Personalized Residuals for Concept-Driven Text-to-Image Generation

Cusuh Ham*
Georgia Institute of Technology
cusuh@gatech.edu

Matthew Fisher
Adobe Research
matfishe@adobe.com

James Hays
Georgia Institute of Technology
hays@gatech.edu

Nicholas Kolkin
Adobe Research
kolkin@adobe.com

Yuchen Liu
Adobe Research
yuliu@adobe.com

Richard Zhang
Adobe Research
rizhang@adobe.com

Tobias Hinz
Adobe Research
thinz@adobe.com

Abstract

We present personalized residuals and localized attention-guided sampling for efficient concept-driven generation using text-to-image diffusion models. Our method first represents concepts by freezing the weights of a pretrained text-conditioned diffusion model and learning low-rank residuals for a small subset of the model's layers. The residual-based approach then directly enables application of our proposed sampling technique, which applies the learned residuals only in areas where the concept is localized via cross-attention and applies the original diffusion weights in all other regions. Localized sampling therefore combines the learned identity of the concept with the existing generative prior of the underlying diffusion model. We show that personalized residuals effectively capture the identity of a concept in ∼3 minutes on a single GPU without the use of regularization images and with fewer parameters than previous models, and localized sampling allows using the original model as strong prior for large parts of the image.

1. Introduction

Large-scale text-to-image diffusion models have demonstrated the ability to generate high-quality images that follow the constraints of the input text [21, 22, 26]. However, these models do not inherently encode any information about the *identity* of a specific concept, thus limiting the control over specifying a particular instance to appear in the generated image. To address this, recent approaches propose techniques to *personalize* these models such that they can generate specific concepts in novel environments and styles.

Given a set of images depicting the desired concept,

personalization approaches differ in which parameters they train and whether they are specific to a single concept (i.e., they need to be separately trained for each new concept) or can generalize to new concepts without retraining. To enable personalization of arbitrary concepts, one can finetune the model's parameters [24] or its inputs [7] directly such that it can reconstruct the training data. These approaches can be applied to any kind of concepts, but the finetuning needs to be done on a per-concept basis and different parameters need to be stored for each. Other approaches train an encoder specific to a particular domain (e.g., faces) and finetune the diffusion model once to use the encoder's embeddings to reconstruct specific concepts within that domain [8, 25, 33]. The advantage of the latter approach is that it does not require retraining for every concept and can instead be used to instantly generate new concepts from the given domain. However, this approach is limited to a single domain and requires a large dataset to train the encoder.

Our approach follows the former setting, i.e., it finetunes the model's parameters for each concept so that there are no constraints on the domain (see Figure 1 for examples using our proposed method). The main challenges of open-domain approaches is the need for regularization to mitigate forgetting of concepts learned in the model's original training, and the computational overhead in finetuning a new set of parameters for each concept. The most common regularization approach is to use images from the same domain as the target concept with the reference images during the finetuning of parameters. The choice of regularization images affects the quality of the final outputs and, as such, is usually model-, training-, and sometimes even concept-dependent. Finally, to address the large overhead of finetuning a whole new model for each concept, many approaches only finetune a subset of parameters (e.g., attention layers weights [16]) or the input to the text-to-image model (e.g., the text embedding representing a specific concept [7]).

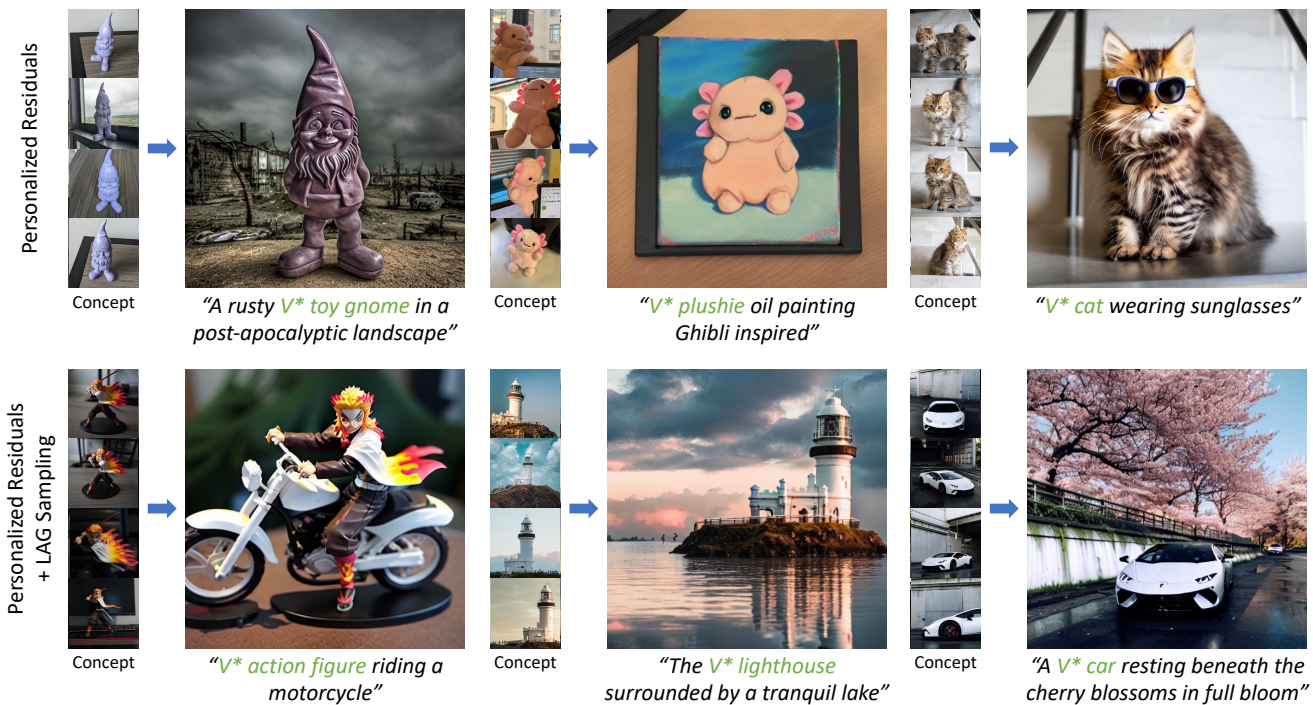Our approach further reduces the number of learnable

---

Figure 1. (Top) Given a set of reference images, we learn *personalized residuals* for a subset of a pretrained diffusion model's weights for efficient concept-driven text-to-image generation. (Bottom) The residuals can be combined with our proposed *localized attention-guided (LAG) sampling*, which leverages the cross-attention maps from the diffusion models to localize the application of the residuals and uses the original, unchanged, diffusion model for generating everything else.

parameters and does not rely on regularization images. While most approaches focus on finetuning the key and value weights of the cross-attention layers, we instead predict a low-rank residual [14] to the weights of the output projection conv layer after each cross-attention layer. This allows us to finetune even fewer parameters (about ∼0.1% of the base model) than previous approaches. Furthermore, we find that this approach does not require any regularization images which makes our approach both simpler, since we do not need to find appropriate strategies to obtain regularization images, and faster, since we do not need additional training iterations for learning from the regularization images. We also show that the choice of macro class for personalizing a given image affects the performance, e.g., using "car" instead of "Lamborghini" as the macro class in Figure 1 affects the quality of the outcome (see supplementary). Based on this, removing the need for regularization images removes an additional dependency and decreases the need for manual selections.

