Concept-as-Tree: A Controllable Synthetic Data Framework Makes Stronger Personalized VLMs

Ruichuan An[1*], Kai Zeng[1*], Ming Lu[1], Sihan Yang[1],
Renrui Zhang[2], Huitong Ji[1], Hao Liang[1], and Wentao Zhang[1,3,4†]

[1]Peking University, [2]CUHK MMLab, [3]Zhongguancun Academy,
[4]Beijing Key Laboratory of Data Intelligence and Security (Peking University)
arctanxarc@gmail.com

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 existing techniques. To reveal the relationship between sample and model performance, we systematically investigate the amount and diversity impact of positive and negative samples (easy and hard) 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, and can be easily extended to multi-concept scenarios. 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 capabilities of VLMs across personalization benchmarks. To the best of our knowledge, this work is the first controllable synthetic data pipeline for VLM personalization. The code is released at https://github.com/zengkaiya/CaT.

Keywords: Vision-Language Model · Personalization · Synthetic Data

1 Introduction

Vision-Language Models (VLMs) [6, 28, 31] have produced impressive results across various tasks, showcasing their potential as AI assistants [11,21,48,53,54]. Despite their success, VLMs still struggle to generate personalized responses,

⋆ Equal Contribution.
† Corresponding Author.

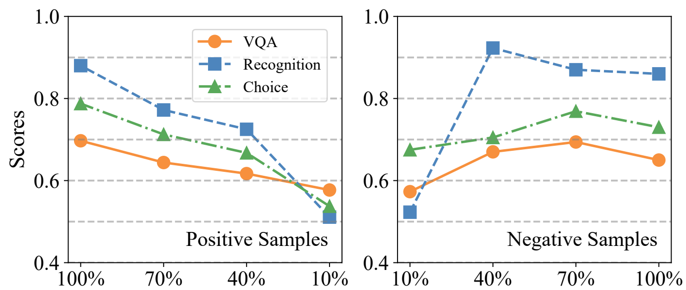

Fig. 1: The performance of Yo'LLaVA on Yo'LLaVA Dataset. (Left) Various tasks show a noticeable decrease in performance when the number of positive samples is limited. (Right) As the number of retrieved negative samples increases, the performance of various tasks does not improve consistently. This might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

such as including user-specific identifiers like ⟨Bob⟩ or ⟨Lina⟩ in the conversation. Several recent studies [1, 2, 19, 38] have explored the personalization of VLMs to address this challenge and seamlessly incorporate VLMs into everyone's daily life. Among these, the test-time fine-tuning paradigm exemplified by Yo'LLaVA [38] demonstrates effectiveness. Its training set includes positive, easy-negative, and hard-negative samples, with detailed examples provided in the appendix.

Although test-time fine-tuning has shown its effectiveness, its success largely hinges on the availability of both positive and negative samples during the process. In real-world applications [8, 43, 47], 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, the acquisition of negative samples often relies on image retrieval [37, 38], a unpredictable process that complicates the assurance of their quality. Preliminary experiments, as illustrated in Fig. 1, corroborate the limitations. Given these challenges, a natural idea is to use synthetic data to provide controllable supplementation for both positive and negative samples. While this approach may help mitigate the aforementioned issues, three significant challenges remain to be addressed:

1. The roles for positive and negative samples in VLM personalization 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.
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 gener-

