurately extract the visual information associated with this concept, such as appearance and action.

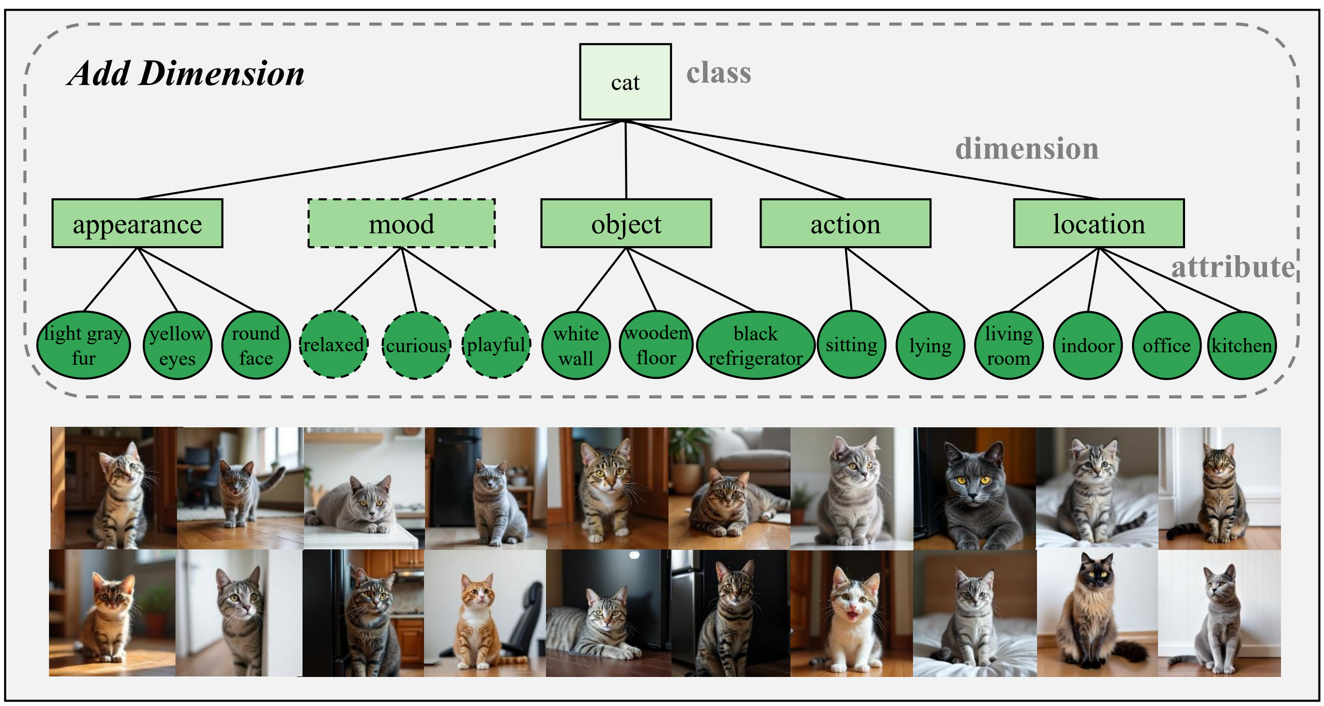

Figure 18. Adding a "mood" dimension to the original concept tree significantly changes the expression and posture of the cat, making it more interesting compared to the previously silent cat in the reference image.