Additionally, many personalization approaches struggle to render specific backgrounds or add new objects often due to some degree of overfitting to the target concept. For these scenarios, we propose a novel localized attention-guided (LAG) sampling scheme, which allows us to use the finetuned residuals with the original model to generate the target concept and the rest of the image, respectively. To achieve this, we use the attention maps from the cross-attention layers of the diffusion model at each timestep to predict the location of the concept in the generated image and then apply the features, produced using the personalized residuals, only in the predicted region such that the rest of the image (e.g., background and other objects) is generated by the original model. Thus, we ensure that we do not lose the capability of generating specific backgrounds or unrelated objects due to overfitting. Furthermore, this sampling approach does not require any additional training or data, and does not increase sampling time as no additional model evaluations are needed.

We evaluate our approach and sampling technique on the CustomConcept101 dataset [16], which was specifically designed to evaluate personalization approaches. We use CLIP and DINO scores to evaluate the text-image alignment (i.e., how well the personalized model can generate the concept in novel scenes and environments) and identity preservation of the personalized model (i.e., how well it can generate the desired concept). We also perform a user study to evaluate human preference for text-image alignment and identity preservation. Our results show that our model performs on par or better compared to current state-of-the-art baselines while using significantly fewer parameters, not re-

lying on regularization images, and being faster to train.

To summarize, our key contributions are a novel and more efficient low-rank personalization approach for text-to-image diffusion models that works for arbitrary domains and concepts, uses fewer parameters than previous approaches, does not rely on regularization images and is, therefore, faster and simpler to train. We also introduce a novel *localized attention-guided (LAG) sampling* approach that allows us to flexibly combine the original pretrained and the finetuned model on the fly to generate different parts of the image, without increasing the sampling time and without requiring additional training or user inputs. Our user study and quantitative evaluations show that our method performs comparably or better than other baselines, and our proposed sampling approach can address challenges with certain types of recontextualization scenarios, such as background changes.

2. Related Work

2.1. Personalization of text-to-image models

The task of text-to-image personalization was proposed by [7], where a few example images of the given concept are used to finetune a "personalized" token embedding while all other parameters of the model frozen. Instead of trying to find an embedding within the existing text conditioning space to represent a concept, DreamBooth [24] finetunes the diffusion model's parameters to directly inject the concept into the learned prior, leading to better performance. Custom Diffusion [16] only finetunes the cross-attention weights in addition to the token embedding to achieve more efficient personalization compared to DreamBooth. Based on these works, other aim to improve the performance and efficiency of personalizing text-to-image models through approaches such as, but not limited to, learning multiple personalized tokens [5, 12], imposing constraints on the trainable parameters (e.g., key-locking [30], orthogonality [19], low-rank [28], singular values only [9]), training hypernetworks and domain-specific encoders [8, 17, 25, 33], and injecting of visual features [10, 32, 33].

2.2. Attention-guided text-to-image synthesis

Attention layers [31] have been shown to play an important role in the success of text-conditioned image synthesis using diffusion models. Recent works propose to manipulate attention maps from these layers for guided synthesis and editing. [4] modifies cross-attention values to guide the generation process so that the subjects specified in an input prompt appear and the attributes are associated to its corresponding subject. [1, 11] enable conditioning on a user-provided layout by guiding the localization of objects via cross-attention manipulation. Given an existing image and a prompt that describes the image, [6, 12] synthesize/edit im-

ages by manipulating the cross-attention map corresponding to the editing target. Similarly, [2] performs edits on existing images albeit through instructions and modifications within self-attention layers.

3. Approach

Our method consists of two components: 1) *Personalized residuals*, which encode the identity of a given concept through a set of learned offsets applied to a subset of weights within a pretrained text-to-image diffusion model, and 2) *Localized attention-guided (LAG) sampling*, which leverages attention maps to localize where the residuals are applied, essentially allowing a single image to be efficiently generated by leveraging both the base diffusion model and the personalized residuals.

3.1. Preliminaries

Diffusion models. Diffusion models [13] consist of a fixed forward noising process that gradually adds noise to an image, and a learned denoising process that iteratively removes noise to produce a valid image. The denoising process is learned through a U-Net [23] ϵ_θ, parameterized by θ, and is conditioned on an image x_t noised to timestep t, and t itself. Text guidance can be incorporated through conditioning on embeddings $c = \tau(y)$ of input prompts y from a text encoder τ, such as CLIP [20].

In this work, we leverage Stable Diffusion, a text-conditioned latent diffusion model (LDM) [22]. An LDM is a variant of a diffusion model that operates in the latent space of a variational autoencoder [15]. The encoder $\mathcal{E}$ embeds an input image x into a latent representation $z = \mathcal{E}(x)$ and a decoder $\mathcal{D}$ maps z back into pixel space $x' = \mathcal{D}(z)$. The diffusion portion of LDM operates on z and is trained using the following objective:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z \sim \mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta(z_t, t, \tau(y))\|_2^2 \right]. \quad (1)$$

Low rank adaptation (LoRA). Low rank adaptation (LoRA) [14] is an efficient method originally proposed for updating large language models through learned residuals instead of directly finetuning their parameters. For a given layer of the pretrained model with weight matrix $W_0 \in \mathbb{R}^{m \times n}$, LoRA learns two matrices A and B whose product forms a residual $\Delta W = AB \in \mathbb{R}^{m \times n}$, where $A \in \mathbb{R}^{m \times r}$, $B \in \mathbb{R}^{r \times n}$, and $r \ll \min(m, n)$ is the rank. The updated weight matrix is then defined as $W' = W_0 + \Delta W$. With small values of r, LoRA has been shown to significantly reduce the number of learnable parameters while retaining or even improving performance.

3.2. Learning residuals for capturing identity

The goal of personalizing text-to-image models is to faithfully capture the identity of a target concept while simulta-

neously avoiding overfitting so that the concept can be re-contextualized into new settings and configurations. Since concepts are often learned using only a few reference images, directly finetuning the weights of a very large generative model can easily lead to overfitting and/or overwriting unnecessary parts of the learned language prior. Instead we propose to use a LoRA-based approach to learn low-rank offsets for a small subset of the diffusion model weights which will represent the target concept. Thus, we are able to recover the full generative capacity of the original model by simply not applying the learned residuals at inference.

The diffusion model contains multiple transformer blocks, which consist of self- and cross-attention layers [31] with a 1×1 conv projection layer on either end (see Figure 2). While several approaches primarily target the cross-attention layers due to their learning of relationships between text and images, we choose to learn offsets for the output projection conv layers because these localized operations can capture finer details than the global operations of cross-attention.

We illustrate the process of learning personalized residuals in Figure 2. Given a pretrained text-to-image diffusion model containing L transformer blocks, we learn $\Delta W_i = A_i B_i \in \mathbb{R}^{m_i \times m_i}$ for the output projection layer $l_{\text{proj\_out,i}}$ with weight matrix $W_i \in \mathbb{R}^{m_i \times m_i \times 1}$ within each transformer block i, where $A_i \in \mathbb{R}^{m_i \times r_i}$ and $B_i \in \mathbb{R}^{r_i \times m_i}$. We reshape the residual such that $\Delta W_i \in \mathbb{R}^{m_i \times m_i \times 1}$ and add to the original weights W_i to produce $W_i' = W_i + \Delta W_i$. The ΔW_i's are updated using the original diffusion objective in Equation (1).

Similar to other works, we associate the concept with a unique identifier token (e.g., V∗), which is initialized using a rarely occurring token embedding. During training, we use the unique token and macro class of the concept in a fixed template for the prompt associated with each reference image (e.g., "a photo of a V∗ macro class"). Personalization approaches that involve direct updates to the diffusion model's weights are susceptible to overwriting parts of the existing generative prior with the new concept and thus explicitly require "prior preservation" through regularization images during training [16, 24]. Since our method does not directly update the diffusion model, we avoid this issue entirely and eliminate the burden on the user to determine an effective set of regularization images, which is not always straightforward. Additionally, the low-rank constraint on the residuals reduces the number of trainable parameters, making our method a simpler and more efficient approach for personalization.

