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

Supplementary Material

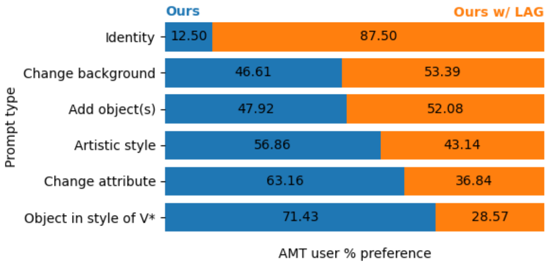


Figure 5. AMT text alignment scores per prompt type.

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* $V*$}. 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* $V*$) 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 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.

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

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.

[2] https://github.com/ChenDarYen/Key-Locked-Rank-One-Editing-for-Text-to-Image-Personalization/

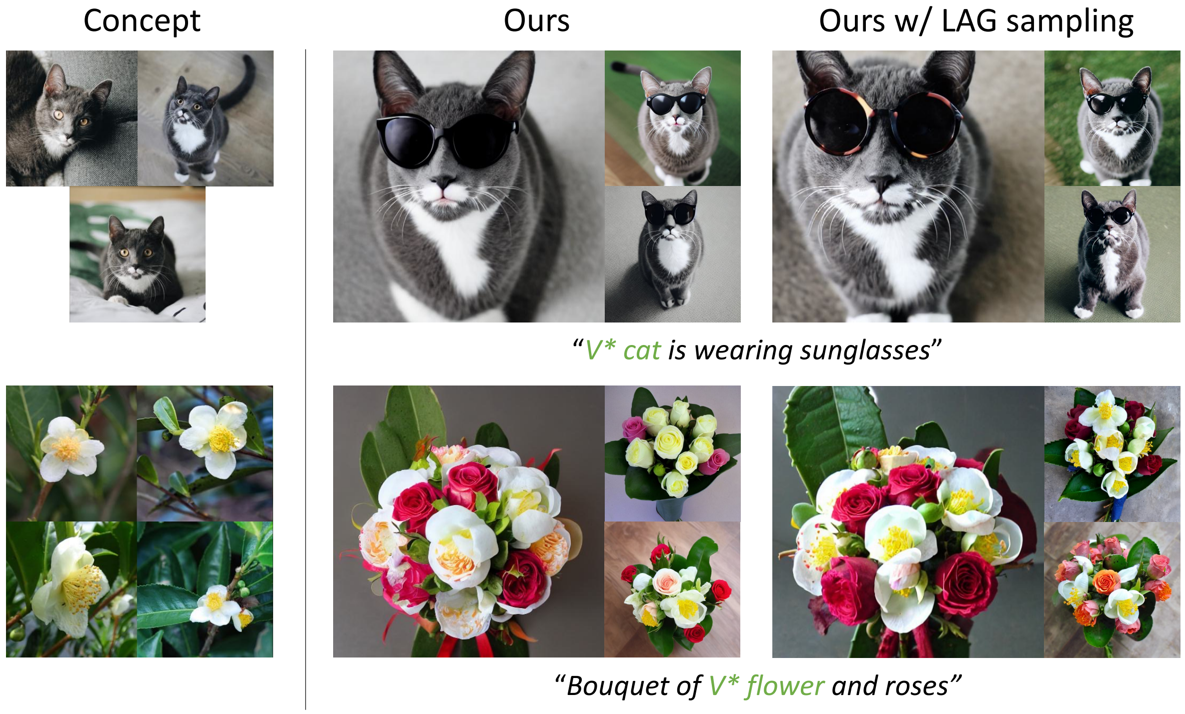

Concept
Ours
Ours w/ LAG sampling
“V* cat is wearing sunglasses”
“Bouquet of V* flower and roses”

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.

