C2P-VAR: CONTINUAL AND COMPOSITIONAL PERSONALIZATION IN VISUAL AUTOREGRESSIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual autoregressive (VAR) models have recently emerged as an efficient paradigm for text-to-image generation, yet their personalization capabilities remain largely limited to static, single-concept settings. In practice, users may continuously introduce new concepts and wish to compose multiple personalized concepts within a single image. Such scenarios pose two fundamental challenges: catastrophic forgetting during sequential personalization and feature interference during multi-concept composition. In this work, we study continual and compositional personalization in VAR models and propose C2P-VAR, a unified framework addressing both challenges. For continual personalization, we introduce C2PVAR-S, which identifies concept-relevant parameters from gradient magnitudes and dynamically updates their selection during training. To preserve previously learned concepts, C2PVAR-S applies regularization only to parameters shared by the current and historical concepts, thereby reducing unnecessary interference without introducing additional model components. For multi-concept personalization, we further propose C2PVAR-M, which employs parallel global and concept-specific branches with spatially localized feature fusion and logit aggregation to achieve controllable concept placement and reduce feature entanglement. Extensive experiments on continual and multi-concept personalization demonstrate that C2P-VAR consistently outperforms existing baselines in subject fidelity, while maintaining competitive text alignment and introducing negligible storage and inference overhead. Our results establish a unified framework for scalable and controllable personalization of visual autoregressive models.

1 INTRODUCTION

Recent advances in text-to-image generation (Peebles & Xie, 2023; Ramesh et al., 2022; Rombach et al., 2022; Xu et al., 2023; Sun et al., 2026) have been largely driven by diffusion models, which achieve remarkable visual quality through iterative denoising processes. More recently, Visual Autoregressive (VAR) models (Tian et al., 2024; Han et al., 2025) have emerged as a promising alternative, reformulating image generation as a coarse-to-fine next-scale prediction problem. By shifting the autoregressive unit from spatial tokens to resolution scales, VAR enables both high-quality synthesis and significantly improved inference efficiency (Zhong et al., 2025).

Despite these advantages, current VAR models remain limited in personalized generation. In practical applications, users often wish to generate images involving specific concepts, such as personal objects or customized styles, which are difficult to fully describe via text prompts alone. Existing personalization approaches address this by fine-tuning models on a small set of images, achieving promising results for single-concept generation. However, real-world personalization is inherently dynamic and incremental. Users continuously introduce new concepts over time, necessitating a continual personalization framework. Unfortunately, existing methods (Zhao et al., 2024a; Yang et al., 2025; Chen et al., 2024) struggle in this setting. First, sequential learning of new concepts leads to catastrophic forgetting, where previously acquired concepts are overwritten (Rebuffi et al., 2017; Dong et al., 2022). Second, composing multiple learned concepts in a single image introduces severe feature entanglement, resulting in incorrect attribute binding and visual artifacts

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Chefer et al., 2023; Feng et al., 2022; Jang et al., 2024; Lee et al., 2023; Ma et al., 2024; Wu et al., 2023; Yu et al., 2022). While continual learning and compositional generation have been extensively studied in diffusion models (Smith et al., 2023; Kumari et al., 2023; Po et al., 2024), directly transferring these techniques to VAR proves ineffective. Our empirical study reveals that existing methods suffer from substantial performance degradation when applied to VAR architectures, due to differences in generation dynamics and representation structure.

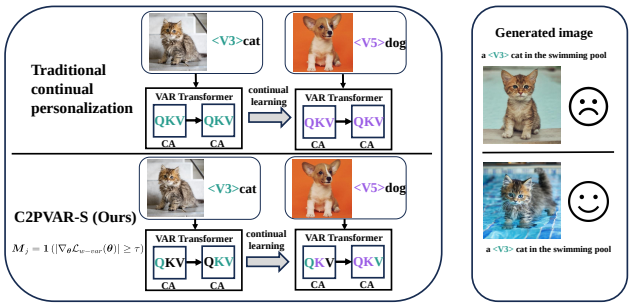


Figure 1: Schematic overview of our method (C2PVAR-S) versus the traditional full fine-tuning continual personalization approach: generating a special cat after learning two concepts.

In this paper, we present the first systematic investigation of continual personalized generation in VAR models. We identify two fundamental challenges: preserving concept knowledge across sequential updates and enabling controllable multi-concept composition. As illustrated in the Figure 1, traditional full-parameter fine-tuning during continual personalization causes subsequently learned concepts to completely overwrite previously acquired ones, thereby resulting in attribute overriding and catastrophic forgetting. To address these challenges, we propose a unified framework with two key components. For continual single-concept learning, we introduce C2PVAR-S, which dynamically identifies parameters most relevant to each concept based on gradient contributions. By enforcing regularization only on overlapping parameters across tasks, C2PVAR-S effectively mitigates catastrophic forgetting while avoiding unnecessary constraints on unrelated parameters. For multi-concept synthesis, we propose C2PVAR-M, which is tailored to the hierarchical generation process of VAR. Specifically, we perform multi-branch feature modeling at critical scales and introduce spatially guided cross-attention fusion using user-provided conditions such as bounding boxes (Li et al., 2023b; Shuai et al.). This design enables precise control over concept placement and significantly reduces feature interference.

Extensive experiments demonstrate that our approach outperforms existing baselines in both long-horizon continual personalization and multi-concept image synthesis. Our results highlight the unique challenges of personalization in VAR models and establish a strong foundation for future research in scalable and controllable generative systems. Our contributions are summarized as follows:

- We present the first study of continual personalized generation in VAR models, revealing the limitations of existing approaches in this setting.
- We propose C2PVAR-S, a parameter-efficient method that mitigates catastrophic forgetting via concept-specific parameter selection and conflict-aware regularization. We introduce C2PVAR-M, that enables precise and disentangled composition within the VAR framework.
- We demonstrate strong empirical performance across both continual learning and compositional generation benchmarks.

2 Related Work

Personalization within the VAR framework Concept personalization (Chen et al., 2024; Gal et al., 2023; Kim et al., 2024; Li et al., 2023a; Motamed et al., 2023; Zhang et al., 2024) aims to adapt pre-trained T2I generation models to synthesize personalized concepts utilizing only a limited number of exemplar images. Common personalized concepts typically encompass specific subjects or distinct visual styles. Pioneering work, such as ARBooth (Chung et al., 2025), achieved single-concept personalization within the VAR architecture for the first time through a Selective Layer Tuning strategy. This approach assigns a unique identifier (Ruiz et al., 2023) to a user-specific concept and specifically finetunes the Feed-Forward Network (FFN) and Cross-Attention (CA) layers of the VAR model. However, this method inherently assumes that the user trains the model on merely a

single specific concept. Consequently, it falls short of satisfying the practical demand for continual concept learning and is fundamentally incapable of facilitating multi-concept generation.

Continual concept personalization Continual concept personalization further aims to incrementally expand the repertoire of concepts learned within a T2I model. A variety of representative works have already been established based on diffusion models (Gu et al., 2023; Kumari et al., 2023). Smith et al. (2023) proposed a self-regularization loss among the LoRA weights of distinct tasks to preserve previously acquired concepts. However, as the number of concepts increases, the newly introduced LoRA weights become heavily constrained by all previously learned LoRA weights, resulting in a substantial degradation in both the learning capacity for new concepts and the robustness against catastrophic forgetting. Po et al. (2024) proposes to represent the weight residual of each concept using a fixed orthogonal basis B_i and a trainable low-rank matrix A_i, while encouraging the residual subspaces of different concepts to be approximately orthogonal. However, as the number of concepts increases, residual interference may still arise, potentially leading to catastrophic forgetting.