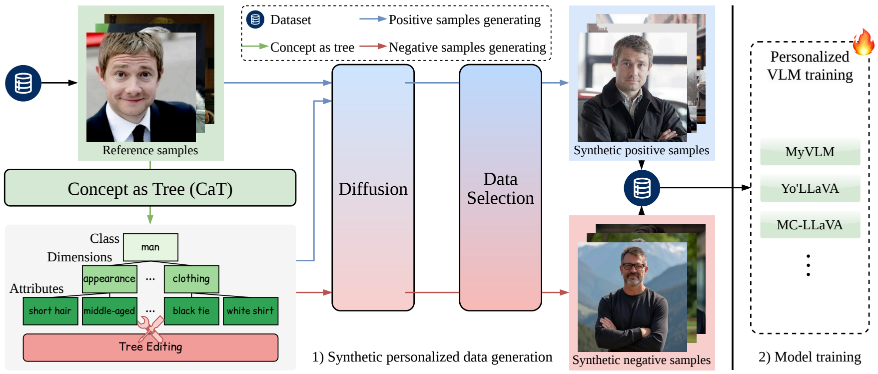

Fig. 2: Overview of unified and controllable data synthesis pipeline and personalized model training. After a systematic analysis of positive and negative samples, we utilize LLM and VLM to construct the concept tree and edit it to achieve controllable generation. We then propose a well-designed data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

ally 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 LLMs and VLMs, we can automate the construction of trees for each concept. To synthesize positive samples, we directly employ the root node as a prompt for diffusion model. Easy negative samples are generated by altering the root node information and constructing a tree for controllable data generation. Hard negative samples are produced according to concept tree whose leaf nodes are strategically adjusted. Moreover, as the number of concepts increases, we naturally merge the trees into forests according to the specific scenarios [3] (e.g., two concepts). Because of the designed tree structure, our framework supports all types of personalized data. Meanwhile, controlling diversity is equivalent to varying the number and types of editing operations performed on the tree in our framework.

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-

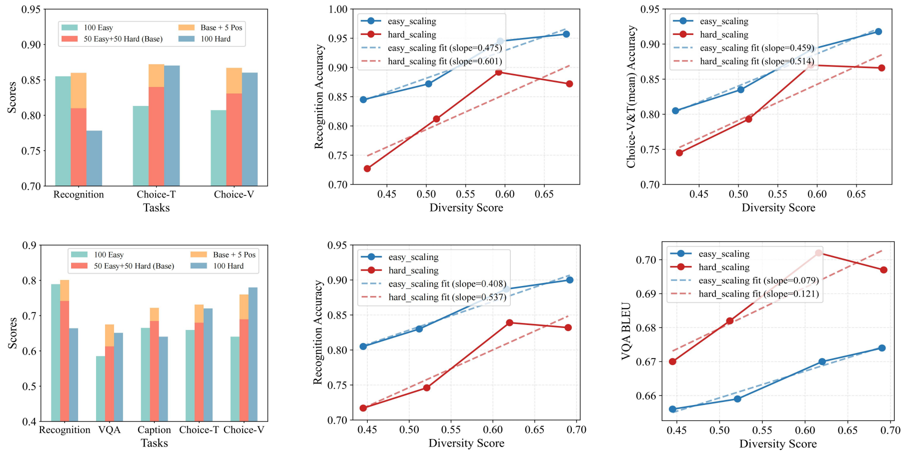

Fig. 3: Explore the role of positive and negative samples in personalization and their demand for diversity. (Left) We evaluate the effect of halving positive and negative samples. Positive samples generally improve performance, while easy and hard negatives show task-specific benefits. (Middle & Right) With fixed sample counts, we vary the diversity of easy and hard negatives. Hard negatives are more sensitive to diversity, and excessive diversity can degrade performance. Diversity scores are computed by clustering retrieved negatives via K-means and measuring distances to cluster centroids. The first line shows results on the Yo'LLaVA dataset; the second line illustrates results on the MC-LLaVa dataset.

agnostic features that possess general characteristics [33]. Therefore, we apply some 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 the 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 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) [25,26,32] have exhibited remarkable capabilities across diverse domains, such as data mining [36], fine-grained understanding [30], and visual question answering [7]. To seamlessly integrate these models into daily life, there has been growing interest in VLM personalization [3, 4, 16, 20, 50]. This task is first introduced by MyVLM [1], employing an additional module strategy for improved injection of concept information into VLMs. YoLLaVA [38] and MC-LLaVA [3] 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 training samples. However, users often employ a limited number of concept images (i.e., 1 to 3), 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 and propose a controllable data synthesis pipeline, thus overcoming challenges.