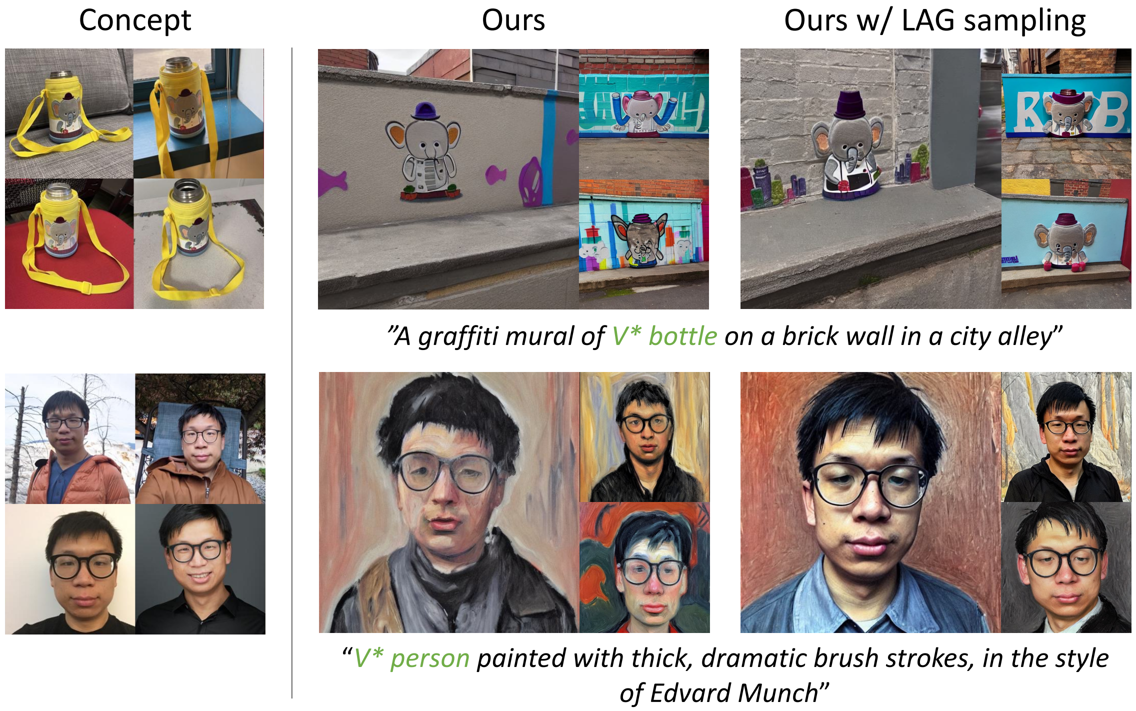

Concept
Ours
Ours w/ LAG sampling
”A graffiti mural of V* bottle on a brick wall in a city alley”
“V* person painted with thick, dramatic brush strokes, in the style of Edvard Munch”

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.