3 METHOD

3.1 PRELIMINARY

Under VAR (Tian et al., 2024), a vision encoder maps an image x to a continuous feature map $F \in \mathbb{R}^{h\times w\times c}$. The quantizer Q progressively converts F into discrete token maps $\{r_s\}_{s=1}^{S}$ at S spatial resolutions. At scale s, the residual feature is

$$f_s = \sum_{i=1}^{s} \mathrm{up}(r_i, (h, w)), \tag{1}$$

where $\mathrm{up}(\cdot,\cdot)$ upsamples a single-scale feature map to the target resolution, and f_s aggregates all scales up to s. During inference,

$$\tilde{f}_s = \mathrm{down}(f_s, (h_{s+1}, w_{s+1})) \tag{2}$$

is prepended as the initial tokens for predicting the next scale. A scale-wise causal mask enables localized bidirectional information flow, while the transformer predicts the residual feature map at the subsequent scale. Infinity (Han et al., 2025) adapts VAR for text-to-image generation by replacing the original VQ quantizer (Van Den Oord et al., 2017) with BSQ (Zhao et al., 2024b). Conditioned on a text prompt c, its likelihood is

$$p(r_1, r_2, \ldots, r_S) = \prod_{s=1}^{S} p(r_s \mid r_1, r_2, \ldots, r_{s-1}; c). \tag{3}$$

3.2 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

For Custom Visual Autoregressive Model (CVAR), given N subject images $X = \{x_n\}_{n=1}^{N}$ and a text prompt c_{sub} containing a concept token (e.g., $\langle S^{*}\mathrm{dog}\rangle$), personalization minimizes

$$\mathcal{L}_{\mathrm{var}} = -\sum_{s=1}^{S} \log p_\theta(r_s \mid r_1, r_2, \ldots, r_{s-1}; c_{\mathrm{sub}}), \tag{4}$$

where r_s denotes the multi-scale token maps extracted from x_n. Most CVARs assume a fixed number of personalized concepts, limiting their ability to accommodate evolving user demands. C2P-VAR therefore decomposes continual concept generation into (i) preserving previously learned concepts during sequential customization and (ii) controllably composing multiple personalized concepts.

To mitigate catastrophic forgetting, C2PVAR-S selects a compact set of concept-specific parameters according to absolute gradient magnitude; only these parameters are updated and regularized, as described in Sec. 3.3. C2PVAR-M performs multi-branch feature modeling and spatially guided localized cross-attention fusion for precise concept composition, as described in Sec. 3.4. The overall framework is shown in Figure 2.

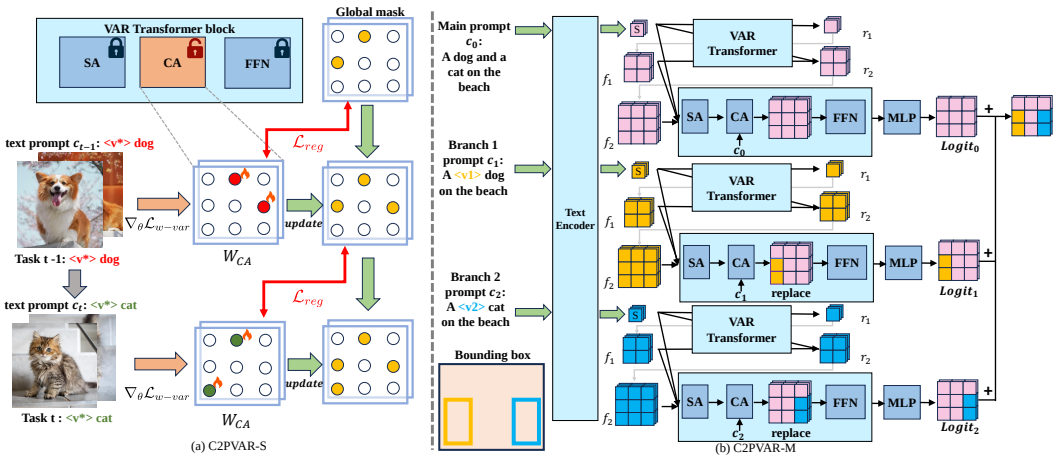


Figure 2: **Overall framework.** (a) C2PVAR-S resolves catastrophic forgetting and enable continual personalization. (b) C2PVAR-M addresses the challenge of concept neglect in multi-concept image synthesis.

3.3 SINGLE-CONCEPT CONTINUAL LEARNING

Concept parameter selection. When full-parameter fine-tuning is adopted, sequentially learning an increasing number of concepts on a single model leads to severe interference: updates for newly introduced concepts tend to overwrite parameters associated with previously learned ones, resulting in catastrophic forgetting. Therefore, it is crucial to develop a parameter selection mechanism that minimizes overlap across concepts and preserves previously acquired knowledge.

Our goal is to identify a mechanism that directs concept personalization toward a subset of model parameters that are most relevant to the target concept, thereby reducing parameter overlap across different concepts. Inspired by gradient-based saliency methods (Smilkov et al., 2017; Fan et al., 2023) for input attribution, we extend this idea to the parameter space and ask: can we construct a weight saliency map to guide continual concept personalization? This perspective allows us to partition model parameters into two subsets: (i) salient parameters that are critical for representing the current concept and should be updated, and (ii) non-salient parameters that remain frozen to preserve previously learned knowledge. Empirically, prior work such as ARBooth (Chung et al., 2025) has shown that cross-attention layers are strongly correlated with personalization objectives. In contrast, we observe that updating feed-forward network (FFN) layers often introduces interference among similar concepts. Based on these observations, we restrict parameter selection to the cross-attention layers of the VAR model.

Formally, consider the t-th task with model parameters $\theta_t \in \mathbb{R}^D$. During each training epoch, we compute the gradient $g \in \mathbb{R}^D$ of the loss function with respect to the parameters in the cross-attention layers:

$$g = \nabla_{\theta_t} \mathcal{L}_{var}(\theta_t), \tag{5}$$

where $\mathcal{L}_{var}(\theta_t)$ refers to Equation 4 defined previously. We define parameters whose absolute gradients exceed the threshold as the essential parameters for the current task (see details in the supplementary material). We then construct a binary importance mask $M \in \{0,1\}^D$ for the model. For the j-th parameter θ_j within the model, the activation rule for its corresponding mask m_j is formulated as follows:

$$m_j = \begin{cases} 1, & \text{if } |g_j| \geq \tau \\ 0, & \text{otherwise} \end{cases}. \tag{6}$$

When $m_j = 1$, it indicates that the given parameter is strongly activated by the current concept and must be included in the protected list.