3.3. Localized attention-guided sampling

With our residual-based personalization approach, we have additional flexibility in how the offsets are applied at inference. We introduce a new *localized attention-guided* (LAG)

sampling method to better combine a newly learned concept with the original generative prior of the diffusion model. As shown in Figure 2, within every transformer block of the diffusion model is a cross-attention layer, which aims to learn the correspondence between text tokens and image regions. Each cross-attention layer computes attention maps A_{y_i} for each token y_i in the prompt, indicating where the token will affect the generated image. The attention maps are produced using the following equation:

$$A(Q, K) = \text{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big), \qquad (2)$$

where $Q = W^Q x$ is the query, $K = W^K y$ is the key, and d_k is the dimension of the query and key.

Given the indices $\mathcal{C}$ of the unique identifier and macro class tokens specifying the concept (e.g., "V∗" and "dog"), we sum the values of the corresponding attention maps $A_{i,\mathcal{C}} = \sum_{j \in \mathcal{C}} A_j$ in transformer block i, and then binarize using its median value to get $M_i = \text{binarize}(A_{i,\mathcal{C}})$. Finally, we compute the output feature $\hat{f}_i$ of each transformer block i as:

$$\hat{f}_i = (1 - M_i) \otimes f_i + M_i \otimes f_i', \qquad (3)$$

where $f_i = W_i x$ is the feature produced using the original conv weight W_i, and $f_i' = W_i' x$ is the feature produced using the updated weight from the personalized residual $W_i' = W_i + \Delta W_i$. Thus, the identity represented through the personalized residuals is only being applied in the regions corresponding to the target concept, and the remaining regions are generated by the original diffusion model. The proposed LAG sampling technique is visualized in Figure 4.

While there exist personalization works using attention guidance (e.g., [10, 33]), they often rely on object masks and/or additional losses at train time to focus on the relevant object location in the reference images, whereas manually-provided object masks or specific training are not needed to enable LAG. Additionally, LAG sampling explicitly merges the features of two layers (personalized/finetuned and original/non-finetuned) on-the-fly based on the cross-attention maps obtained during inference and has negligible impact on the sampling speed. In contrast, other synthesis/editing works (see Section 2.2) use cross-attention values to up- or down-weight the influence of specific tokens at specific image locations.

LAG sampling can be beneficial in scenarios where the learned residuals overfit to the reference images and have not effectively disentangled the target concept from the background, which can occur as a consequence of ambiguities of the target concept given the reference images or model biases (e.g., furniture often photographed indoors). By leveraging the attention maps from the tokens denoting
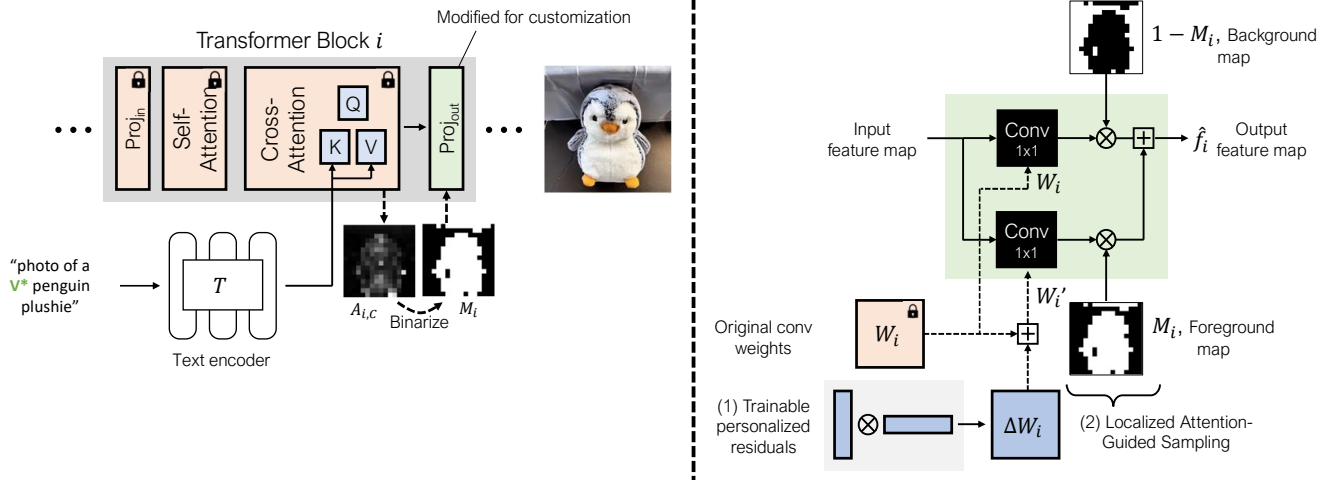
Figure 2. Overview of our proposed work. (1) *Personalized residuals*: We learn low-rank residuals for the output projection layer within each transformer block in the diffusion model. The residuals contain relatively few parameters, are fast to train, and do not require any regularization images during training. (2) *Localized attention-guided sampling*: We optionally apply the personalized residuals only in the areas that the cross-attention layers have localized the concept via predicted attention maps. Thus, we can combine the newly learned concept with the original generative prior of the base diffusion model within a single image.

the concept, we can localize the residuals so that they do not affect the background, which can instead be generated using the base model.

4. Experiments

In this section, we describe our experimental setup and evaluation protocols, and visualize examples using the proposed personalized residuals with and without localized attention-guided sampling.

4.1. Training details

We build upon Stable Diffusion v1.4 [22]. For each transformer block i, we compute the rank r_i for its output projection convolution layer with weight matrix $W_i \in \mathbb{R}^{m_i \times m_i \times 1}$ as $r_i = 0.05 m_i$, totalling 1.2M trainable parameters ($\sim$0.1% of Stable Diffusion). Each of the low-rank matrices are randomly initialized. We train our method for 150 iterations with a batch size of 4 and learning rate of 1.0e-3 on 1 A100 GPU ($\sim$3 minutes) across all experiments.

4.2. Baselines

We focus on comparisons to open-domain (i.e., does not require encoders limited to a single given domain) approaches with publicly available code. Specifically, we compare our method against four baselines: Textual Inversion [7], DreamBooth [24], Custom Diffusion [16], and ViCo [10]. Textual Inversion freezes the entire diffusion model and optimizes only the unique identifier token V* for each concept. ViCo optimizes V* as well as newly added cross-attention layers to the diffusion model to incorporate visual information from the reference images while keeping the rest of the

model frozen. DreamBooth finetunes the entire diffusion model using the reference images and a set of regularization images, which are generated within the same domain as the target concept using the original model. While DreamBooth was originally proposed using Imagen [26], we use an open-source version built on Stable Diffusion[1]. Custom Diffusion finetunes only the key and value weights of the cross-attention layers in addition to the identifier token embedding, and uses a set of real regularization images sampled from LAION-400M [27].

We use the recommended settings described by each paper. For Textual Inversion and ViCo, which initialize the identifier token embedding to a single word that best represents the concept, we use our best discretion to pick a word most similar to the macro class given by CustomConcept101.

4.3. Evaluation metrics

Following the protocol described in [16], we leverage the CustomConcept101 dataset, consisting of 101 concepts across 16 broader categories. For every concept we generate 50 samples for each of the 20 prompts given by the dataset. We use DDIM sampling [29] with $N = 50$ steps, $\eta = 0.0$, and a guidance scale of 6.0 for all methods. We set the same random seed for sampling across each method so that the "choice" of starting noise does not impact the results. Results of our method with LAG sampling are explicitly labeled as such.

We evaluate each method for text alignment and image

---

[1]https://github.com/XavierXiao/Dreambooth-Stable-Diffusion

Table 1. Quantitative evaluations for text and image alignment using the similarity of CLIP and DINO features. We report the number of parameters for each method in addition to scores from the base Stable Diffusion model, which is not trained for personalization, for reference.