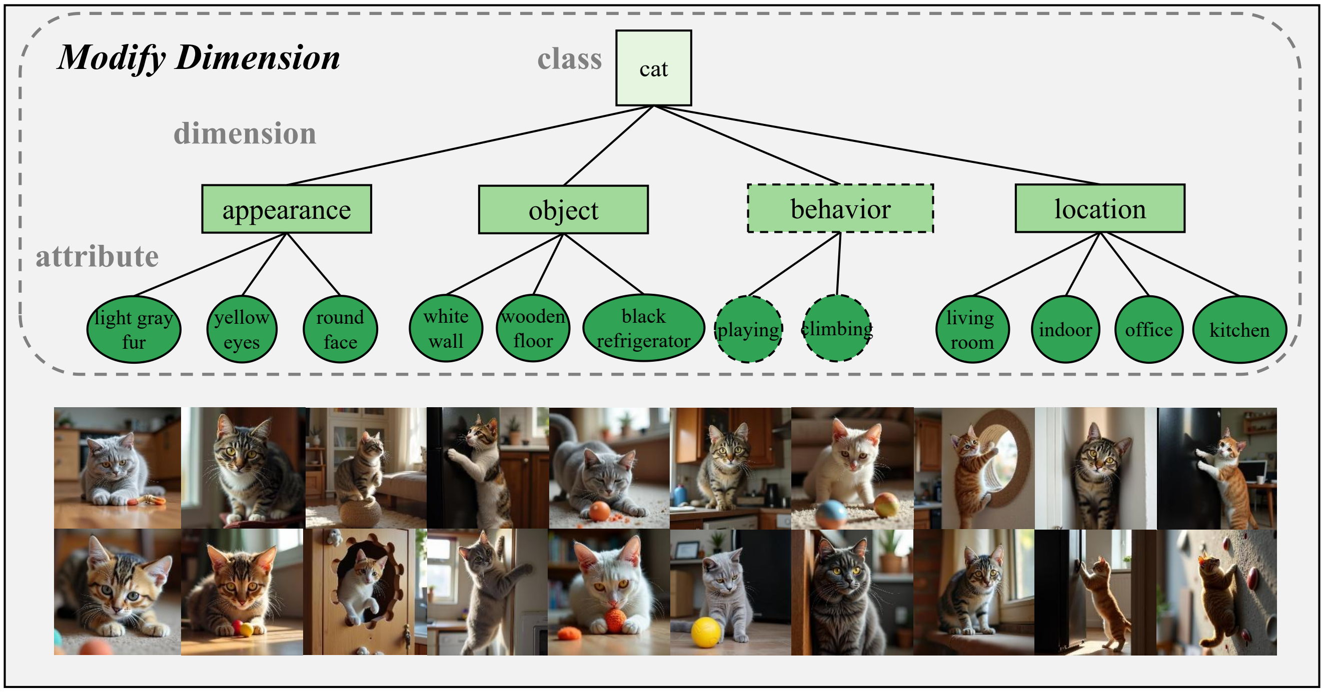

Figure 19. By changing the "action" dimension in the original concept tree to the "behavior" dimension, cats are no longer just squatting or lying down, bringing about a variety of posture changes.