Dynamic mask updating. We observe that important parameters vary across different training stages, and using a fixed mask increases interference among similar concepts, leading to catastrophic

forgetting (Table 3). To address this, we periodically refresh the mask during training. For task t, the gradient distribution is recalculated every e epochs to generate a phase-specific mask M_t^{ke}, where $k \in \mathbb{N}$. After training, all phase-specific masks are merged through a logical OR operation to obtain the final task mask M_t during E training epochs:

$$M_t = \bigvee_{k=0}^{\lfloor E/e \rfloor} M_t^{ke}. \tag{7}$$

Cross-task conflict regularization. Although task-specific masks identify important parameters, they cannot prevent interference across tasks. When learning task t, updates for the new concept may overwrite parameters important to previous tasks. To mitigate this, we introduce a global mask-based regularization mechanism. Before training task t, we first aggregate the masks of all previous tasks to form a historical mask $M_{<t}$:

$$M_{<t} = \bigvee_{i=1}^{t-1} M_i. \tag{8}$$

During training, parameters with $M_{<t,j} = 0$ are freely updated, while overlapping parameters with $M_{<t,j} = 1$ are constrained by an L_2 regularization term. The overall objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{var}(\theta_t) + \lambda \left\| M_{reg} \odot (\theta_t - \theta_{old}) \right\|_2^2, \tag{9}$$

where θ_{old} denotes the model weights of the previous task, λ is the regularization coefficient, $M_{reg} = M_{<t} \wedge M_t^{ke}$ denotes the overlap between the current update mask and historical mask, and $\odot$ denotes the Hadamard product.

Scale-wise weighted loss. We observe that coarse scales have a larger impact on generation quality than fine scales in VAR. Therefore, we prioritize learning at coarse scales by introducing a scale-weighted cross-entropy loss $\mathcal{L}_{w-var}$ to replace $\mathcal{L}_{var}$:

$$\mathcal{L}_{w-var} = -\sum_{s=1}^{S} w_s \log p_\theta(r_s \mid r_1, r_2, \ldots, r_{s-1}; c_{\text{sub}}). \tag{10}$$

3.4 Spatially Guided Concept Composition

When generating scenes containing multiple objects, traditional CVAR models often suffer from feature confusion (unintended blending between objects) and subject neglect (only one object appears). To address this, we propose C2PVAR-M for multi-concept generation. Given B customized concepts, the inference condition consists of one global condition and B local conditions. The global condition y_{global}, describing the overall scene, is assigned to the 0-th branch to establish the image layout. Each local condition contains a prompt y_i and a bounding box b_i, where $i \in \{1, \ldots, B\}$. The prompt y_i includes the special token learned during single-concept training and is assigned to the corresponding local branch.

We observe that when the resolution scale reaches $s \geq 3$, the global spatial structure is largely determined, making this stage suitable for spatial intervention. During inference at $s \geq 3$, the global and local branches first independently fuse text features through cross-attention. To constrain each concept within its target region while preserving global consistency elsewhere, we replace local features outside the target mask with global features.

Let $f_G \in \mathbb{R}^{L_q \times d}$ denote the global branch feature map at the cross-attention layer, and $f_i \in \mathbb{R}^{L_q \times d}$ denote the feature map of the i-th local branch. Let $\mathbf{1}$ be an all-ones tensor with the same shape as b_i. The fused local feature is computed as:

$$f_i^F = b_i \odot f_i + (\mathbf{1} - b_i) \odot f_G. \tag{11}$$

This operation preserves local concept features inside the mask while inheriting global features outside it. To better integrate personalized concepts with the background, we further perform logits-level fusion. Let L_G and L_i denote the predicted logits of the global branch and the i-th local branch, respectively. We introduce a hyperparameter α (set to 0.05 in our experiments) to control the influence of global features in local regions. The smoothed local logits are computed as:

$$\tilde{L}_i = \alpha L_G + (1 - \alpha) L_i. \tag{12}$$

We define the background mask as $b_G = \mathbf{1} - \bigvee_{i=1}^{B} b_i$. The final merged logits are then obtained by:

$$L_M = b_G \odot L_G + \sum_{i=1}^{B} \left(b_i \odot \tilde{L}_i \right). \tag{13}$$

Finally, the merged logits L_M are synchronized across all branches for prediction at the next scale.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUPS

Datasets. We construct the first challenging benchmark for concept incremental learning in VAR, consisting of 15 sequential concept customization tasks. Within this dataset, 13 customization tasks pertain to distinct object concepts collected from Chung et al. (2025), while the remaining two tasks involve different style concepts collected from websites. We establish approximately 3 to 5 text-image pairs for each task. Notably, we incorporate several semantically similar concepts (e.g., the dogs in V1 and V5), thereby rendering the dataset substantially more challenging within a continual personalization setting. More details are provided in the Appendix A.3.

Baselines. We compare our proposed method against two representative continual learning methods, three typical diffusion-based continual personalization approaches, and two standard finetuning baselines, which include LWF (Li & Hoiem, 2017), MAS (Aljundi et al., 2018), Continual Diffusion (Smith et al., 2023), Orthogonal Adaptation (Po et al., 2024), CNS (Liao et al., 2025), ARBooth (Chung et al., 2025), and LoRA (Hu et al., 2022). Further details are provided in the Appendix A.1.

Evaluation metrics. Following Gal et al. (2022); Kumari et al. (2023); Nam et al. (2024); Ruiz et al. (2023), we evaluate both subject fidelity and text prompt fidelity. To assess subject fidelity, we employ DINO (Caron et al., 2021) and CLIP (Radford et al., 2021) to measure the image similarity between the generated images and the reference subjects, denoted as DINO and CLIP-I, respectively. To evaluate text prompt alignment, we compute the CLIP image-text similarity by comparing the visual features of the generated images with the textual features of the prompts (substituting the special token with its corresponding class token), denoting this metric as CLIP-T. Additionally, for each generated multi-concept image, we first crop the region corresponding to each concept according to the bounding boxes specified during the generation process. We then compute the DINO and CLIP-I between each cropped region and the corresponding reference image, and finally average the scores across all concepts. CLIP-T is still computed between the complete generated image and the multi-concept prompt to evaluate the overall text-image consistency.

4.2 QUALITATIVE COMPARISONS

To validate the effectiveness of our model under concept incremental learning, we conduct qualitative comparisons on single-concept customization, multi-concept customization, and custom style transfer. As shown in Figure 3, our model achieves superior single-concept customization by mitigating catastrophic forgetting while preserving the unique attributes of previously learned concepts. For multi-concept customization, we incorporate the region-controllable sampling module from Gu et al. (2023) into baseline methods for fair comparison. As shown in Figure 4, due to the architectural differences between diffusion models and VAR, prompts in VAR influence the generation process from the beginning rather than only through cross-attention layers. Therefore, directly applying the CA-based control strategy of Gu et al. (2023) leads to severe concept neglect, where the generated image is dominated by the main prompt and fails to synthesize the target concepts. In contrast, our model effectively resolves this issue through C2PVAR-M. We provide more qualitative and quantitative results in Appendix.