Method	# params	CLIP text	CLIP image	DINO image
Textual Inversion	768	0.6150	0.7259	0.4700
ViCo	51.3M	0.7403	0.7111	0.4678
DreamBooth	983M	0.7536	0.7424	0.5212
Custom Diffusion	19M	**0.7664**	0.7074	0.4669
Ours	1.2M	0.7193	**0.7594**	**0.5671**
Ours w/ LAG sampling	1.2M	0.7220	0.7424	0.5411
Stable Diffusion	983M	0.8126	0.6207	0.2920

Table 2. Human preference evaluations for text and image alignment through Amazon Mechanical Turk. We perform bootstrap resampling over the 1250 responses collected for each task.

Ours vs.	Textual Inversion	ViCo	DreamBooth	Custom Diffusion	Ours w/ LAG
Text	**81.85** ±4.15%	37.40 ±7.46%	41.34 ±5.08%	**50.99** ±5.46%	**58.57** ±6.34%
Image	**61.96** ±4.76%	**62.11** ±5.80%	**51.33** ±4.65%	**63.27** ±5.59%	26.26 ±4.91%

alignment. *Text alignment* is measured as the similarity between the CLIP [20] text feature of the input prompt and the CLIP image feature of the resulting generated image. *Image alignment* is measured as the similarity between image features from either CLIP or DINO [3] of the reference images and corresponding generated images.

Additionally, we evaluate both text and image alignment using human evaluations through user studies on Amazon Mechanical Turk (AMT). For each text alignment case, we display a text prompt and a pair of corresponding generated images, and ask users *"Which image is more consistent with the given text prompt?"*. For each image alignment case, we display 3 reference images for a concept and a pair of corresponding generated images, and ask *"Which image better preserves the identity of the subject in the provided reference images?"*. For both studies, each pair of images contains one from {Textual Inversion, ViCo, DreamBooth, Custom Diffusion, Ours w/ LAG sampling} and one from ours with normal DDIM sampling. Users can select either image or neither (*"Not sure"*).

4.4. Results

We visualize samples generated by each method for various types of prompts in Figure 3. Textual Inversion fails to reliably capture the concept's identity and/or the prompt whereas all other methods, including ours, are able to better preserve the concept's identity while also adhering to the prompt. We highlight that our method is able to achieve these results while having significantly fewer learnable parameters and requiring less training time compared to ViCo, DreamBooth, and Custom Diffusion, as well as not leveraging regularization images.

We compare examples using our proposed personalized residuals with and without localized attention-guided sampling in Figure 4. We illustrate how LAG sampling affects the output image by using the same starting noise map z_T to sample each pair of {w/o LAG, w/ LAG} images. We highlight scenarios where LAG sampling performs better than normal sampling in Figure 4a and vice versa in Figure 4b.

Quantitative evaluations for text and image alignment using CLIP and DINO are shown in Table 1. We include results using the original Stable Diffusion model, which has no notion of any of the concepts, for reference. We show that our method performs similarly with and without LAG sampling averaged across the whole dataset, demonstrating higher image alignment and slightly lower text alignment than the more computationally-heavy baselines.

However, as seen by the results of 1250 responses collected through AMT user studies for both text and image alignment in Table 2, we show that the CLIP text alignment scores do not necessarily correlate to human preference. We observe that our method performs similarly to Custom Diffusion for text alignment, which was assigned the highest CLIP text score, and outperforms all baselines for image alignment. Again, we note that our method achieves similar performance to the better performing baselines while being significantly more computationally efficient. We also compare our method with and without LAG sampling in the user studies and show that LAG is preferred for image alignment but not text alignment. Further analysis comparing the two sampling approaches can be found in the supplementary.

We also train and evaluate our method using CLIP similarity to select the "most representative" macro class among the 117k nouns in WordNet [18] for each concept. In Table 4, we show that using the WordNet macro class leads to further improvements in image alignment while decreasing text alignment, the latter of which may not necessarily reflect human preference as previously demonstrated. See the supplementary for additional discussions.

Ablation studies. We perform ablation studies on changing the targets for where the residuals are applied, removing the macro class from the prompt, including regularization images (sampled from LAION) during training, updating the concept identifier token embedding $V\ast$, and varying the rank of the residuals. Results are shown in Table 3 (see Table 5 for results on changing the rank).