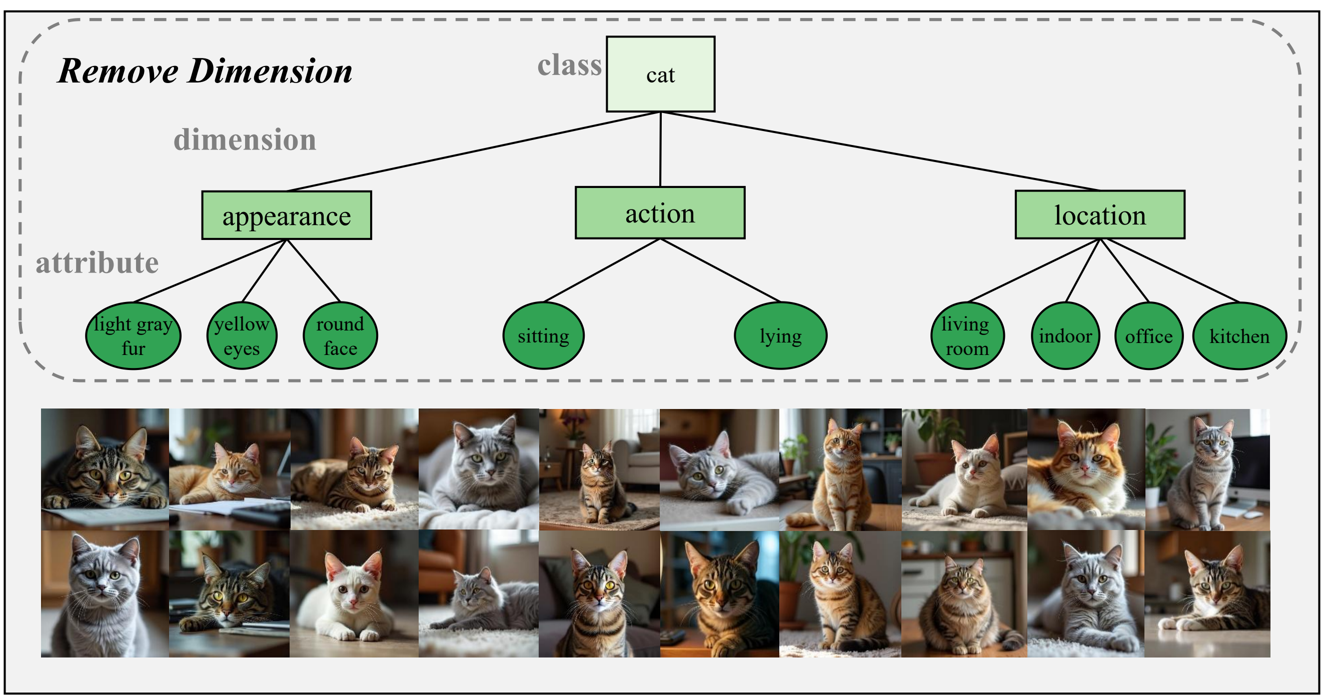

Figure 20. Removing the "object" dimension from the original concept tree makes the objects around the cat become monotonous. The diversity of images has significantly decreased.

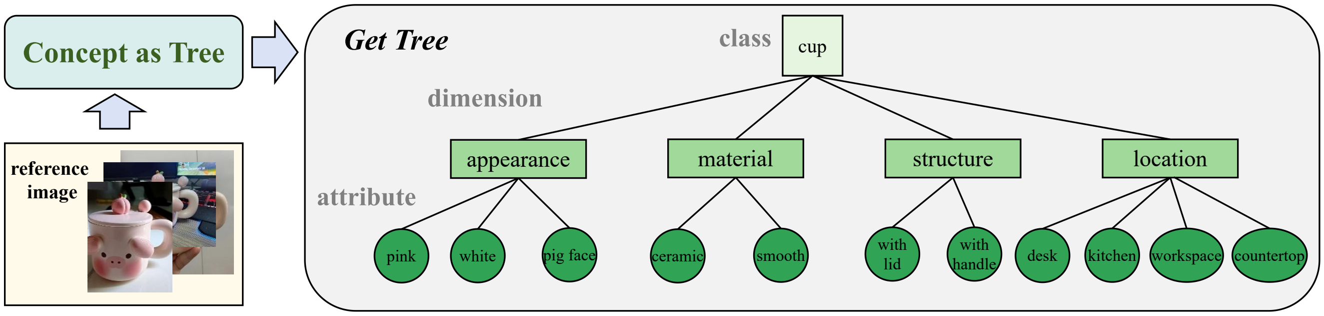

Figure 21. Using the CaT framework, we generate a concept tree for the "cup" concept. The CaT framework can accurately extract the visual information associated with this concept, such as material and structure.