Learning from Synthetic Data. The use of synthetic data [18,24,35] has been studied across various computer vision tasks, such as classification [42], segmentation [49], 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, 27, 34, 39]. 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 [14]. To better support the learning of VLMs, direct generation of text-image pairs is pursued [29]. However, existing generation methods only consider class information [44]. These methods rely on simplistic prompt templates, attributes, and class information, either manually defined or LLM-generated [12, 45]. 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 generated samples.

Data Quality and Selection. The development of LLMs and VLMs relies heavily on data [10, 17, 52]. The well-known principle is “garbage in, garbage out." High-quality data can significantly enhance model performance [41]. To ensure data quality, data selection and filtering are common and indispensable practices [9,51]. Primarily, there are rule-based [5] and model-based methods [13]. 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, 46]. 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.

3 Method

We present a comprehensive data synthesis pipeline to address the challenge of personalizing Vision-Language Models (VLM). This pipeline views the concept information as a tree structure and controls the generation of positive and negative samples via 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 part 1, we systematically study the impact of positive and negative samples on VLM personalization tasks. Building on this, we introduce the CaT framework and related tree operations in part 2. Furthermore, to ensure the quality of the synthetic samples, we propose a filtering strategy based on the PCS score in part 3.

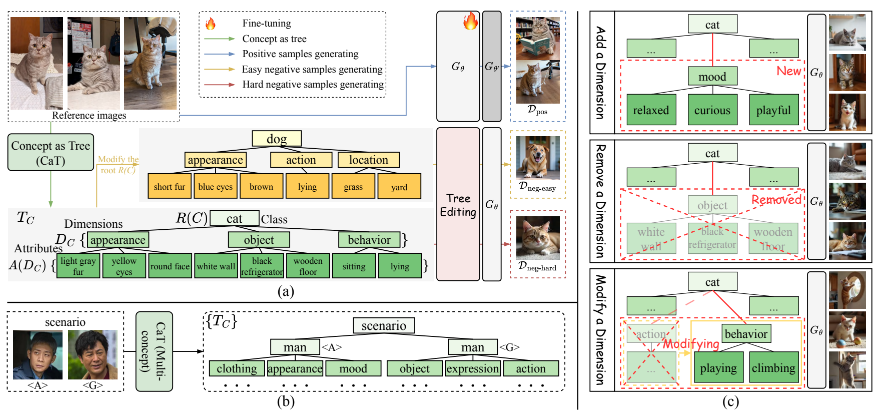

Fig. 4: Concept tree synthesis and editing operations for sample generation. (a) A concept tree T_C is constructed using the CaT framework and reference images. Positive samples are synthesized with a fine-tuned diffusion model guided by the root node $R(C)$. Easy negatives are generated by changing the class of $R(C)$, while diverse hard negatives are created by applying three editing operations to the original tree, using an unfine-tuned model. (b) Multiple concept trees can also be merged to support multi-concept personalization data generation. (c) The three operations are visualized: adding a "mood" dimension alters emotions; removing the "object" simplifies the scene; modifying a dimension changes behavior.

3.1 Impact of Positive and Negative Samples

In this section, we examine the effects of positive and negative samples respectively and their diversity on the personalization of VLM. Based on the positive samples provided by existing datasets 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 samples yields the best performance. Furthermore, we find that easy negative samples are essential. Using only hard negative samples can help the model perform well in conversational tasks, but recognition ability significantly declines. However, recognition tasks are the first step for VLMs acting as personal assistants, representing an essential capability that is crucial. Hence, easy negative samples are irreplaceable due to their diversity and broad sourcing, which offers valuable context for the model's learning process.

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 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:

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

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 SSDLLM [36], 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 pretrained VLM generates textual descriptions 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 $\mathcal{D}_{\text{neg-hard}} = G_\theta(T'_C)$.