We show that changing where the residuals are applied to either the key and value weights of the cross-attention layers (like Custom Diffusion) or the input projection conv layer (rather than the output) slightly decreases the scores across all three metrics compared to our proposed approach. We hypothesize that the output projection layer achieves noticeably higher identity preservation because it refines the feature map at the end of each block. Additionally, learning
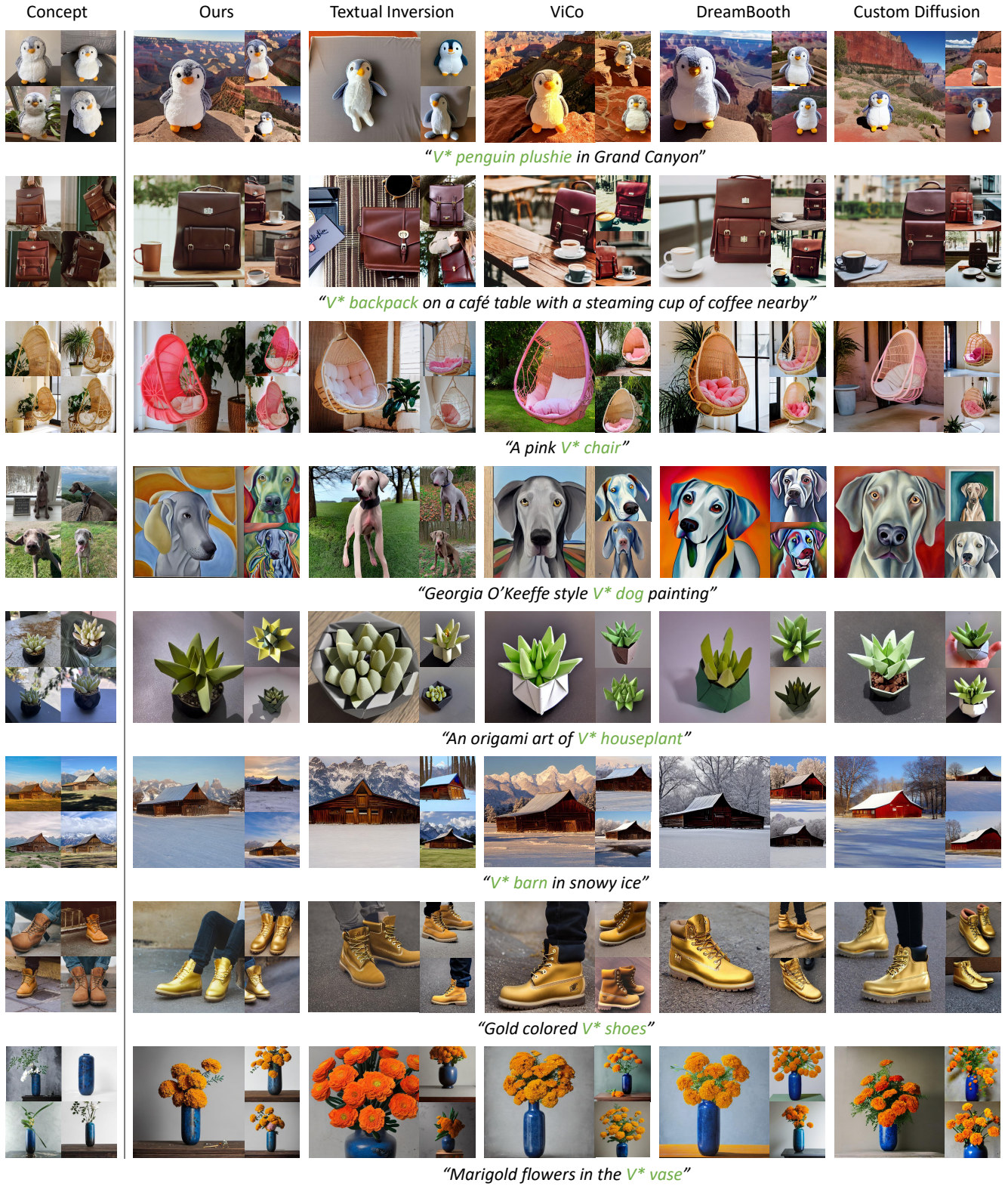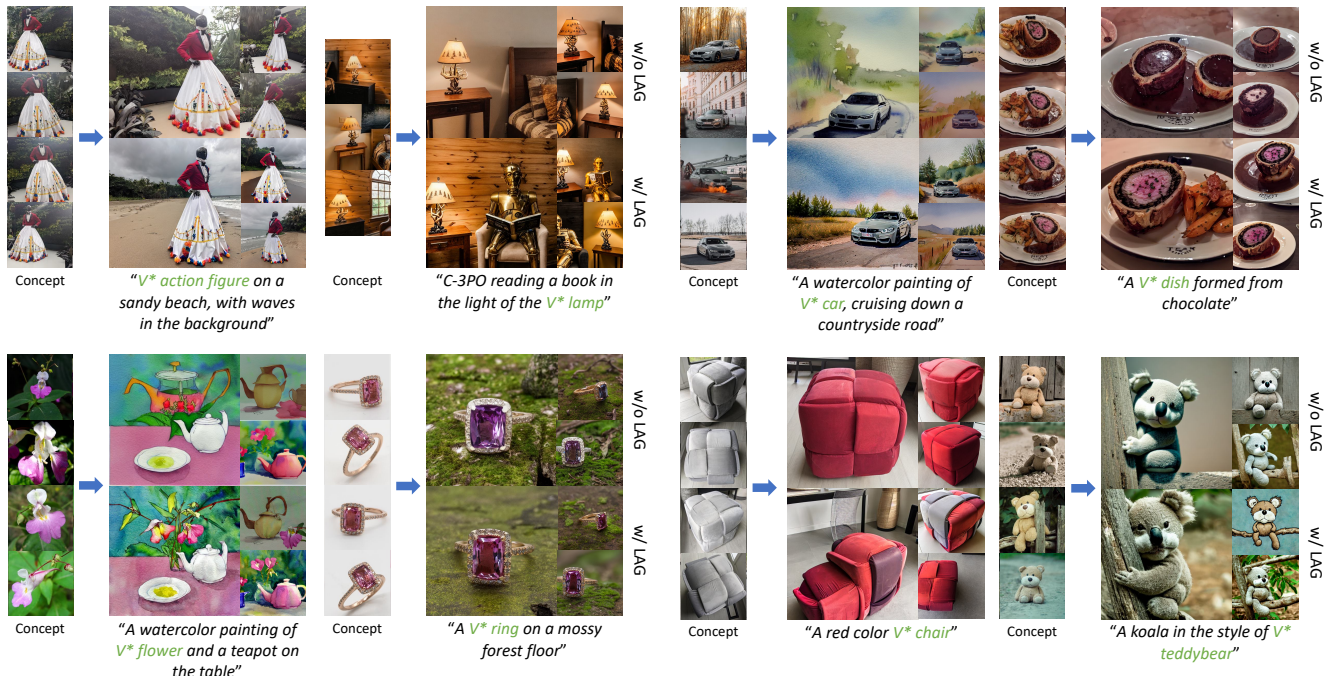
Figure 3. Qualitative comparison of our proposed approach with the baselines.

(a) Examples where LAG produces results that are better aligned with the concept and prompt.

(b) Examples where normal sampling produces results that are better aligned with the concept and prompt.

Figure 4. Comparison of image generated with and without LAG sampling. We use the same starting noise map to generate corresponding pairs of images to directly visualize how LAG sampling affects the output image.

Table 3. We evaluate our method using two different targets for the residuals and altering various training settings.

Method	CLIP text	CLIP image	DINO image
KV weights	0.7172	0.7508	0.5353
$l_{\text{proj\_in}}$ weights	0.7136	0.7460	0.5333
KV + $l_{\text{proj\_out}}$ weights	0.7049	0.7733	0.5868
KV + $l_{\text{proj\_in}}$ + $l_{\text{proj\_out}}$ weights	0.6739	0.7870	0.6040
w/o macro class	0.6605	0.6521	0.3798
w/ reg images	0.7204	0.6771	0.3830
Update token embedding	0.6673	0.8000	0.6194
Ours	0.7193	0.7594	0.5671

residuals for multiple layers simultaneously leads to overfitting to the reference images as demonstrated by the higher image alignment scores and lower text alignment.

Omitting the macro class leads to significant drops across all metrics, demonstrating that the additional information is useful to our method for knowing what within the reference images is important to model. Similar to the effect of using regularization images for DreamBooth and Custom Diffusion, regularization images slightly improves text alignment but decreases image alignment. On the other hand, updating the token embedding for V* leads to overfitting as shown by the increase in image alignment and decrease in text alignment.

5. Conclusion

We introduce personalized residuals, a method for concept-driven synthesis using text-to-image diffusion models. Previous approaches to personalization are often slow to train, have high computational demands, require regularization images, and/or have difficulty recontextualizing the target concept. Through our proposed LoRA-based approach that learns a small set of residuals to represent the identity of a concept, we reduce the number of learnable parameters and training time and remove the reliance on domain regularization while maintaining flexibility with editing. We also introduce localized attention-guided sampling which applies the personalized residuals only in regions where the concept is localized via the cross-attention mechanism. We evaluate our method across several metrics to show that we are able to efficiently enable personalization.

Limitations and future work. We show that localized sampling is not always the best choice (e.g., changing the color of a concept) and relies on the cross-attention layers to produce high-quality attention maps, which is not always the case. Our approach can be sensitive to the choice of macro class and inherits the pretrained model's biases and limitations, such as mixing up the relationship between attributes in the prompt. Finally, we leave multi-concept generation through LAG sampling as future work.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 3

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3

[3] 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*, pages 9650–9660, 2021. 6

[4] 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. 3

[5] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022. 3

[6] Dave Epstein, Allan Jabri, Ben Poole, Alexei A Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *arXiv preprint arXiv:2306.00986*, 2023. 3

[7] 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. 1, 3, 5

[8] 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. 1, 3

[9] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 3

[10] Shaozhe Hao, Kai Han, Shihao Zhao, and Kwan-Yee K Wong. Vico: Detail-preserving visual condition for personalized text-to-image generation. *arXiv preprint arXiv:2306.00971*, 2023. 3, 4, 5

[11] Yutong He, Ruslan Salakhutdinov, and J Zico Kolter. Localized text-to-image generation for free via cross attention control. *arXiv preprint arXiv:2306.14636*, 2023. 3

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 3

[15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2014. 3

[16] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1, 2, 3, 4, 5

[17] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 3

[18] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 6

[19] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *arXiv preprint arXiv:2306.07280*, 2023. 3

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[21] 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. 1

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3, 5

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[24] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1, 3, 4, 5

[25] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 1, 3

[26] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 5

[27] 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. 5

[28] 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. 3

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[30] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3, 1

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4

[32] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 3

[33] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 1, 3, 4

Personalized Residuals for Concept-Driven Text-to-Image Generation
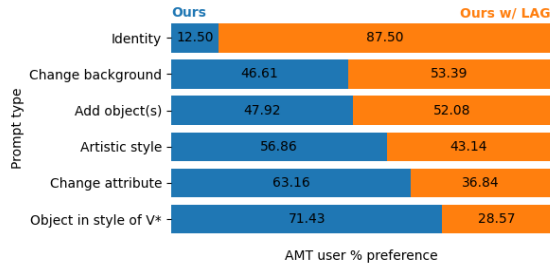
Supplementary Material

Figure 5. AMT text alignment scores per prompt type.