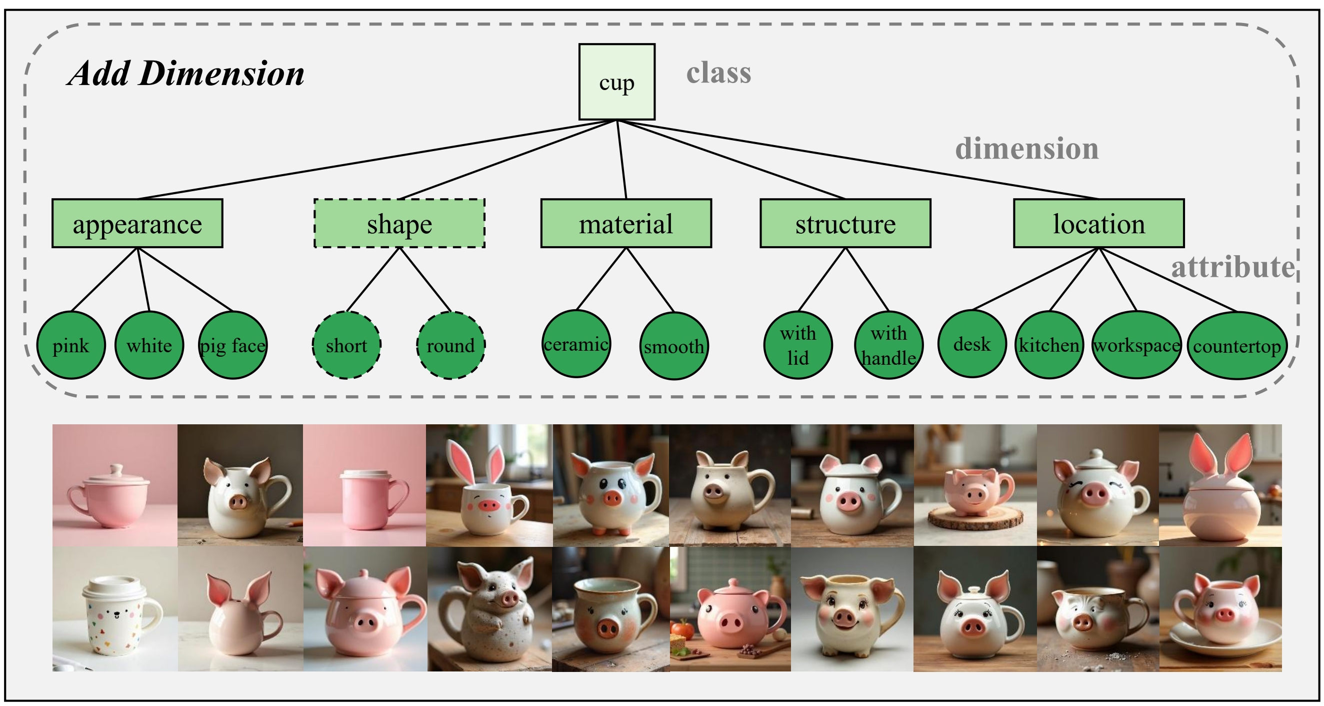

Figure 22. Adding a "shape" dimension to the original concept tree, the height and weight of the cups are no longer consistent, bringing about an increase in diversity.