4.3 QUANTITATIVE COMPARISONS

We conduct quantitative comparisons exclusively on single-concept inference. For each concept, we evaluate it using 20 prompts and generate 10 images per prompt, resulting in a total of 200 images, over which the metrics are averaged. As presented in Table 1, we report the results of the first 8 concepts in the main paper for clarity, while the results for all 15 concepts are provided

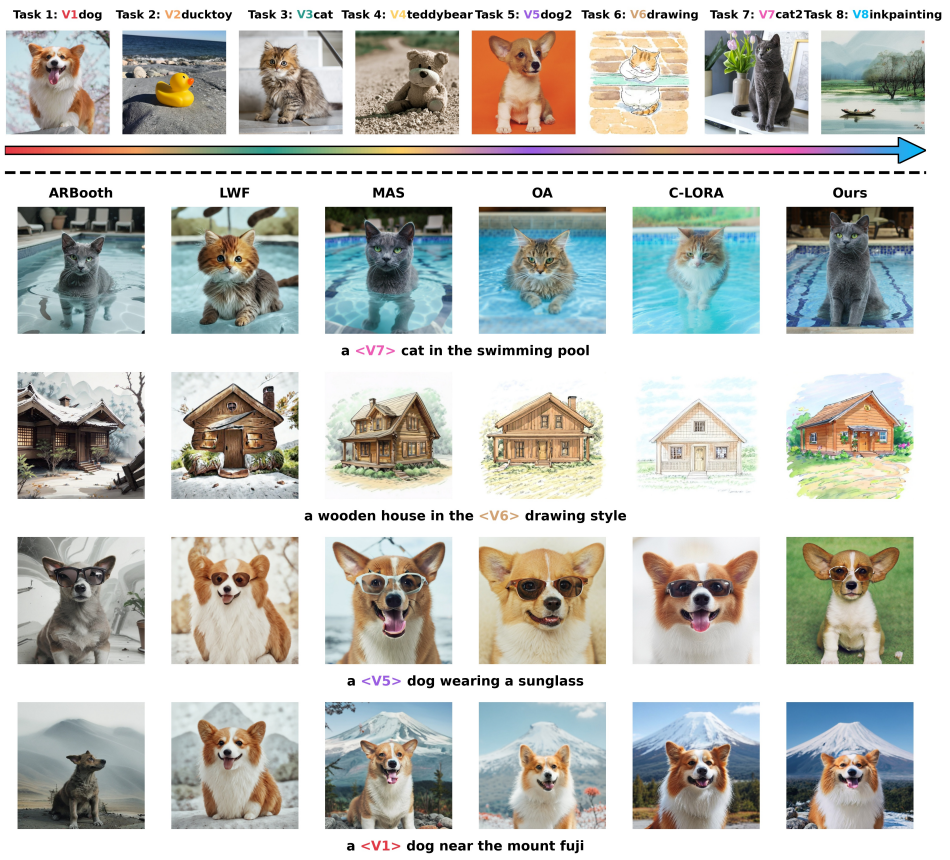


Figure 3: Qualitative comparison of single-concept customization with different baselines.

Table 1: Quantitative comparisons of single-concept customization across 8 tasks. We report the DINO metric for each individual concept, alongside the average DINO, CLIP-I, and CLIP-T metrics.

Methods	Single Concept (DINO↑)								Overall Metrics		
	V1	V2	V3	V4	V5	V6	V7	V8	Avg. DINO↑	Avg. CLIP-I↑	Avg. CLIP-T↑
ARBooth	40.78	45.71	62.36	66.03	61.62	17.80	78.97	40.16	51.68	75.99	31.28
LoRA	40.32	33.40	73.80	64.52	58.78	19.76	73.68	40.56	50.60	76.63	31.48
LWF	74.02	69.10	84.03	68.94	63.90	24.23	59.09	29.09	59.05	80.78	29.01
MAS	66.88	59.94	57.05	65.77	74.55	29.73	74.97	28.83	57.21	77.52	**32.65**
Orthogonal Adaptation	79.06	57.52	80.17	70.40	67.23	22.12	60.68	27.21	58.05	78.88	31.58
Continual Diffusion	82.49	57.59	83.82	73.44	69.28	23.30	63.66	26.78	60.04	78.68	31.85
CNS	37.16	47.83	70.78	63.36	35.30	22.15	60.13	25.73	45.31	75.61	32.58
C2PVAR-S (Ours)	81.89	69.17	82.86	75.50	80.36	41.80	78.21	35.57	**68.17**	**82.71**	31.14

in the Appendix. Our method outperforms diffusion-based baselines in terms of both DINO and CLIP-I, demonstrating enhanced subject alignment. The results in both the main table and the Appendix further demonstrate that our method effectively mitigates catastrophic forgetting in both short-horizon and long-horizon continual learning settings, achieving the best overall performance. We additionally report the standard deviations of the three evaluation metrics across the 8 tasks in Appendix A.5.

It is worth noting that the CLIP-T metric places a stronger emphasis on the overall alignment between the image background and the text prompt. However, our task definition inherently prioritizes the quality and fidelity of the customized subject. Consequently, there is typically a trade-off between the subject alignment metrics (DINO and CLIP-I) and the text alignment metric (CLIP-T). Taking this trade-off into consideration, our proposed method achieves the optimal overall performance.



Figure 4: Qualitative comparison of multi-concept customization, where Main Prompt indicates the global text prompt, and Branch Prompt denotes the region text prompt.

In addition, we construct 45 multi-concept prompts, create an appropriate bounding box for each concept, and generate 10 images for each prompt, with the average results reported for evaluation. All prompts follow one controlled evaluation protocol. As shown in Table 2, our method achieves significant improvements across all three metrics, demonstrating the best performance in both subject fidelity and text alignment for multi-concept personalization.

Table 2: Quantitative comparisons of multi-concept personalization. We report the average DINO, CLIP-I, and CLIP-T.

Methods	Avg. DINO↑	Avg. CLIP-I↑	Avg. CLIP-T↑
ARBooth	42.34	73.15	31.95
LoRA	41.91	74.66	32.51
LWF	53.51	78.89	29.82
MAS	57.14	79.12	33.83
Orthogonal Adaptation	56.52	80.44	32.77
Continual Diffusion	52.61	79.36	33.03
CNS	44.55	79.16	34.12
C2PVAR-M (Ours)	**68.37**	**83.35**	**35.24**

Concurrently, as shown in Appendix Table 14, we evaluate the memory and time consumption of various methods when integrating eight distinct concepts. LoRA-based methods generally require additional memory and computational time because multiple concept-specific weights must be stored and composed during inference. In contrast, C2PVAR-S requires zero additional storage space and completely avoids the inference-time overhead associated with LoRA-weight composition. This design provides a more efficient deployment profile while maintaining high generative performance, thereby achieving an optimal balance between fusion efficiency and generation quality.

Table 3: Ablation study on the effectiveness of Regularization, Scale Weighting, and Dynamic Masking. We report the DINO scores for each concept and the overall average.

Components			Single Concept(DINO ↑)								
Reg.	Weight	Mask	V1	V2	V3	V4	V5	V6	V7	V8	Avg.
			58.80	56.10	73.98	64.77	64.93	21.21	65.37	29.41	54.32
✓			71.76	60.58	79.84	77.29	81.37	33.32	78.17	33.96	64.54
✓	✓		74.74	63.52	76.72	76.08	80.36	38.31	78.75	33.79	65.28
✓	✓	✓	81.89	69.17	82.86	75.50	80.36	41.80	78.21	35.57	**68.17**

5 Ablation Study

Impact of layer selection in multi-concept synthesis. To validate the necessity of our continuous multi-scale intervention (operating consistently from $s \geq 3$), we conducted an ablation study by applying the regional mask intervention only at a single high-resolution scale (e.g., exclusively at $s = 7$ or $s = 11$). Figure 5 shows that isolated intervention fails to synthesize the specific concepts and leads to severe image distortion, with the cross-attention exhibiting a negligible effect. Furthermore, owing to the sparse token count at lower resolution scales ($s = 1$ and $s = 2$), the bounding boxes are inevitably downscaled to encompass merely one or two tokens. This extreme spatial compression causes the features injected by the final processing branch to completely override the features established by all preceding branches, thereby rendering the method ineffective.