Multi-Concept Scenarios. Our CaT framework seamlessly supports multi-concept scenarios by merging different concept trees according to the specific scenario, as illustrated in the Fig. 4(b). We generate training data for multiple concepts using the same approach as that used for single concepts. The scalability of the tree structure allows our framework to adapt flexibly to the continuous addition of new concepts, which is a significant advantage of our CaT framework.

Tree Editing Operations. We define three fundamental tree editing operations $F(R(C), D'_C)$ (see Fig. 4(c)), 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.

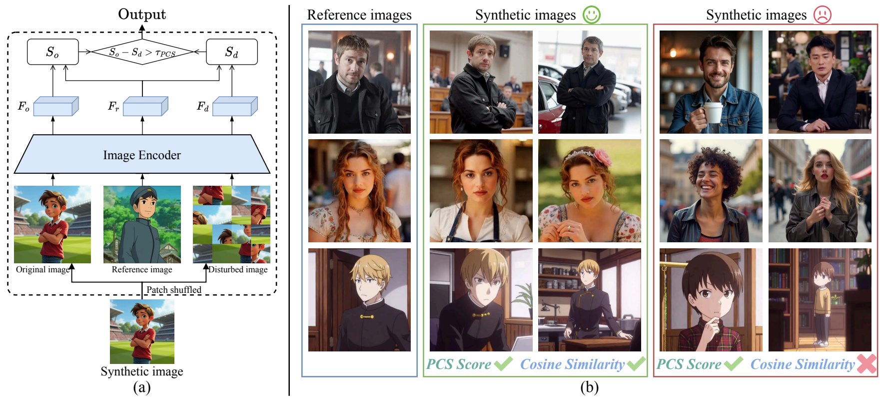

Fig. 5: PCS-based filtering and visualization of high-quality image selection. (a) We apply patch shuffling to synthetic images and extract features F_o, F_d, and F_r from original, disturbed and reference images. Image similarities S_o and S_d are computed to select images with high PCS scores. (b) Images rich in CS information are in green and those rich in CA information are in red. Cosine similarity can not distinguish between two types, but PCS score effectively filters out unqualified ones.

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, while the latter encompasses other elements present in the image, such as background. We consider synthetic data predominantly containing CS information to be regarded as high-quality samples. Conversely, samples with more CA information may be categorized as low-quality samples.

Building upon this, we first mix patches from the reference and original images, subsequently calculating the similarity between the altered images and the reference images using the CLIP model. The difference in similarities before and after perturbation is defined as the Perturbation-based Concept-Specific (PCS) score. Typically, the PCS score of an image is expected to be high due to the alteration of patches, thereby disrupting the original Concept-Specific (CS) information. This disruption causes CLIP to struggle to discern the conceptual relationship between the synthetic image and the reference image, resulting in a significant decrease in similarity between the altered and reference images. However, if the similarity is based on Conceptual-Agnostic (CA) information rather than CS information, the similarity remains relatively unchanged, leading to a low PCS score. Therefore, we claim that the PCS score can be used to determine the information components in the image. A high PCS score signifies more CS information, while a low PCS score indicates 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 [40] image encoder f_θ: $F_I = f_\theta(I)$ and then calculate the cosine similarity S_o and S_d, respectively. The cosine similarity $S_x \in [-1, 1]$ is calculated as:

$$S_x = \frac{F_1 F_2^\top}{\parallel F_1 \parallel \cdot \parallel F_2 \parallel}, \tag{4}$$

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 proposed perturbation-based filtering method, we can effectively identify and filter out images with more CA information using the PCS score, while retaining high-quality images with more CS information for better personalization.