6. Additional experimental results

We explore the difference in normal and LAG sampling by using ChatGPT to categorize each prompt into {*add object(s)*, *artistic style*, *change attribute*, *change background*, *identity*, *object in style of* $V\star$}. We note that a prompt may fall into multiple categories, but we only use one as determined by ChatGPT. We split the AMT evaluations for text alignment by category in Figure 5. We observe that LAG sampling performs best for *identity*, *change background*, and *add object(s)*, which are tasks in which the target object is somewhat independent of the rest of the image. Tasks that require modifying the target (*artistic style*, *change attribute*, *object in style of* $V\star$) perform better with normal DDIM sampling.

In Figures 6 to 11 we directly compare examples from each of the six prompt categories using the two sampling methods by generating corresponding pairs using the same starting noise maps. Additional qualitative samples can be found in Figures 12 and 13.

We plot CLIP/DINO image alignment scores against CLIP text scores, averaged across concepts within the the 16 categories of CustomConcept101, for each method from Section 4.

Additionally, we compare our method to an unofficial implementation[2] of Perfusion [30] (an official version is not publicly available). We followed the experimental setup and hyperparameter values described by the original authors, but note that we were unable to reproduce the quality of the results shown in the paper: CLIP text 0.6879, CLIP image 0.5669, DINO image 0.2228.

7. Effect of macro class choice

For each concept in CustomConcept101, we compute the mean CLIP image embedding of its reference images and

---

[2]

Table 4. We compute the nearest neighbor (NN) in CLIP embedding space for each concept among all WordNet nouns. We compare our method using different combinations of macro classes during training and sampling.

Macro class choice		CLIP text	CLIP image	DINO image
Training	Sampling			
CustomConcept101	CustomConcept101	0.7193	0.7594	0.5671
	WordNet NN	0.7155	0.7594	0.5671
WordNet NN	CustomConcept101	0.6626	0.7798	0.5904
	WordNet NN	0.6869	0.7798	0.5904

calculate the cosine similarity against the CLIP text embedding for each of the 117k nouns within WordNet. We train our method and/or sample using the WordNet noun with the highest similarity and compare with using the provided macro class from CustomConcept101 during training and/or sampling in Table 4. We observe that using the WordNet nearest neighbor as the macro class leads to higher image alignment and lower text alignment compared to the CustomConcept101-provided macro class.

Selecting the "best" macro class for concepts can be challenging and given that it can lead to noticeable changes in alignment metrics, an automatic heuristic for choosing a suitable macro class would be helpful to users. We leave the designing of such a heuristic as future work.

8. Ablation study: rank value

Table 5. Quantitative evaluations for varying the rank of the learned residuals. m_i is the dimension of the weight of the projection layer in transformer block i.

Rank	CLIP text	CLIP image	DINO image
1	0.7398	0.6809	0.4148
8	0.7054	0.7402	0.5239
16	0.6926	0.7573	0.5513
32	0.6832	0.7701	0.5713
64	0.6704	0.7798	0.5865
128	0.6544	0.7938	0.6053
$0.025m_i$	0.6889	0.7622	0.5595
Ours ($0.05m_i$)	0.7193	0.7594	0.5671

We evaluate different values for the rank of the learned residuals in Table 5 and observe that text alignment is inversely proportional to the rank and image alignment is directly proportional. Since the dimensions of the conv weight matrix varies across the transformer blocks within the U-Net, we believe that calculating the rank with respect to the dimensions is the better approach over setting a fixed value across all layers, which is empirically validated by the results with our proposed formula achieving a better balance of image and text alignment.

Figure 6. Samples for *add object(s)* prompts using personalized residuals with and without LAG sampling where corresponding pairs are generated using the same input noise map.
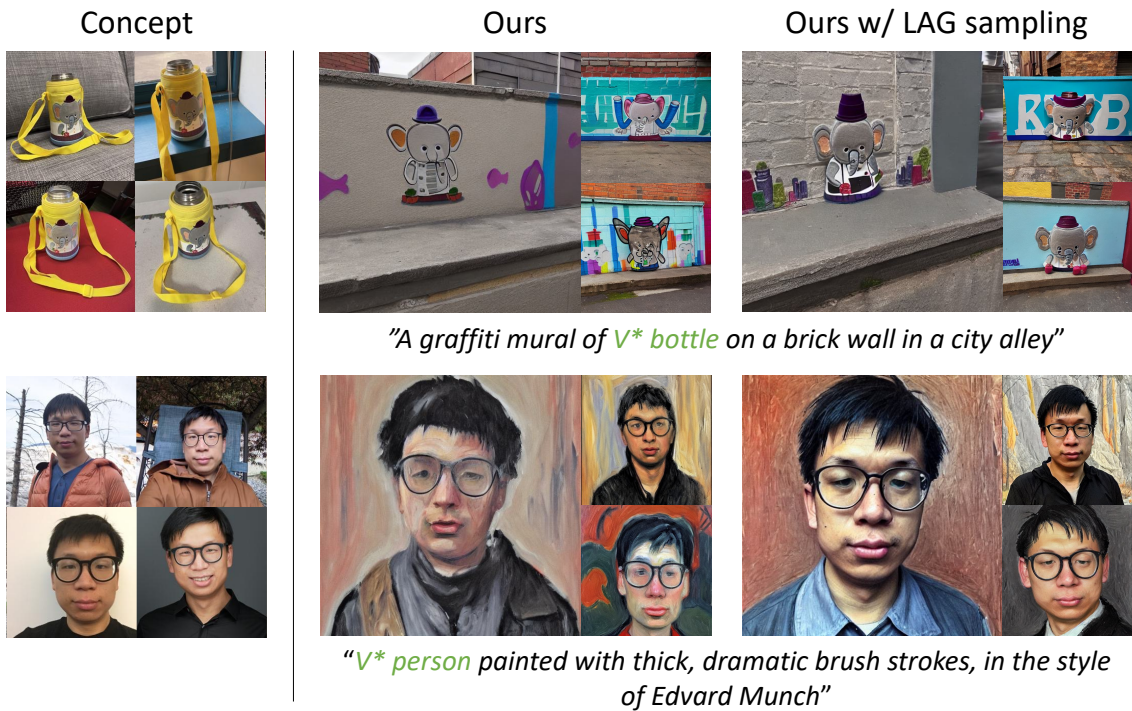