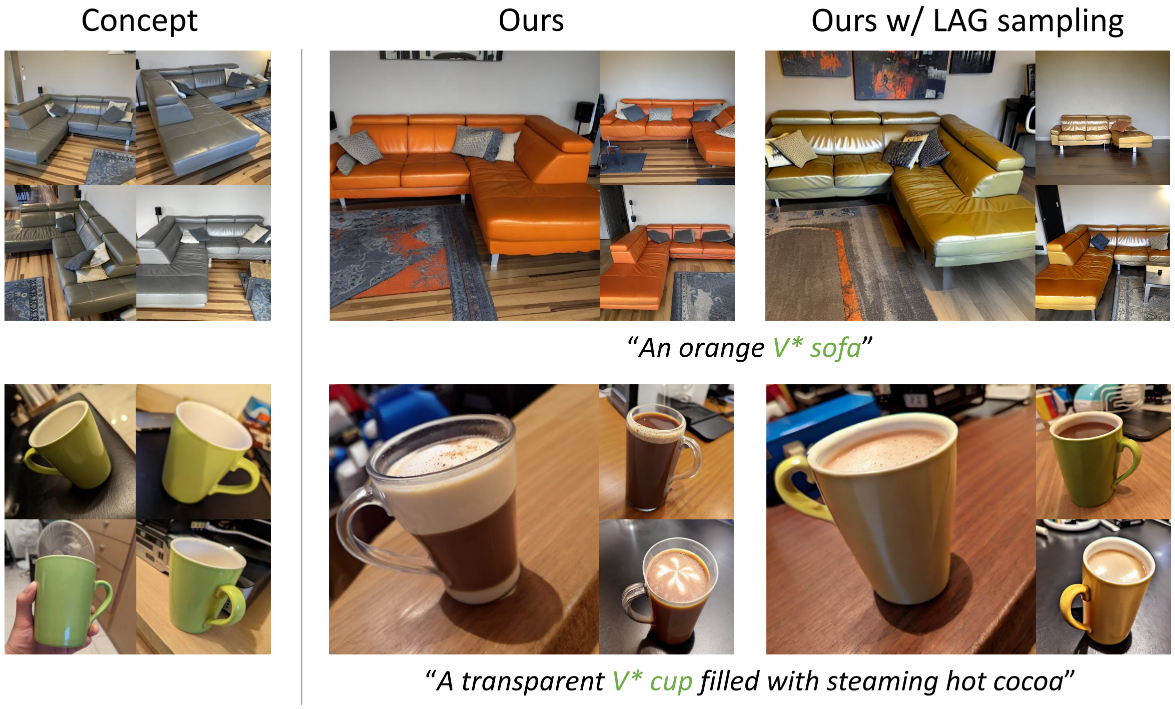


Figure 8. Samples for *change attribute* prompts using personalized residuals with and without LAG sampling where corresponding pairs are generated using the same input noise map.

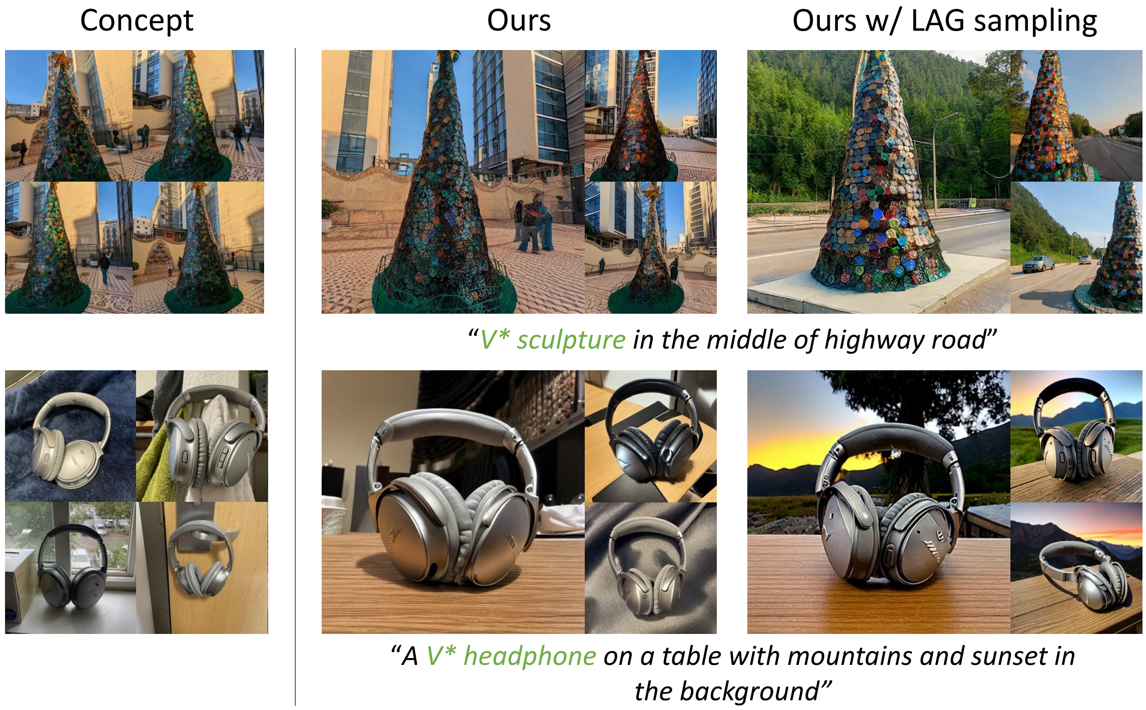

Concept
Ours
Ours w/ LAG sampling
"V* sculpture in the middle of highway road"
"A V* headphone on a table with mountains and sunset in the background"

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.