Figure 5: Ablation study on the intervention scale s

Effectiveness of regularization, dynamic masking and scale weighting. We investigate the effects of regularization, dynamic mask updating and scale weighting. As shown in Table 3, without regularization loss, a subset of these overlapping parameters is overwritten by newly introduced concepts, consequently leading to catastrophic forgetting and a degradation in overall performance. We set $w_{>8} = 0.5$ and set the other weight to 1.0. As shown in the third row of the table, scale weighting effectively increases the DINO metrics, indicating an enhancement in the model's learning capacity. Meanwhile, dynamic mask updating effectively mitigates mutual interference among semantically similar concepts. The DINO scores for the V1dog and V3cat exhibit substantial improvements, demonstrating the successful alleviation of catastrophic forgetting.

6 Conclusion

We address two challenges in VAR-based personalization: catastrophic forgetting during concept-incremental customization and feature entanglement in multi-concept synthesis. We present C2P-VAR, the first continual and compositional personalization framework for VAR. C2PVAR-S updates task-relevant parameter subspaces with conflict-aware regularization, mitigating forgetting without data replay. C2PVAR-M enables spatially controlled feature fusion and branch-wise logit aggregation during inference. Experiments show that C2P-VAR outperforms state-of-the-art diffusion methods in both tasks while maintaining low computational and storage overhead.

AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing of the manuscript, including improving the clarity and readability of existing text. We did not use generative AI tools to generate or develop research ideas, methodology, experiments, analysis, or scientific claims. All AI-assisted text was carefully reviewed and verified by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

Our work does not involve human subjects, human participant studies, or the collection of personally identifiable information. All concept images used in our experiments are obtained from publicly available datasets or publicly accessible sources, and are used solely for research and evaluation purposes in accordance with the usage conditions and licenses of their respective sources. We do not collect, process, or release any private or sensitive personal data.

Our work investigates continual learning and multi-concept personalization for generative models. While generative models may have potential dual-use risks, we are not aware of any specific harmful applications introduced by the proposed method beyond the general risks associated with image generation and personalization. We follow standard research practices and report our experimental results according to the stated evaluation protocols. We have no known conflicts of interest or external sponsorship that could influence this work.

REPRODUCIBILITY STATEMENT

All implementation details, including the hyperparameter settings and experimental configurations, are described in Appendix A.2 and Appendix A.4, providing the necessary details for reproducing our experiments. The datasets used for experiments are already publicly available.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *European conference on computer vision*, pp. 144–161. Springer, 2018.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6593–6602, 2024.

Jiwoo Chung, Sangeek Hyun, Hyunjun Kim, Eunseo Koh, MinKyu Lee, and Jae-Pil Heo. Fine-tuning visual autoregressive models for subject-driven generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19174–19184. IEEE, 2025.

Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. Federated class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10164–10173, 2022.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

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

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36:15890–15902, 2023.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15733–15744, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. *Advances in Neural Information Processing Systems*, 37: 100895–100937, 2024.

Jimyeong Kim, Jungwon Park, and Wonjong Rhee. Selectively informative description can reduce undesired embedding entanglements in text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8312–8322, 2024.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1931–1941, 2023.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023a.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22511–22521, 2023b.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Yu-Chien Liao, Jr-Jen Chen, Chi-Pin Huang, Ci-Siang Lin, Meng-Lin Wu, and Yu-Chiang Frank Wang. Continual personalization for diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15511–15520. IEEE, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Wan-Duo Kurt Ma, Avisek Lahiri, John P Lewis, Thomas Leung, and W Bastiaan Kleijn. Directed diffusion: Direct control of object placement through attention guidance. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4098–4106, 2024.

Saman Motamed, Danda Pani Paudel, and Luc Van Gool. Lego: Learning to disentangle and invert personalized concepts beyond object appearance in text-to-image diffusion models. *arXiv preprint arXiv:2311.13833*, 2023.

Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8100–8110, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular customization of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7964–7973, 2024.

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*, pp. 8748–8763. PmLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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*, pp. 22500–22510, 2023.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

X Shuai, H Ding, X Ma, R Tu, YG Jiang, and D Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. arxiv 2024. *arXiv preprint arXiv:2406.14555*.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *arXiv preprint arXiv:2304.06027*, 2023.

Yi Sun, Yimin Zhou, Xinhao Zhong, Zhiqi Zhang, Junhao Li, and Bin Chen. Flowerase-opd: Multi-concept erasure via anchored on-policy distillation in flow matching models. *arXiv preprint arXiv:2608.07620*, 2026.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2096–2105, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Xiaofei He, et al. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *IEEE Transactions on Image Processing*, 34:8145–8158, 2025.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Yanbing Zhang, Mengping Yang, Qin Zhou, and Zhe Wang. Attention calibration for disentangled text-to-image personalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4764–4774, 2024.

Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 273–290. Springer, 2024a.

Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. *arXiv preprint arXiv:2406.07548*, 2024b.

Xinhao Zhong, Yimin Zhou, Zhiqi Zhang, Junhao Li, Yi Sun, Bin Chen, Shu-Tao Xia, Xuan Wang, and Ke Xu. Closing the safety gap: Surgical concept erasure in visual autoregressive models. *arXiv preprint arXiv:2509.22400*, 2025.

A Appendix

A.1 Baseline method

ARBooth (Chung et al., 2025) We clone the code base of ARBooth from official GitHub repository. The learning rate of text embedding is 2e-5 and Infinity model are 2e-5 while training steps is set to 300. For multi-concept personalization, we naively feed all the trained special tokens we need to the model.

LoRA (Hu et al., 2022) We naively integrate LoRA into the CA and FFN layers of the Infinity model, with the rank r set to 64, and conduct continual training on the same pair of LoRA matrices. The learning rate of text embedding is 2e-5 and LoRA matrices are 2e-5 while training steps is set to 300. For multi-concept personalization, we naively feed all the trained special tokens we need to the model.

LWF (Li & Hoiem, 2017) We adapt LwF to the Infinity model. For each new concept, the student model is initialized from the checkpoint of the previous task, while a frozen copy of the same checkpoint is used as the teacher model. During training, we align the autoregressive output distributions of the teacher and the currently optimized student model with a temperature-scaled KL divergence loss, following the standard LwF objective:

$$\mathcal{L} = \mathcal{L}_{\text{new}} + \lambda_{\text{LwF}} \tau^2 \sum_t \text{KL}\left(\text{softmax}\left(\frac{z_t^T}{\tau}\right) \middle\| \text{softmax}\left(\frac{z_t^S}{\tau}\right)\right), \tag{14}$$

where z_t^T and z_t^S denote the teacher and student autoregressive logits at token position t, and τ is the distillation temperature. The learning rates of both the text embedding and the Infinity model are set to 2e-5, and each concept is trained for 300 steps. For multi-concept personalization, we naively feed all the trained special tokens required by the prompt to the model.

MAS (Aljundi et al., 2018) We adapt MAS to the Infinity model as an importance-regularization baseline. Before optimizing the new concept, we load the accumulated MAS importance weights from previous tasks and penalize changes to important historical parameters with a quadratic regularization term as follows:

$$\mathcal{R}_{\text{MAS}}(\theta) = \lambda_{\text{MAS}} \sum_i \Omega_i \left(\theta_i - \theta_i^*\right)^2, \tag{15}$$

where Ω_i denotes the accumulated importance of parameter θ_i, and θ_i^* denotes its value after learning previous tasks. After finishing each task, we estimate the current-task importance by computing the gradient magnitude of the squared autoregressive logits with respect to the trainable parameters, and merge it into the historical importance with a decay factor. The learning rates of both the text embedding and the Infinity model are set to 2e-5, and each concept is trained for 300 steps. For multi-concept personalization, we naively feed all the trained special tokens we need to the model.