4 Experiments

4.1 Experimental Setup

Implementation 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 generated by CaT. Baseline (Real+Syn) is articulated as comprising 1 to 3 original concept images and several synthesized positive samples, ensuring the same amount of the original 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. 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 the 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. Within the MC-LLaVA framework, despite the use of purely synthesized data, performance still surpasses the baseline, which

Dataset	MC-LLaVA Evaluation					Yo'LLaVA Evaluation			MyVLM Evaluation	
Method/Task	**Rec**	**Choice-V**	**Choice-T**	**VQA**	**Caption**	**Rec**	**Choice-V**	**Choice-T**	**Rec**	**Caption**
GPT-4o+Prompt	0.746	0.888	0.712	0.728	0.836	0.856	0.932	0.897	0.891	0.969
MyVLM(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(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(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	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. Synthetic data improves all methods across three datasets. MC-LLaVA, trained on original positive data combined with all synthetic data, achieves performance comparable to GPT-4o. Choice-T is a text-based task; MyVLM cannot perform it as it requires concept images for embedding loading. Numbers after "+" denote the new SOTA's improvement over the original baseline. Evaluation tasks: Real = Original Data, Syn = Synthetic Data, P = Plus. Evaluation metrics: Rec = Recognition Accuracy, Choice-V = Choice-V Accuracy, Choice-T = Choice-T Accuracy, VQA = BLEU, Caption = Recall. **Higher** values are preferred for all metrics.

is attributed to the 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.

Our CaT framework also supports the synthesis of multi-concept personalized training data. By applying attribute constraints across different concept trees within the same scene, we ensure that the synthesized positive samples do not share overlapping attributes. Experimental results in Tab. 2 show that this significantly enhances the capability of multi-concept personalized models.

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.

Dataset	**MC-LLaVA Evaluation (Multi-Concept)**				
Method/Task	**Rec**	**Choice-V**	**Choice-T**	**VQA**	**Caption**
GPT-4o+Prompt	0.822	0.889	0.680	0.651	0.816
Yo'LLaVA(Real)	0.729	0.602	0.594	0.557	0.611
Yo'LLaVA(Syn)	0.715	0.613	0.579	0.541	0.625
Yo'LLaVA(Real+Syn)	0.732	0.608	0.597	0.569	0.628
Yo'LLaVA(Real+Syn)(P)	0.760	0.634	0.618	0.580	0.647
	+3.1%	**+3.2%**	**+2.4%**	**+2.3%**	**+3.6%**
MC-LLaVA(Real)	0.845	0.905	0.695	0.611	0.763
MC-LLaVA(Syn)	0.832	0.918	0.682	0.598	0.776
MC-LLaVA(Real+Syn)	0.849	0.911	0.698	0.624	0.780
MC-LLaVA(Real+Syn)(P)	0.881	0.933	0.714	0.636	0.798
	+3.6%	**+2.8%**	**+1.9%**	**+2.5%**	**+3.5%**
RAP-MLLM	0.688	0.690	0.656	0.423	0.748

Table 2: Results of synthetic data used in multi concept scenarios. Our CaT framework also well supports multi-concept scenarios. When trained with a combination of original positive data and all synthetic data, both Yo'LLaVA and MC-LLaVA gain an impressive improvement. The numbers following the "+" symbol represent the improvement of the new SOTA compared to original baseline. Real = Original Data; Syn = Synthetic Data; P = Plus. **Higher** values are preferred for all metrics.

4.3 Ablations and Analysis

The Use of PCS Score Metric. We compare filtering strategy based on the PCS score with the cosine similarity-based filtering strategy that relies on image features. The results presented in Tab. 3 show that the performance of synthetic data without any filtering is even worse than the original Yo'LLaVA [38] 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 3: Ablation of different filtering metrics.

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. 4, 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, removing decreases the diversity of generated results due to a reduction in combinable attributes. In accordance with Fig. 3, 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 framework can provide precise control over this process, ensuring that optimal diversity leads to stable improvements.

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 4: Ablation study on tree editing operations (TEO). The type and number of operations can both alter the diversity of synthesized data.

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. 5, 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 observation experiments.