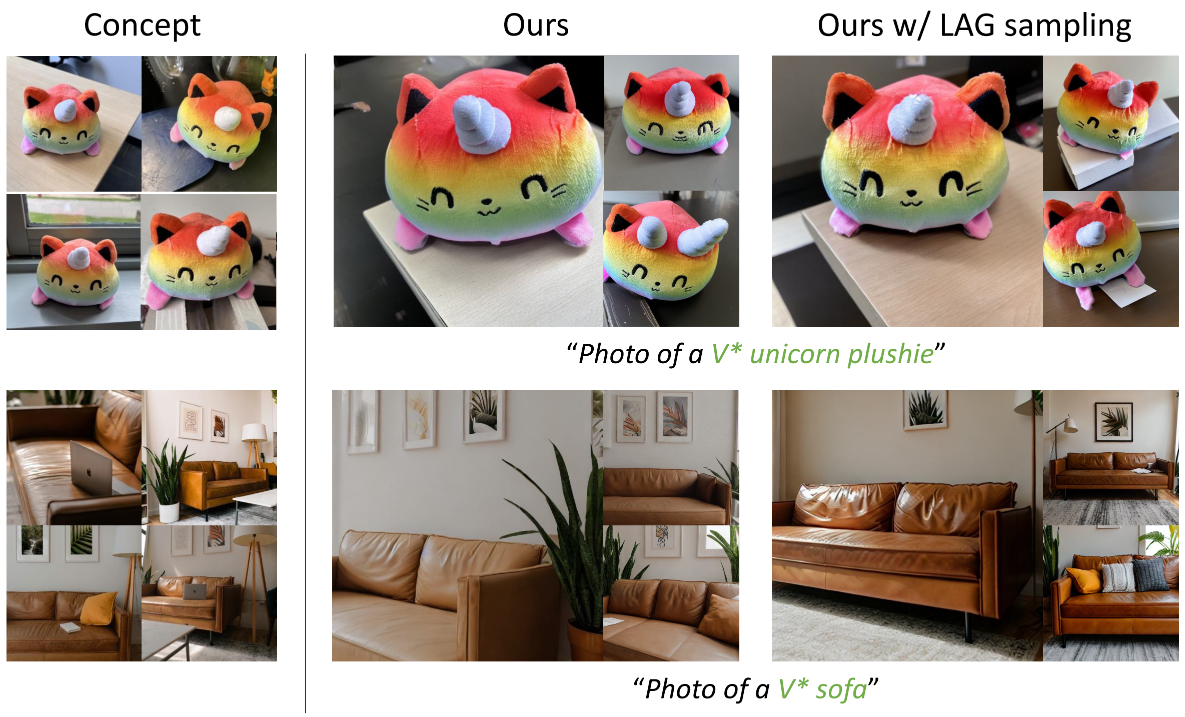


Figure 10. Samples for *identity* prompts using personalized residuals with and without LAG sampling where corresponding pairs are generated using the same input noise map.