Orthogonal Adaptation (Po et al., 2024) We adapt Orthogonal Adaptation to the Infinity model by ourselves because official code is not available. For each adapted linear layer, the weight residual is parameterized as $\Delta W_i = A_i B_i^T$, where B_i is fixed and sampled from a shared randomized orthogonal basis, while only A_i and the concept embedding are optimized. We apply the adapters to the cross-attention and feed-forward linear layers, train each concept independently for 300 steps, and set the learning rates of both the text embedding and adaptation parameters to 2e-5. During inference, multiple concepts are merged by summing their learned residual weights with a fixed merge coefficient.

Continual Diffusion (Smith et al., 2023) We adapt Continual Diffusion to the Infinity model by ourselves, as the official code is unavailable. We follow the self-regularization loss presented in Continual Diffusion to fulfill continual personalization. For the t-th concept, we follow the C-LoRA self-regularization objective in Continual Diffusion. Let ΔW_t^ℓ denote the LoRA-induced weight update of layer ℓ for the current concept, and let $\Delta W_{<t}^\ell = \sum_{i=1}^{t-1} \Delta W_i^\ell$ denote the accumulated LoRA updates from previous concepts. The self-regularization loss is formulated as below:

$$\mathcal{L}_{\text{C-LoRA}} = \sum_{\ell \in \mathcal{A}} \left\| \Delta W_{<t}^\ell \odot \Delta W_t^\ell \right\|_F^2, \tag{16}$$

where $\odot$ denotes element-wise multiplication and $\mathcal{A}$ is the set of adapted layers. We apply the adapters to the cross-attention and feed-forward linear layers. The learning rates of both the text embedding and LoRA parameters are set to 2e-5, and each concept is trained for 300 steps.

CNS (Liao et al., 2025) We adapt CNS to the Infinity model by ourselves, as the official code is unavailable. For each new concept, we train only the selected rows of the W_k and W_v matrices in cross-attention layers. Following CNS, given a caption embedding c and a key/value weight matrix W, we compute the neuron importance as $S(W, c) = |W| \odot (\mathbf{1} \cdot \|c\|_2)$ denotes the channel-wise text-embedding intensity and is broadcast to all rows of W_k and W_v. We use ImageNet calibration captions to identify general neurons and the current concept captions to identify concept-specific neurons. We set the row selection ratio to 5% and use λ_1=1.0 and λ_2=1.0 for the overlap regularization terms following the paper setting. The learning rates of both the text embedding and the Infinity model are set to 2e-5, and each concept is trained for 300 steps. For multi-concept personalization, we naively feed all the trained special tokens we need to the model.

A.2 Implementation details

We conduct all experiments utilizing the Infinity-2B model (Han et al., 2025), which is pretrained on the LAION (Schuhmann et al., 2021), COYO (Byeon et al., 2022), and OpenImages (Kuznetsova et al., 2020) datasets. We adopt the default scale configuration ($S = 13$). We employ the AdamW optimizer (Loshchilov & Hutter, 2017) ($\beta_0 = 0.9, \beta_1 = 0.97$), setting the learning rate to 2e-5 for the text embeddings and 2e-5 for the concept neurons. The model is finetuned for 300 iterations at a resolution of 1024 with a batch size of 1 on a single NVIDIA A6000 GPU. We empirically set λ to 1, and detailed ablation experiments are presented in Table 13.

A.3 Datasets

We collect 13 object concepts from Chung et al. (2025) and 2 style concepts from online sources, resulting in a total of 15 concepts: V1 dog, V2 duck toy, V3 cat, V4 teddy bear, V5 dog, V6 style, V7 cat, V8 style, V9 clock, V10 dog, V11 monster toy, V12 sunglasses, V13 teapot, V14 vase, and V15 wooden barn. For multi-concept image evaluation, we further construct 45 prompts with corresponding bounding boxes, including 25 prompts featuring two object concepts arranged vertically, 15 prompts featuring two object concepts arranged horizontally, and 5 prompts involving three concepts. Each multi-concept prompt specifies the target concepts and their spatial arrangement, together with the corresponding bounding boxes used to control their locations during generation. An example of the constructed multi-concept prompts and their corresponding bounding boxes is provided in the Table 4. We evaluate our framework and all competing baselines under the same dataset and evaluation settings.

Table 4: An example of the multi-concept prompts and corresponding bounding boxes used in our evaluation.

Item	**Example**
Task	V1 dog + V2 duck toy
Branch prompts	"a V1 dog sitting in a magical garden"; "a V2 duck toy sitting in a magical garden"
Main prompt	"A photo of a dog sitting on the left and a duck toy sitting on the right in a magical garden"
Bounding boxes	dog: $[0.1, 0.25, 0.45, 0.8]$; duck toy: $[0.65, 0.25, 0.95, 0.8]$

A.4 Threshold of the neuron selection

We identify the top 5% of parameters as the crucial concept parameters for each task. We conduct an additional ablation study on the threshold for selecting crucial concept parameters. As shown in the Table 5, a threshold of 5% achieves the best trade-off among the three evaluation metrics, demonstrating the most balanced overall performance.

Table 5: Ablation study on the threshold for concept parameters selection. We report the average DINO, CLIP-I, and CLIP-T scores across all tasks.

Threshold	**Avg. DINO ↑**	**Avg. CLIP-I ↑**	**Avg. CLIP-T ↑**
5%	68.17	82.71	31.14
1%	66.15	81.69	31.31
10%	67.65	83.02	30.69

A.5 Additional quantitative comparisons

We have added additional quantitative comparisons in this section. As shown in Table 6, Table 7, our method outperforms diffusion-based baselines in terms of CLIP-I, demonstrating enhanced sub-

ject alignment. It is worth noting that the CLIP-T metric places a stronger emphasis on the overall alignment between the image background and the text prompt. However, our task definition inherently prioritizes the quality and fidelity of the customized subject. Consequently, there is typically a trade-off between the subject alignment metrics (DINO and CLIP-I) and the text alignment metric (CLIP-T). As shown in Table 8, Table 9 and 10, our method continues to outperform all baselines in terms of both DINO and CLIP-I in the long-horizon continual learning setting, demonstrating its strong robustness and effectiveness under prolonged continual learning. As shown in Table 11, we additionally report the standard deviations of the three evaluation metrics across the 8 tasks, obtained from three independent runs with different random seeds. These results further demonstrate the reliability of our evaluation and the robustness of our method.

As shown in Table 14, we evaluate the memory and time consumption of various methods when integrating eight distinct concepts. It is noteworthy that LoRA-based methods necessitate memory and computational time resources that substantially exceed those required by C2PVAR-S. In contrast, C2PVAR-S requires zero additional storage space and completely circumvents the inference-time overhead typically associated with LoRA weights composition, thereby demonstrating an optimal balance between fusion efficiency and high generative performance.

Table 6: Quantitative comparisons of single-concept customization across 8 tasks (CLIP-I metric).