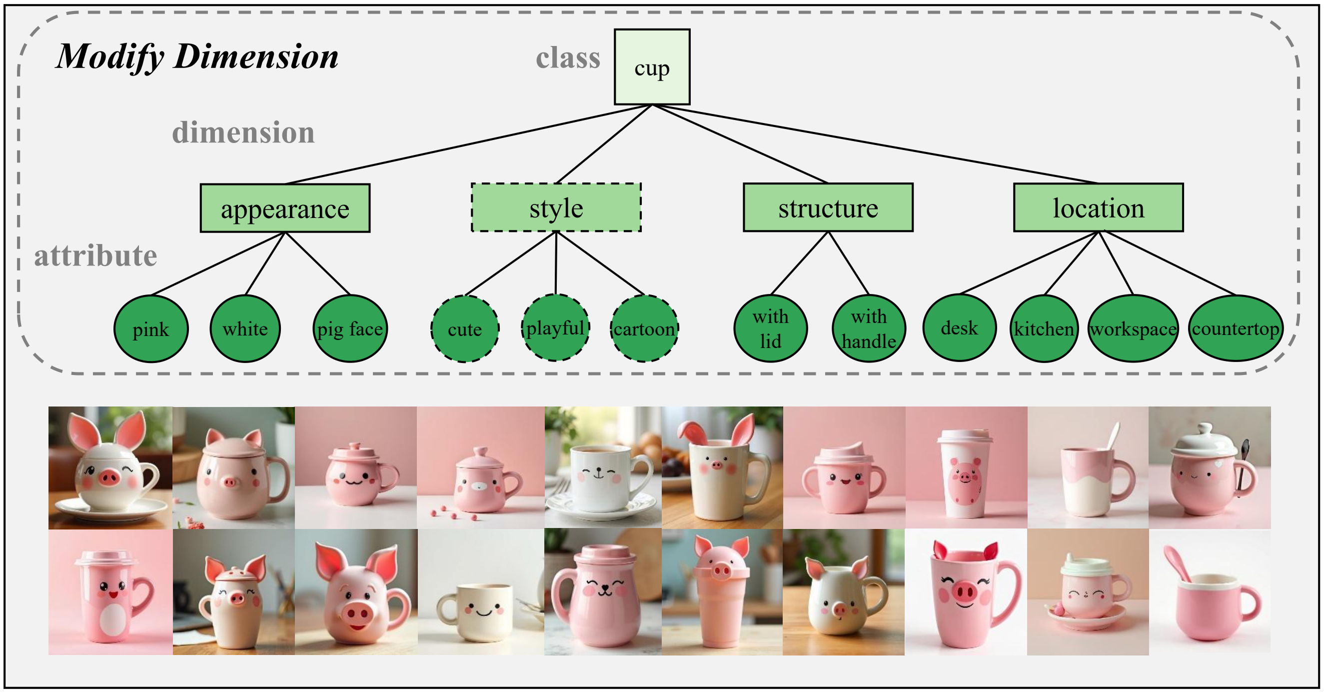

Figure 23. Changing the “material” dimension in the original concept tree to the “style” dimension makes the expression on the cup more vivid and adorable.

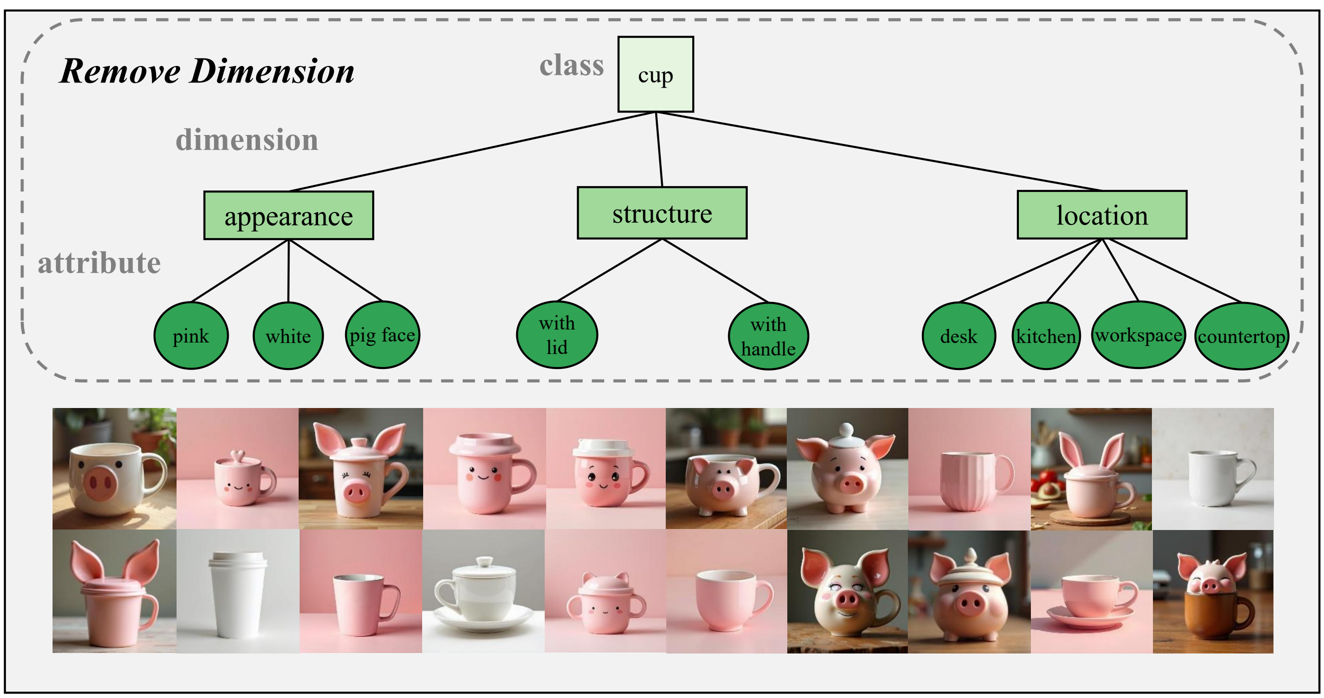

Figure 24. After removing the “material” dimension from the original concept tree, most of the generated images only have one material, smoothness. The diversity of images has significantly decreased.