Figure 7. Samples for *artistic style* prompts using personalized residuals with and without LAG sampling where corresponding pairs are generated using the same input noise map.
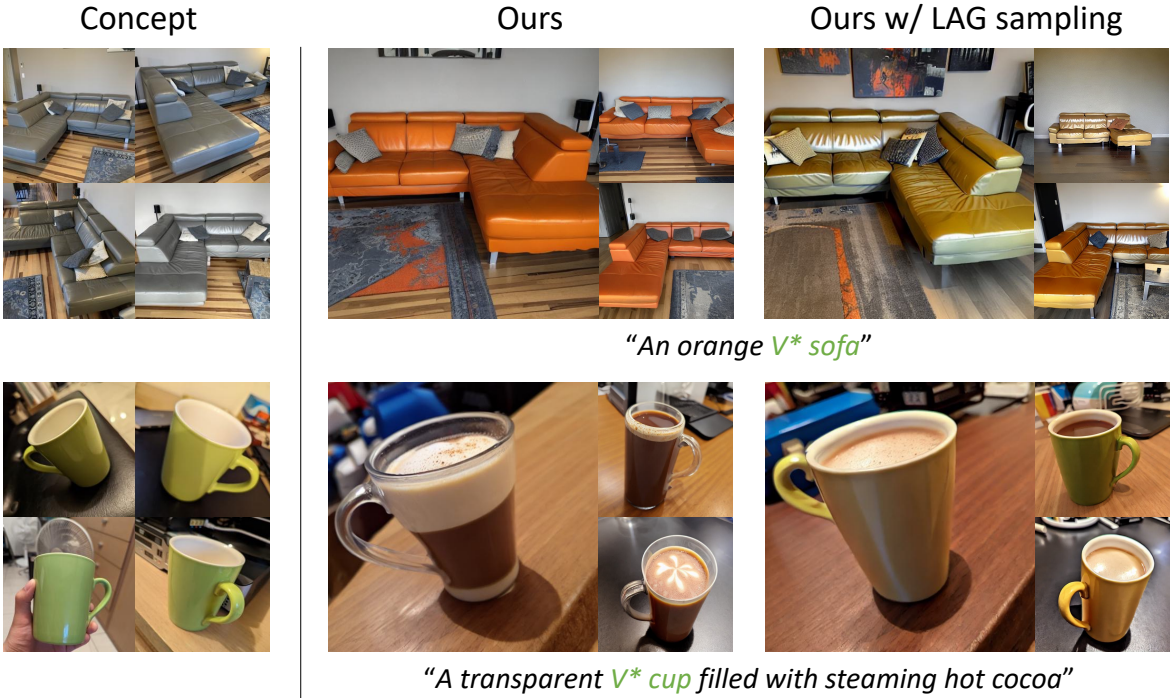
Figure 8. Samples for *change attribute* prompts using personalized residuals with and without LAG sampling where corresponding pairs are generated using the same input noise map.

Figure 9. Samples for *change background* prompts using personalized residuals with and without LAG sampling where corresponding pairs are generated using the same input noise map.

Concept	Ours	Ours w/ LAG sampling

"Photo of a V unicorn plushie"*

"Photo of a V sofa"*

Figure 10. Samples for *identity* prompts using personalized residuals with and without LAG sampling where corresponding pairs are generated using the same input noise map.
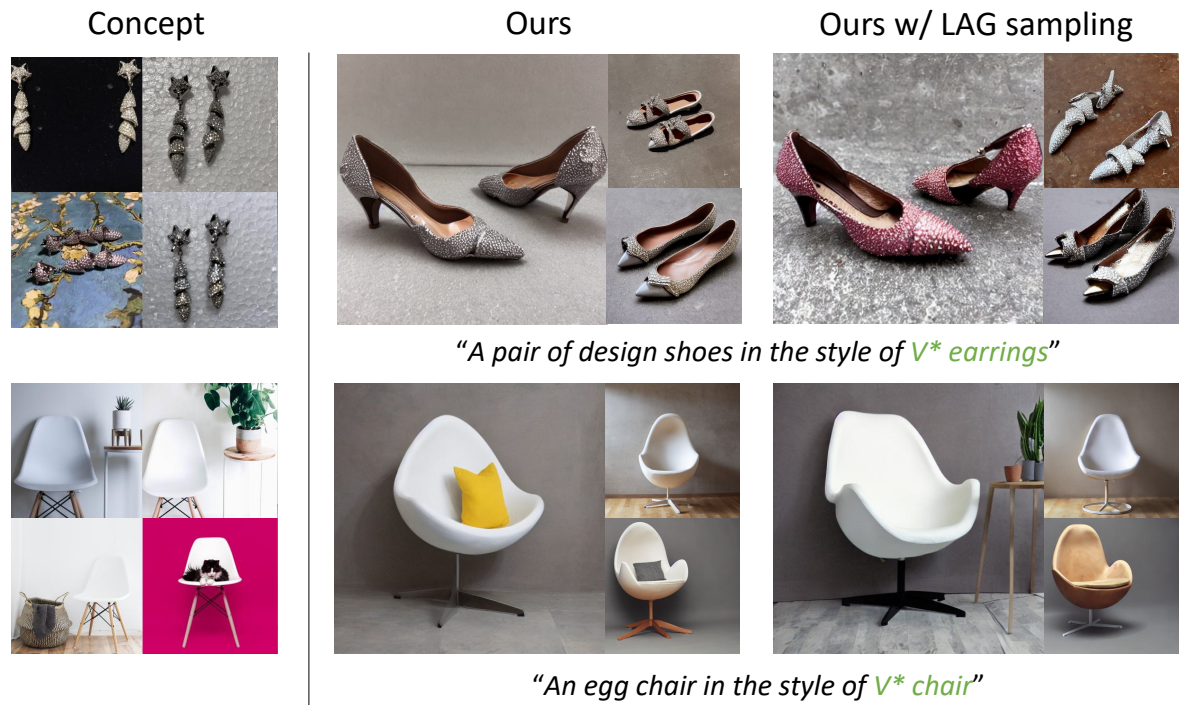

Concept	Ours	Ours w/ LAG sampling

"A pair of design shoes in the style of V earrings"*

"An egg chair in the style of V chair"*

Figure 11. Samples for *object in style of* $V\star$ prompts using personalized residuals with and without LAG sampling where corresponding pairs are generated using the same input noise map.

Concept	Ours	Ours w/ LAG sampling

"Print of V* houseplant on a sweater"

"V* bear oil painting Ghibli inspired"

"A teapot in the style of V* vase"

"Japanese ukiyo-e style depiction of the V* waterfall"
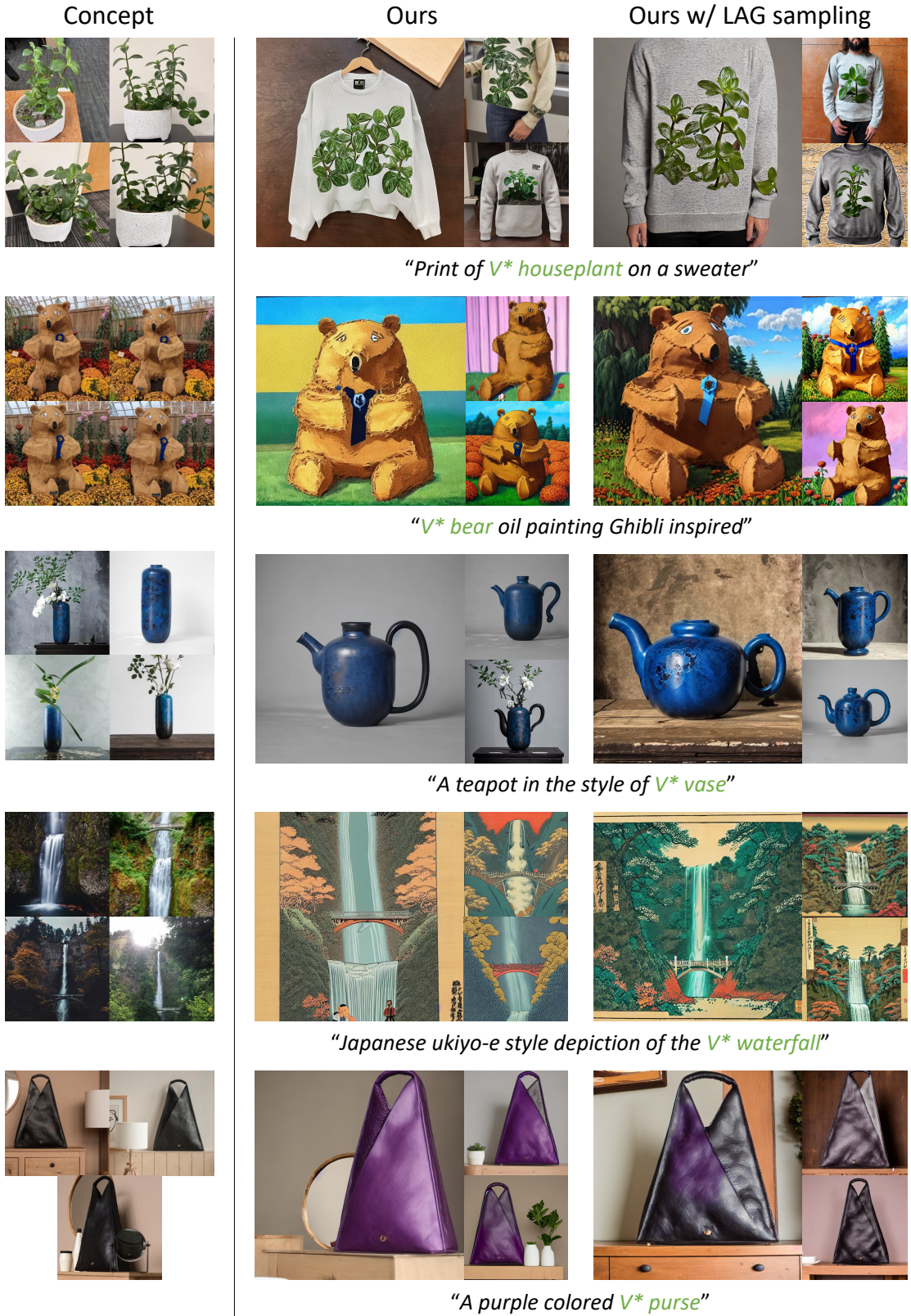
"A purple colored V* purse"