Methods	**Single Concept (CLIP-I↑)**								
	V1	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**Avg.**
ARBooth	76.26	77.43	77.67	79.51	77.48	61.68	80.87	77.03	75.99
LoRA	79.67	72.80	81.46	81.90	78.88	61.36	80.02	76.94	76.63
LWF	86.37	84.18	89.21	88.64	85.87	62.25	79.22	70.52	80.78
MAS	84.33	82.49	77.53	81.81	82.35	60.74	80.92	70.02	77.52
Orthogonal Adaptation	86.15	83.46	85.78	84.30	83.59	62.68	75.82	69.30	78.88
Continual Diffusion	87.79	82.95	83.66	84.91	83.50	62.07	76.66	67.91	78.68
CNS	79.20	81.00	79.24	84.07	76.10	60.47	76.80	68.00	75.61
C2PVAR-S (Ours)	85.75	84.83	86.04	89.87	87.32	67.13	84.67	76.06	**82.71**

Table 7: Quantitative comparisons of single-concept customization across 8 tasks (CLIP-T metric).

Methods	**Single Concept (CLIP-T↑)**								
	V1	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**Avg.**
ARBooth	30.52	33.02	30.54	32.53	30.59	31.33	30.75	30.96	31.28
LoRA	30.08	32.41	30.85	33.06	30.86	31.76	31.91	30.92	31.48
LWF	26.09	31.96	28.26	29.24	26.92	30.58	28.83	30.17	29.01
MAS	30.69	33.92	31.86	34.10	30.71	34.06	31.83	34.01	32.65
Orthogonal Adaptation	28.70	34.24	29.33	32.23	29.84	33.43	32.32	32.51	31.58
Continual Diffusion	30.05	33.02	31.74	32.19	30.07	33.29	32.36	32.08	31.85
CNS	31.26	33.34	32.22	33.17	31.30	34.17	32.28	32.92	32.58
C2PVAR-S (Ours)	31.26	33.72	29.85	31.41	29.53	32.33	29.71	31.31	31.14

Table 8: Quantitative comparisons of single-concept customization across 15 tasks (DINO metric).

Methods	**Single Concept (DINO↑)**															
	V1	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**V9**	**V10**	**V11**	**V12**	**V13**	**V14**	**V15**	**Avg.**
ARBooth	44.73	44.62	60.64	65.24	27.09	18.32	58.54	28.48	58.64	62.70	42.78	53.98	54.45	56.60	84.98	50.79
LoRA	31.59	34.88	57.33	63.21	27.35	20.08	48.25	32.84	58.85	41.73	42.63	56.30	52.20	51.08	82.15	46.70
LWF	67.01	65.32	78.75	55.83	64.13	25.72	64.44	20.21	36.18	21.19	27.71	58.06	46.01	47.12	60.93	49.24
MAS	65.09	64.72	62.24	67.66	41.38	33.31	75.20	31.79	69.26	72.70	26.87	67.35	48.81	48.44	72.01	56.46
Orthogonal Adaptation	70.64	54.95	74.98	63.90	54.32	20.21	60.70	21.80	49.63	48.53	23.84	53.00	45.00	29.63	44.22	47.69
Continual Diffusion	76.59	50.32	80.51	72.00	59.62	19.42	62.67	28.01	68.45	46.97	36.49	64.14	54.50	37.01	56.23	54.19
CNS	36.40	48.01	71.13	62.57	34.99	22.75	60.04	25.49	59.40	26.10	28.15	59.82	48.72	37.92	67.87	45.96
C2PVAR-S (Ours)	84.06	63.50	87.11	75.76	80.71	40.80	78.93	37.45	71.25	80.14	56.04	64.52	67.14	67.36	78.44	**68.88**

Table 9: Quantitative comparisons of single-concept customization across 15 tasks (CLIP-I metric).

Methods	Single Concept (CLIP-I↑)															
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	Avg.
ARBooth	74.74	79.63	74.70	77.95	70.08	55.87	72.11	68.15	77.03	78.54	68.19	78.46	73.50	76.79	86.25	74.13
LoRA	73.28	76.04	76.95	82.50	71.74	58.76	75.10	73.12	78.65	75.08	68.73	79.20	74.72	75.41	82.76	74.80
LWF	85.88	81.31	86.10	81.77	84.01	62.50	82.69	66.69	70.93	73.61	63.22	79.97	75.75	75.81	70.33	76.04
MAS	81.56	82.33	77.85	83.51	75.50	62.64	80.76	68.55	82.11	83.12	61.95	82.76	74.12	77.00	78.60	76.82
Orthogonal Adaptation	80.20	85.40	80.39	81.63	79.80	61.67	76.71	68.00	79.76	74.62	61.00	79.14	77.92	72.85	65.06	74.94
Continual Diffusion	85.82	81.05	81.34	83.19	80.01	58.97	76.53	67.98	81.65	77.56	60.93	80.95	77.52	74.57	67.42	75.70
CNS	78.93	80.84	79.20	83.81	75.88	60.73	76.59	67.91	81.17	78.30	62.71	80.18	73.79	72.99	77.02	75.34
C2PVAR-S (Ours)	86.83	82.57	90.85	88.65	89.16	66.56	87.49	77.04	84.54	84.98	76.12	81.97	84.65	82.70	80.14	**82.95**

Table 10: Quantitative comparisons of single-concept customization across 15 tasks (CLIP-T metric).

Methods	Single Concept (CLIP-T↑)															
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	Avg.
ARBooth	30.22	33.54	31.22	33.23	30.41	28.75	30.90	28.22	29.81	28.25	31.09	33.21	34.97	33.82	32.33	31.33
LoRA	29.83	33.26	32.31	32.77	31.23	31.43	32.46	30.71	30.91	30.21	31.32	34.85	35.42	34.13	32.81	32.24
LWF	26.01	30.87	27.55	27.89	26.81	29.06	27.15	28.01	24.28	25.13	24.41	30.25	30.77	27.31	30.83	27.76
MAS	30.56	34.39	31.41	34.00	30.64	33.02	31.61	32.95	29.48	29.97	31.63	33.24	37.90	35.58	33.42	32.65
Orthogonal Adaptation	27.67	31.72	29.13	32.39	29.05	33.10	30.15	30.44	28.60	28.29	28.33	30.07	33.29	30.74	30.25	30.22
Continual Diffusion	30.36	34.37	32.07	32.55	30.76	32.70	32.60	31.60	29.86	30.65	29.91	30.52	34.54	32.54	30.88	31.73
CNS	31.22	33.29	32.27	33.43	31.35	34.32	32.32	33.09	29.25	30.79	32.68	34.23	38.04	34.83	33.97	33.01
C2PVAR-S (Ours)	31.04	33.50	28.66	31.99	28.04	32.44	27.57	32.06	27.98	28.45	31.04	31.31	31.75	30.74	32.98	30.64

A.6 Additional qualitative comparisons

We have added additional qualitative comparisons in this section, as shown in Figure 6. Our proposed method inherently facilitates the seamless transfer of personalized style concepts to personalized object concepts. Moreover, the image generation process merely requires the inclusion of both the style special token and the object special token within the text prompt, entirely circumventing the need for any auxiliary multi-concept synthesis strategies. This capability arises because distinct concept tokens can directly activate their respective concept parameters within the model, thereby enabling the generation of the specific object accurately rendered in the designated style.

A.7 Pseudocode of the framework

In this section, we provide the detailed pseudocode for our proposed framework. Algorithm 1 outlines the continual fine-tuning process with C2PVAR-S, which fundamentally resolves catastrophic forgetting. Algorithm 2 details the multi-concept synthesis inference process.

Table 11: Quantitative comparisons of single-concept customization across 8 tasks. We report the average DINO, CLIP-I, and CLIP-T metrics with standard deviations.