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 (+1.7%)	0.816 (+1.5%)	0.716 (+1.3%)	0.656 (+1.3%)	0.721 (+2.0%)
w/ Syn Easy Neg	0.873 (+1.5%)	0.824 (+0.8%)	0.721 (+0.5%)	0.667 (+1.1%)	0.727 (+0.6%)
w/ Syn Hard Neg	0.885 (+1.2%)	0.845 (+2.1%)	0.738 (+1.7%)	0.680 (+1.3%)	0.754 (+2.7%)

Table 5: Ablation study on data component.

Quality of Synthesized Data. As discussed in Sec. 3.3, a higher PCS score indicates that an image contains more concept-specific 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.

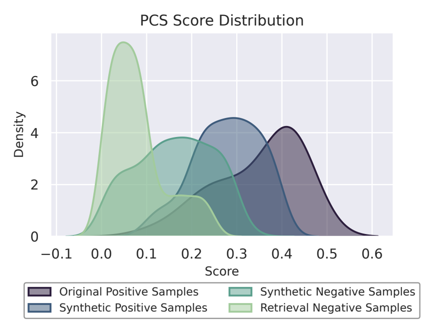

Fig. 6: The distribution of PCS scores across different datasets.

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 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 practical human assistants.

Acknowledgements

This work is supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM113), National Natural Science Foundation of China (92470121, 62402016), National Key R&D Program of China (2024YFA1014003), Zhongguancun Academy (C20250204, C20250602), Beijing Major Science and Technology Project (Z251100008125043, Z251100008425023), and High-performance Computing Platform of Peking University.

References

1. Alaluf, Y., Richardson, E., Tulyakov, S., Aberman, K., Cohen-Or, D.: Myvlm: Personalizing vlms for user-specific queries. In: European Conference on Computer Vision. pp. 73–91. Springer (2025)
2. An, R., Yang, S., Lu, M., Zeng, K., Luo, Y., Chen, Y., Cao, J., Liang, H., She, Q., Zhang, S., Zhang, W.: Mc-llava: Multi-concept personalized vision-language model (2024), `https://arxiv.org/abs/2411.11706`
3. An, R., Yang, S., Lu, M., Zeng, K., Luo, Y., Chen, Y., Cao, J., Liang, H., She, Q., Zhang, S., et al.: Mc-llava: Multi-concept personalized vision-language model. arXiv preprint arXiv:2411.11706 (2024)
4. An, R., Yang, S., Zhang, R., Shen, Z., Lu, M., Dai, G., Liang, H., Guo, Z., Yan, S., Luo, Y., et al.: Unictokens: Boosting personalized understanding and generation via unified concept tokens. arXiv preprint arXiv:2505.14671 (2025)
5. Awotunde, J.B., Chakraborty, C., Adeniyi, A.E.: 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)
6. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
7. Cao, J., Zhang, Y., Huang, T., Lu, M., Zhang, Q., An, R., Ma, N., Zhang, S.: Move-kd: Knowledge distillation for vlms with mixture of visual encoders. arXiv preprint arXiv:2501.01709 (2025)
8. Chawla, N.V., Davis, D.A.: Bringing big data to personalized healthcare: a patient-centered framework. Journal of general internal medicine **28**, 660–665 (2013)
9. Chen, R., Wu, Y., Chen, L., Liu, G., He, Q., Xiong, T., Liu, C., Guo, J., Huang, H.: Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. arXiv preprint arXiv:2402.12501 (2024)