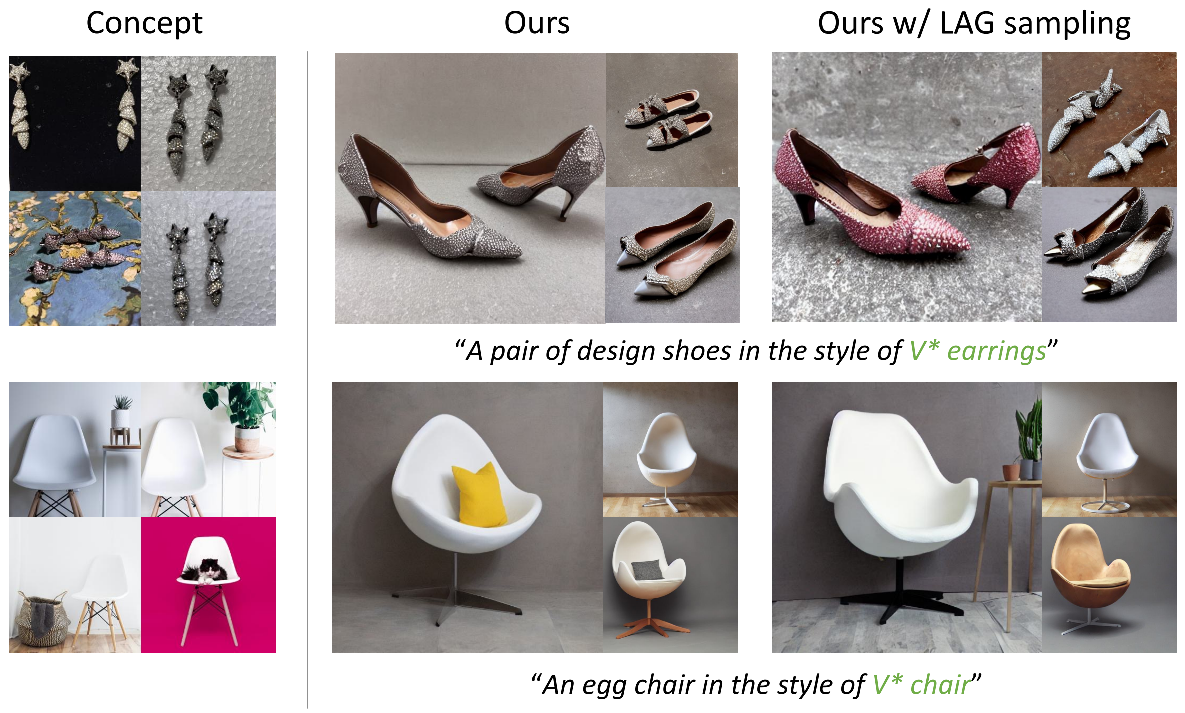

Concept
Ours
Ours w/ LAG sampling
"A pair of design shoes in the style of V* earrings"
"An egg chair in the style of V* chair"