Methods	Avg. DINO↑	Avg. CLIP-I↑	Avg. CLIP-T↑
ARBooth	51.68 ± 0.82	75.99 ± 0.37	31.28 ± 0.14
LoRA	50.60 ± 0.65	76.63 ± 0.39	31.48 ± 0.53
LWF	59.05 ± 0.44	80.78 ± 0.80	29.01 ± 0.30
MAS	57.21 ± 0.38	77.52 ± 0.25	$\mathbf{32.65 \pm 0.15}$
Orthogonal Adaptation	58.05 ± 0.29	78.88 ± 0.18	31.58 ± 0.58
Continual Diffusion	60.04 ± 0.19	78.68 ± 0.14	31.85 ± 0.07
CNS	45.31 ± 0.26	75.61 ± 0.58	32.58 ± 0.12
C2PVAR-S (Ours)	$\mathbf{68.17 \pm 0.47}$	$\mathbf{82.71 \pm 0.07}$	31.14 ± 0.14

Algorithm 1 C2PVAR-S

Input: Pre-trained Infinity model weights θ, Task datasets $\mathcal{T} = \{\mathcal{D}_1, \ldots, \mathcal{D}_T\}$
Hyper-parameters: learning rates η, selection ratio p , update interval e, penalty λ , iterations N.

Initialize $M_{<t} \leftarrow \mathbf{0}$, $\theta_{old} \leftarrow \theta$
for each task $t = 1$ to T **do**
  $M_t \leftarrow \mathbf{0}$, $k \leftarrow 0$
  for epoch $ep = 0$ to E **do**
    if $ep \bmod e == 0$ **then**
      $g \leftarrow \nabla_\theta \mathcal{L}_{w-var}(\theta; \mathcal{D}_t)$ {Compute gradients of CA layers }
      $M_t^{ke} \leftarrow |g| \geq \text{Percentile}(|g|, 100 - p)$
      $M_t \leftarrow M_t \vee M_t^{ke}$ {Update task mask}
      $k \leftarrow k + 1$
    end if
    for iteration $i = 1$ to N **do**
      $M_{reg} \leftarrow M_{<t} \wedge M_t^{ke}$ {Detect overlapping parameters }
      $\mathcal{L}_{total} \leftarrow \mathcal{L}_{w-var}(\theta) + \lambda \left\| M_{reg} \odot (\theta - \theta_{old}) \right\|_2^2$ {Eq.9}
      $\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L}_{total}$
    end for
  end for
  $M_{<t} \leftarrow M_{<t} \vee M_t$, $\theta_{old} \leftarrow \theta$ {Update global mask}
end for
return θ_T

Algorithm 2 C2PVAR-M

Input: Global prompt y_{global}, local prompts and bboxes $\{(y_i, b_i)\}_{i=1}^B$, Scales S
Hyper-parameters: Blending factor α, intervention threshold s_{start}.

Initialize $B + 1$ parallel branches and compute background mask $b_G \leftarrow \mathbf{1} - \bigvee_{i=1}^B b_i$
for each scale $s = 1$ to S **do**
  if $s \geq s_{start}$ **then**
    for each VAR transformer block **do**
      $f_G \leftarrow \text{transformer}(y_{global})$
      for each local branch $i = 1$ to B **do**
        $f_i \leftarrow \text{transformer}(y_i)$
        $f_i^F \leftarrow b_i \odot f_i + (\mathbf{1} - b_i) \odot f_G$ {Eq.11 }
      end for
    end for
    Obtain L_G and L_i after all VAR transformer blocks
    $L_M \leftarrow b_G \odot L_G + \sum_{i=1}^B (b_i \odot (\alpha L_G + (1 - \alpha) L_i))$ {Eq.13 }
    Synchronize all branches with L_M
  else
    Execute standard VAR next-scale prediction
  end if
  Sample tokens for scale s
end for
return Final generated image X

A.8 Additional ablation study

Sequence. We also conduct an ablation study regarding the sequence of concepts in continual personalization. Specifically, order 2 is defined as (ducktoy, dog2, cat, drawing, teddybear, cat2, dog, inkpainting), order 3 as (ducktoy, drawing, teddybear, cat, dog2, cat2, inkpainting, dog), and order 4 as (dog, drawing, ducktoy, cat, dog2, inkpainting, teddybear, cat2). As presented in the Table

Table 12: Ablation study on task order in continual learning. We additionally compare the average performance across eight concepts under three different learning sequences. Metrics are reported as the average scores over eight tasks.

Task Order	Avg. DINO ↑	Avg. CLIP-I ↑	Avg. CLIP-T ↑
Order 1 (Default)	68.17	82.71	31.14
Order 2	67.16	81.40	31.62
Order 3	68.59	83.14	30.93
Order 4	68.18	82.53	30.93

Table 13: Ablation study on the cross-task conflict regularization coefficient λ. Metrics are reported as the average scores over eight tasks.

Coefficient (λ)	Avg. DINO ↑	Avg. CLIP-I ↑	Avg. CLIP-T ↑
$\lambda = 0.001$	65.18	80.46	31.78
$\lambda = 0.1$	67.54	82.68	30.87
$\lambda = 1.0$ (Default)	**68.17**	**82.71**	31.14
$\lambda = 1000.0$	67.67	82.26	**31.29**

12, our experiments fully demonstrate that alterations in the training sequence exert a negligible impact on the model's continual learning capabilities, thereby substantiating the strong robustness of our proposed C2PVAR-S continual learning method.

Regularization coefficient. We evaluate the impact of varying λ on the average performance across all customized concepts. $\lambda = 1.0$ is the default setting used in our framework, which achieves the optimal overall performance taking all three metrics into consideration.

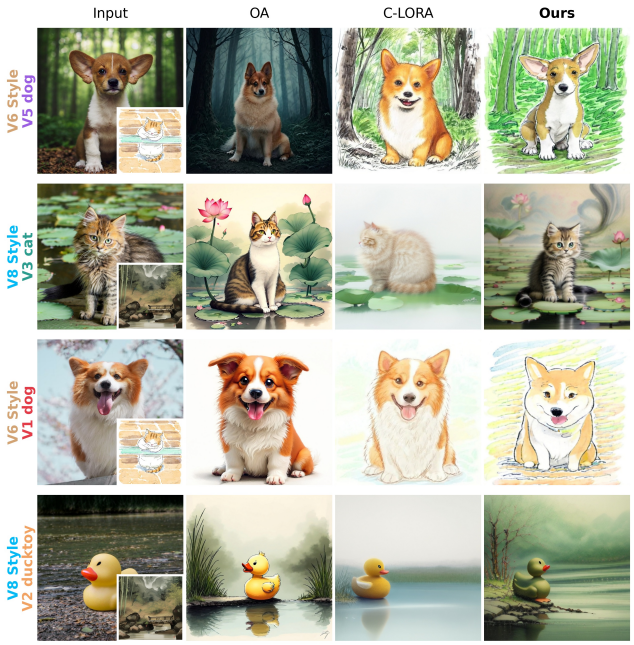


Figure 6: Qualitative comparison of custom style transfer

Table 14: Comparisons of computational resources. Memory requirements for GPU/CPU and computation time for customization indicate the additional costs for fusing concept weights previously learned.

Methods	**Computational Resources**	
	Memory(MB)↓	**Time(s)↓**
ARBooth	0/0	0
LoRA	137/52	1.75
LWF	0/0	0
MAS	0/0	0
Orthogonal Adaptation	1278/35	11.44
Continual Diffusion	3303/60	12.51
CNS	0/0	0
C2PVAR-S (Ours)	**0/0**	**0**