NTU-DLCV-2024 Autumn EnglishOrSpanish

Challenge 2-Multiple Concept Personalization

范宇清 GIEE R12943180

黃子青 GIEE R11943004

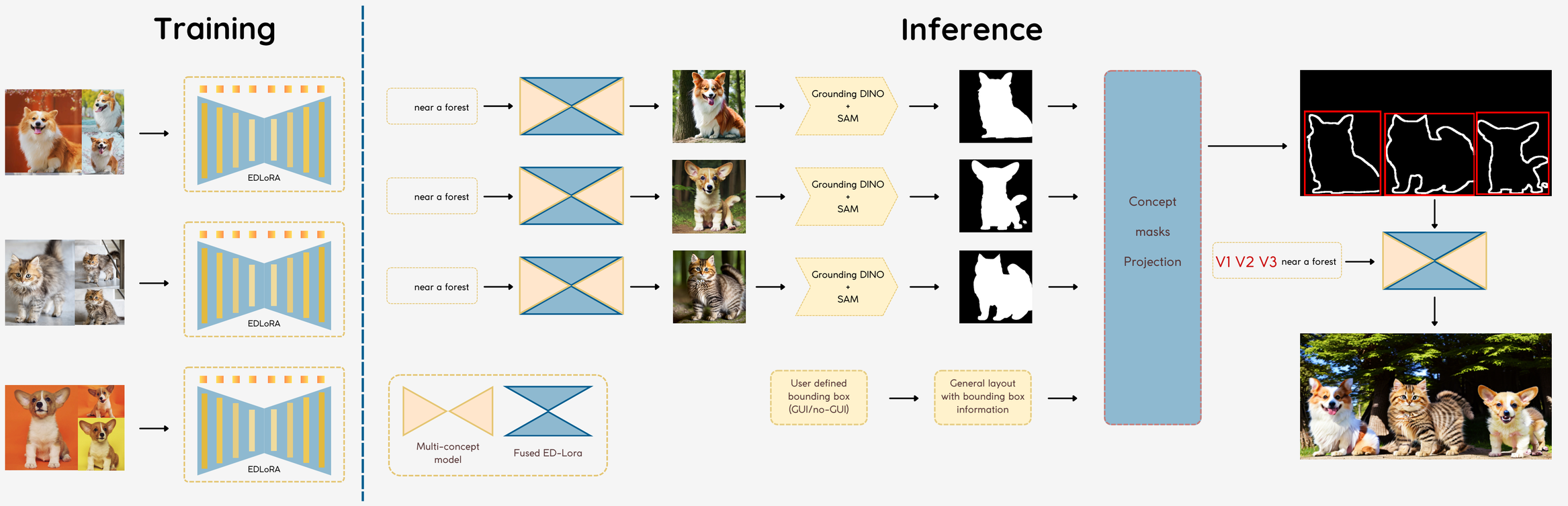
張根齊 GIEE R13943015

黃逸平 GSAT R13k41003

INTRODUCTION

Current research in Multiple Concept Personalization faces challenges in integrating new concept images effectively. Our approach leverages a Mix of Show-based architecture to combine various new concept images, enabling deeper and more accurate personalization. Additionally, we developed an intuitive GUI that allows users to manually draw bounding boxes, creating regional control masks for precise content customization. Through multiple experiments, we discovered that existing approaches in Multiple Concept Personalization do not offer a unified perfect solution, underscoring the need for more adaptable and comprehensive strategies.

METHODOLOGY



EXPERIMENTS

Impact of bbox on image quality

Object-Specific Bounding Boxes:

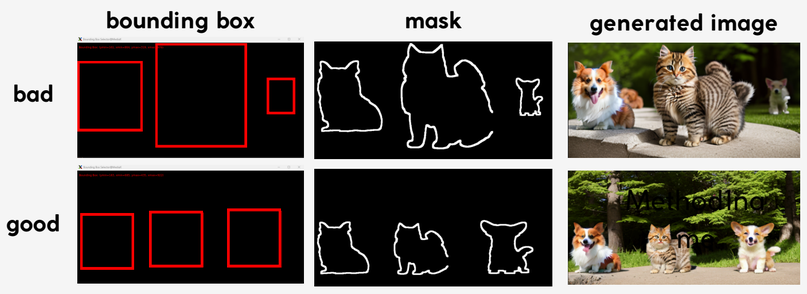
- Animal Detection: Square-shaped bounding boxes provide more accurate and visually appealing results when framing animals like cats and dogs which means that optimal bbox shape varies with the type of object, enhancing personalization precision, and this is the reason we develop user-difined bounding box GUI.

Multiple Bounding Boxes:

- Uniform Size: Using equal-sized bounding boxes for multiple objects results in more harmonious and consistent personalization.