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

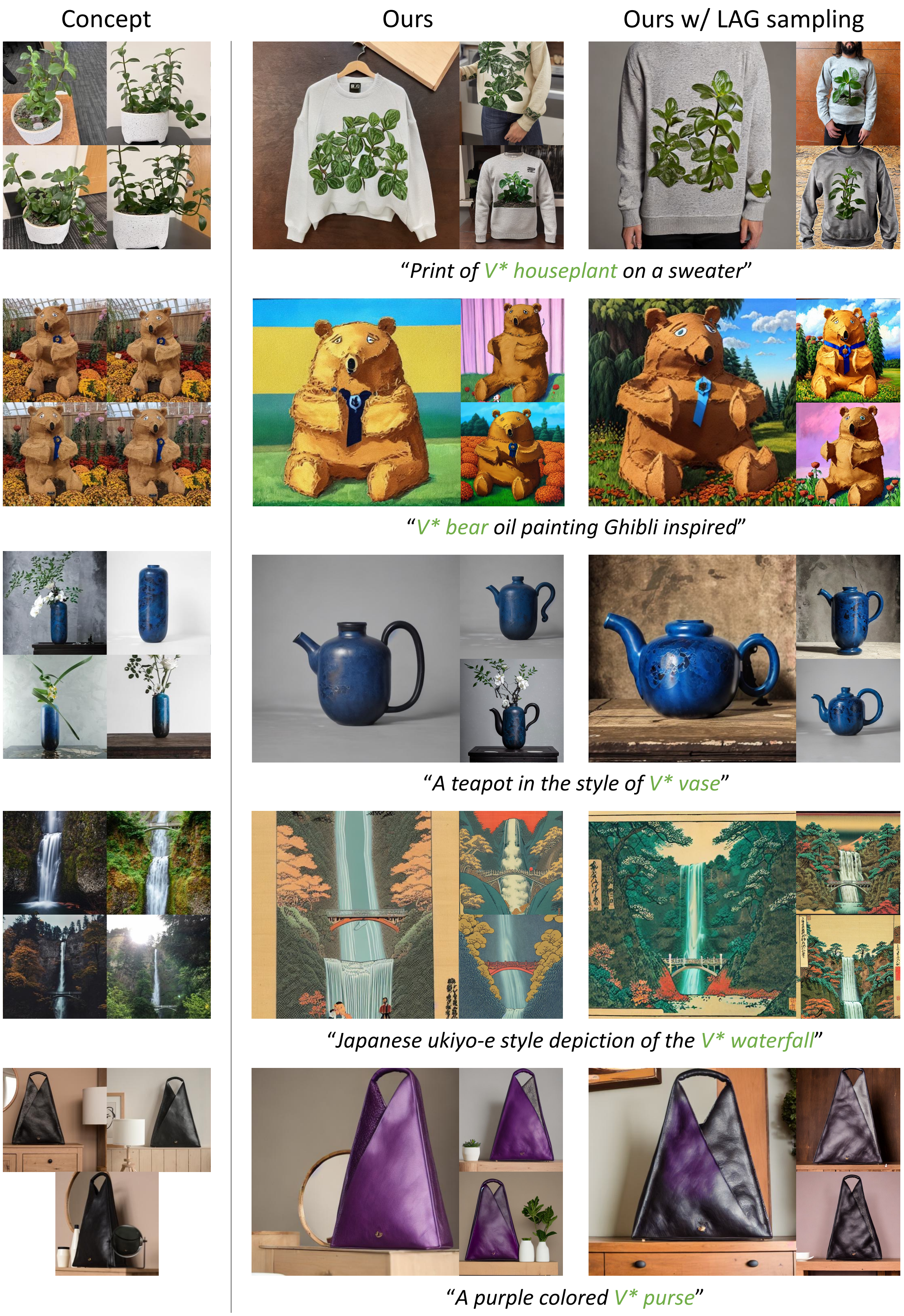

Concept
Ours
Ours w/ LAG sampling
“Print of V* houseplant on a sweater”
“V* bear oil painting Ghibli inspired”
“A teapot in the style of V* vase”
“Japanese ukiyo-e style depiction of the V* waterfall”
“A purple colored V* purse”