10. Dong, H., Kang, Z., Yin, W., Liang, X., Feng, C., Ran, J.: Scalable vision language model training via high quality data curation. arXiv preprint arXiv:2501.05952 (2025)
11. Dorka, N., Marecki, J., Anwar, A.: Training a vision language model as smartphone assistant. arXiv preprint arXiv:2404.08755 (2024)
12. Dunlap, L., Umino, A., Zhang, H., Yang, J., Gonzalez, J.E., Darrell, T.: Diversify your vision datasets with automatic diffusion-based augmentation. Advances in neural information processing systems **36**, 79024–79034 (2023)
13. Fabris, M., Ceccato, R., Zanella, A.: Efficient sensors selection for traffic flow monitoring: An overview of model-based techniques leveraging network observability. Sensors **25**(5), 1416 (2025)
14. Fan, L., Krishnan, D., Isola, P., Katabi, D., Tian, Y.: Improving clip training with language rewrites. Advances in Neural Information Processing Systems **36**, 35544–35575 (2023)
15. Feng, C., Zhong, Y., Jie, Z., Xie, W., Ma, L.: Instagen: Enhancing object detection by training on synthetic dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14121–14130 (2024)
16. Feng, J., Xu, B., Chen, J., Dai, M., Wu, C., Li, H., Zeng, B., Xie, Y., Liang, H., Lu, M., et al.: M2a: Multimodal memory agent with dual-layer hybrid memory for long-term personalized interactions. arXiv preprint arXiv:2602.07624 (2026)
17. Goyal, S., Maini, P., Lipton, Z.C., Raghunathan, A., Kolter, J.Z.: Scaling laws for data filtering–data curation cannot be compute agnostic. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22702–22711 (2024)
18. Guo, X., Chen, Y.: Generative ai for synthetic data generation: Methods, challenges and the future. arXiv preprint arXiv:2403.04190 (2024)
19. Hao, H., Han, J., Li, C., Li, Y.F., Yue, X.: Remember, retrieve and generate: Understanding infinite visual concepts as your personalized assistant (2024), `https://arxiv.org/abs/2410.13360`
20. Hao, H., Han, J., Li, C., Li, Y.F., Yue, X.: Rap: Retrieval-augmented personalization for multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14538–14548 (2025)
21. Hong, M., Zhang, C.J., Chen, C., Lian, R., Jiang, D.: Dialogue language model with large-scale persona data engineering (2025), `https://arxiv.org/abs/2412.09034`
22. Hong, R., Agnew, W., Kohno, T., Morgenstern, J.: 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. pp. 1–17 (2024)
23. Huang, R., Xu, J., Jiang, X., An, R., Yang, Y.: Can modifying data address graph domain adaptation? (2024), `https://arxiv.org/abs/2407.19311`
24. Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S.N., Weller, A.: Synthetic data–what, why and how? arXiv preprint arXiv:2205.03257 (2022)
25. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
26. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024)

27. Li, H., Dong, Q., Tang, Z., Wang, C., Zhang, X., Huang, H., Huang, S., Huang, X., Huang, Z., Zhang, D., et al.: Synthetic data (almost) from scratch: Generalized instruction tuning for language models. arXiv preprint arXiv:2402.13064 (2024)
28. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
29. Liang, H., Sun, L., Wei, J., Huang, X., Sun, L., Yu, B., He, C., Zhang, W.: Synth-empathy: Towards high-quality synthetic empathy data. arXiv preprint arXiv:2407.21669 (2024)
30. Lin, W., Wei, X., An, R., Gao, P., Zou, B., Luo, Y., Huang, S., Zhang, S., Li, H.: Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. arXiv preprint arXiv:2403.20271 (2024)
31. Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al.: Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575 (2023)
32. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
33. Liu, J., Yang, S., Jia, P., Zhang, R., Lu, M., Guo, Y., Xue, W., Zhang, S.: Vida: Homeostatic visual domain adapter for continual test time adaptation. arXiv preprint arXiv:2306.04344 (2023)
34. Liu, Z., Liang, H., Huang, X., Xiong, W., Yu, Q., Sun, L., Chen, C., He, C., Cui, B., Zhang, W.: Synthvlm: High-efficiency and high-quality synthetic data for vision language models. arXiv preprint arXiv:2407.20756 (2024)
35. Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., Wang, H.: On llms-driven synthetic data generation, curation, and evaluation: A survey. arXiv preprint arXiv:2406.15126 (2024)
36. Luo, Y., An, R., Zou, B., Tang, Y., Liu, J., Zhang, S.: Llm as dataset analyst: Subpopulation structure discovery with large language model. arXiv preprint arXiv:2405.02363 (2024)
37. Meghwani, H.: Enhancing retrieval performance: An ensemble approach for hard negative mining. arXiv preprint arXiv:2411.02404 (2024)
38. Nguyen, T., Liu, H., Li, Y., Cai, M., Ojha, U., Lee, Y.J.: Yo'llava: Your personalized language and vision assistant. arXiv preprint arXiv:2406.09400 (2024)
39. Peng, W., Xie, S., You, Z., Lan, S., Wu, Z.: Synthesize diagnose and optimize: Towards fine-grained vision-language understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13279–13288 (2024)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
41. Ridzuan, F., Zainon, W.M.N.W.: A review on data quality dimensions for big data. Procedia Computer Science **234**, 341–348 (2024)
42. Shipard, J., Wiliem, A., Thanh, K.N., Xiang, W., Fookes, C.: 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. pp. 769–778 (2023)
43. Song, H., Wang, Y., Zhang, K., Zhang, W.N., Liu, T.: Bob: Bert over bert for training persona-based dialogue models from limited personalized data. arXiv preprint arXiv:2106.06169 (2021)