Figure 12. Samples generated using personalized residuals with and without LAG sampling.

Concept	Ours	Ours w/ LAG sampling

"*Rose flowers in V* wooden pot on a table*"

"*A funky Picasso-style cubist painting of V* violin*"

"*V* plushie sitting at the beach with a view of the sea*"

"*V* canal scene painting by artist Claude Monet*"
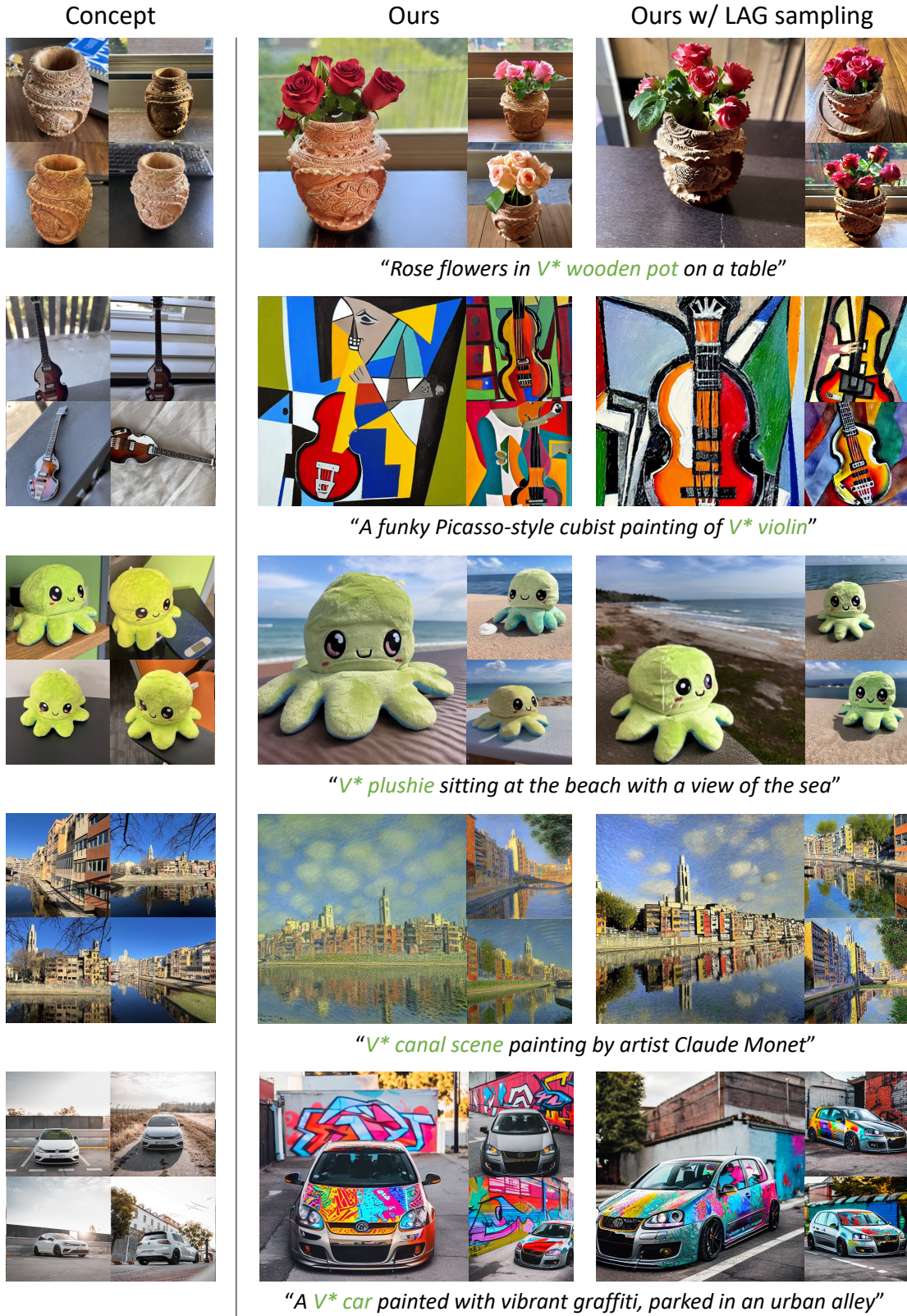
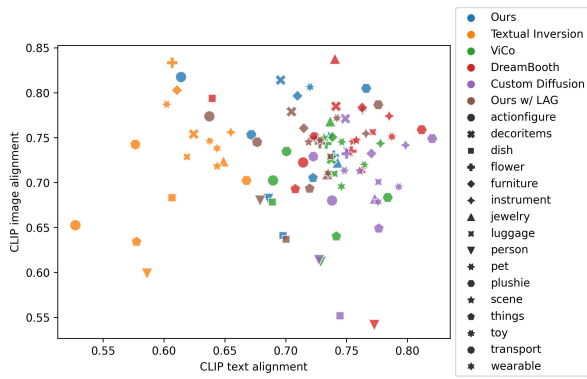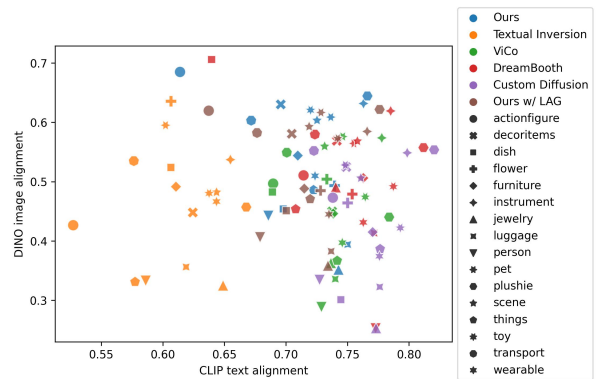"*A V* car painted with vibrant graffiti, parked in an urban alley*"

Figure 13. Samples generated using personalized residuals with and without LAG sampling.

(a) Plot of CLIP image alignment vs. CLIP text alignment.

(b) Plot of DINO image alignment vs. CLIP text alignment.

Figure 14. For each method, we plot the either CLIP or DINO image alignment scores against CLIP text alignment scores averaged across the concepts within each of the 16 categories of CustomConcept101.