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

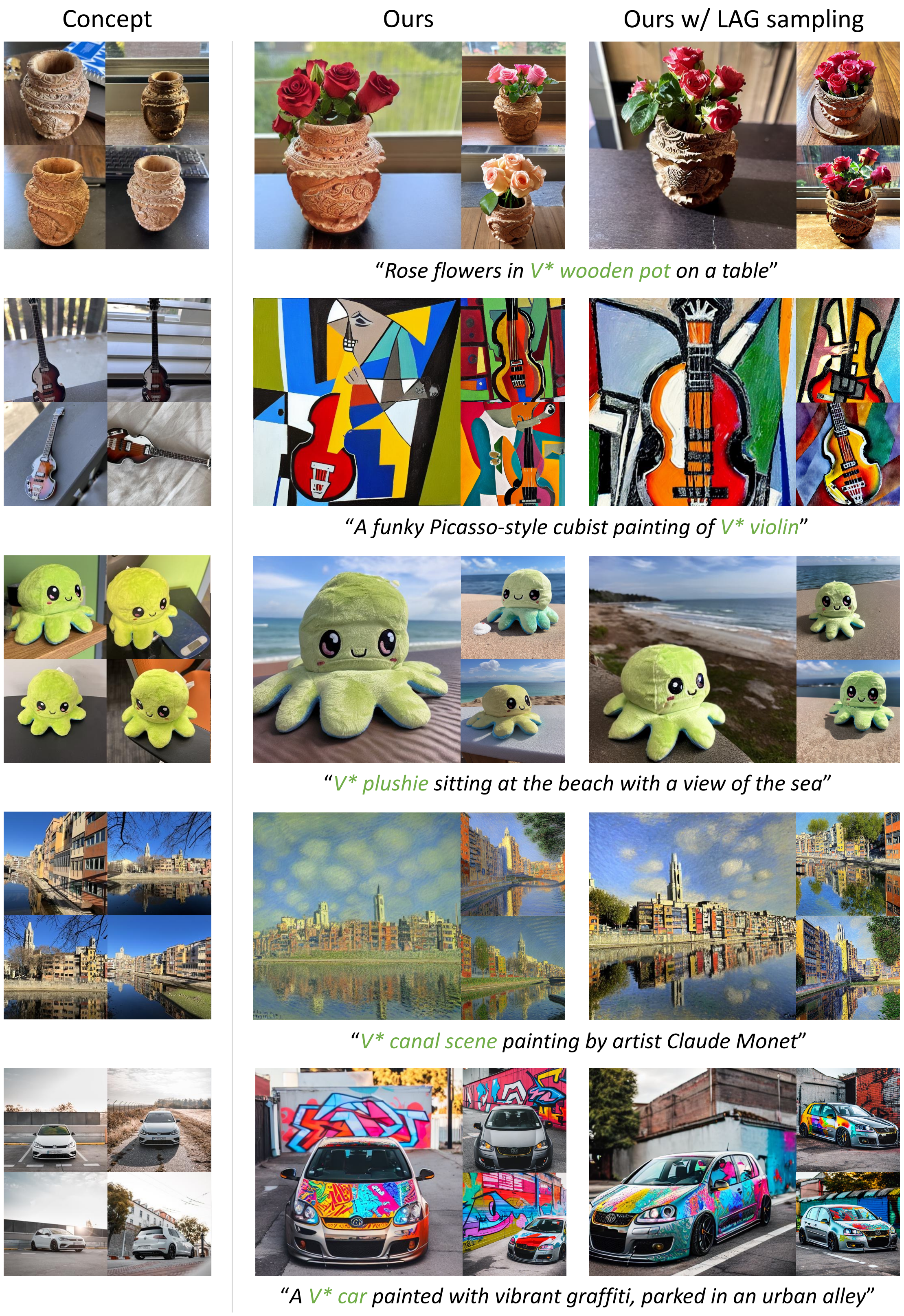
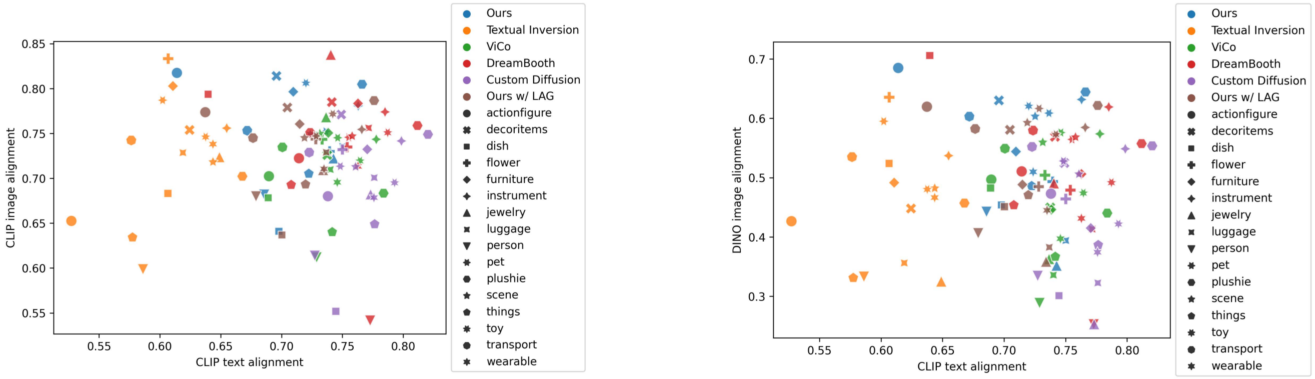

Concept
Ours
Ours w/ LAG sampling
"Rose flowers in V* wooden pot on a table"
"A funky Picasso-style cubist painting of V* violin"
"V* plushie sitting at the beach with a view of the sea"
"V* canal scene painting by artist Claude Monet"
"A V* car painted with vibrant graffiti, parked in an urban alley"

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.