Bounding Box Size:

- Generation Quality: Larger bounding boxes generally lead to higher quality content generation, as they provide more contextual information resulting in more accurate and visually coherent personalized content.



Orthogonal loss & Elastic LoRAcombination

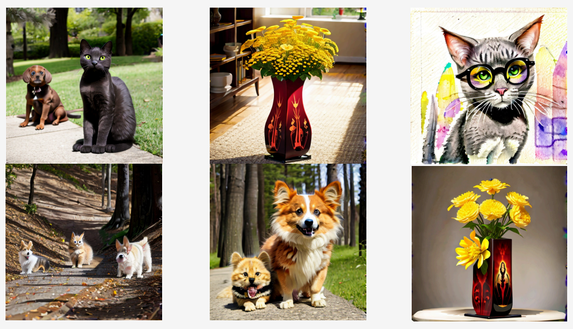
In one of our experiments, we hope to "separate" concepts by adding a subspace orthogonal regularizer. In the g-th task, the orthogonal loss is performed by $\sum_{i=1}^{g-1}\sum_{l=1}^{L}(\mathbf{A}_i^l(\mathbf{A}_g^l)^T)$. $\mathbf{A}_i^l$ represents the LoRA up weight of i-th concept and l-th attention layer.

Furthermore, a MSE between $\mathbf{A}_g^l$ and all-ones matrix is used to keep weight from zero matrix

At inference stage, a weighted average LoRA based on the response between text embedding and concept embedding is computed. The i-th response $R_i = max(\mathbf{e}_{i_{avg}}(\mathbf{e}_{prompt})^T)$, where $\mathbf{e}_{prompt}$ is the text encoder output of prompt and $\mathbf{e}_{i_{avg}}$ is the mean of all special embeddings of concept i.

However, if the count of concepts increase, the orthogonality of new concepts may decrease and lead to confusing.

not really good result



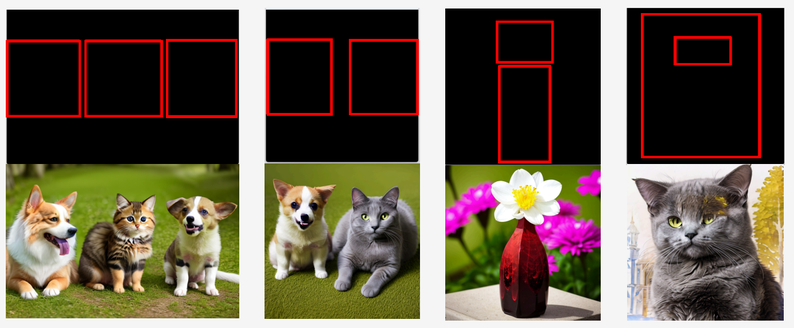
LoRA Composer

LoRA Composer is also a training-free framework that integrates multiple LoRAs into same diffusion model. It does not require masks but only bounding boxes to control the spatial placement of each concept.

By adding some constraints on cross-attention, it effectively addresse challenges like concept vanishing, ensuring each concept is precisely visualized without interference.

However, while LoRA Composer excels at mitigating concept vanishing, it struggles with handling overlapping bounding boxes. When bounding boxes overlap, the generation quality inevitably becomes less ideal, leading to suboptimal and less coherent visual outcomes. Additionally, concept confusion can sometimes result, compromising the distinctiveness of each concept.

not really good when bboxes overlap



RESULTS

	Prompt 1		Prompt 2		Prompt 3		Prompt 4		Average	
	CLIP-I	CLIP-T	CLIP-I	CLIP-T	CLIP-I	CLIP-T	CLIP-I	CLIP-T	CLIP-I	CLIP-T
Ours	78.81	31.50	73.96	31.03	77.09	27.82	63.28	34.33	73.29	31.17
Orthogonal	74.54	29.98	67.95	29.07	72.27	29.87	62.34	34.80	69.28	30.93
LoRA Composer	79.12	30.96	74.28	30.34	79.42	30.71	65.36	26.64	74.54	29.66

CONCLUSION

Our survey and experiments highlight the lack of a universal solution for Multiple Concept Personalization. While some methods work well with specific cases, they fail to generalize. For instance, regional control methods struggles with style concepts, and LoRA Composer performs poorly with overlapping bounding boxes. A potential solution could integrate various composition methods to one system.

To address this, we developed a Mix-of-Show based architecture with an intuitive GUI for the prompts of this challenge. However, the quality of generated images still heavily depends on the location and size of user-provided bounding boxes. We also observed discrepancies between human preference and CLIP scores, suggesting the need for improved evaluation metrics in future work.