44. Wang, S., Song, L., Shimizu, R., Goto, M., wu, H.: Attributed synthetic data generation for zero-shot image classification. In: Synthetic Data for Computer Vision Workshop @ CVPR 2024 (2024), https://openreview.net/forum?id=k4Xnh0EPus
45. Wang, S., Song, L., Zhang, J., Shimizu, R., Luo, A., Yao, L., Chen, C., McAuley, J., Wu, H.: Template matters: Understanding the role of instruction templates in multimodal language model evaluation and training. arXiv preprint arXiv:2412.08307 (2024)
46. Wang, W., Mrini, K., Yang, L., Kumar, S., Tian, Y., Yan, X., Wang, H.: Finetuned multimodal language models are high-quality image-text data filters. arXiv preprint arXiv:2403.02677 (2024)
47. Welch, C., Gu, C., Kummerfeld, J.K., Perez-Rosas, V., Mihalcea, R.: 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). pp. 1742–1752 (2022)
48. Xu, B., Feng, J., Lu, S., Luo, Y., Yan, S., Liang, H., Lu, M., Zhang, W.: Jarvis: Towards personalized ai assistant via personal kv-cache retrieval. arXiv preprint arXiv:2510.22765 (2025)
49. Yang, L., Xu, X., Kang, B., Shi, Y., Zhao, H.: Freemask: Synthetic images with dense annotations make stronger segmentation models. Advances in Neural Information Processing Systems **36**, 18659–18675 (2023)
50. Yang, S., Ji, H., Lu, S., Chen, J., Xu, B., Lu, M., Zhang, Y., Dong, W., Zhang, W.: Small-large collaboration: Training-efficient concept personalization for large vlm using a meta personalized small vlm. arXiv preprint arXiv:2508.07260 (2025)
51. Yi, C., He, Y., Zhan, D.C., Ye, H.J.: Bridge the modality and capability gaps in vision-language model selection. Advances in Neural Information Processing Systems **37**, 34429–34452 (2025)
52. Zeng, K., Wu, Z., Xiong, K., Wei, X., Guo, X., Zhu, Z., Ho, K., Zhou, L., Zeng, B., Lu, M., et al.: Rethinking driving world model as synthetic data generator for perception tasks. arXiv preprint arXiv:2510.19195 (2025)
53. Zhang, Z., Zhang, Y., Ding, X., Yue, X.: Vision search assistant: Empower vision-language models as multimodal search engines. arXiv preprint arXiv:2410.21220 (2024)
54. Zheng, Y., An, R., Lin, X., Liu, Y., Yang, S., Zhang, H., Li, H., Zhang, Q., Zhang, R., Li, G., et al.: Pearl: Personalized streaming video understanding model. arXiv preprint arXiv:2603.20422 (2026)