Attention Calibration for Disentangled Text-to-Image Personalization

Supplementary Material

6. Experiments

Additional qualitative results. Further comparisons, including Textual Inversion (TI) [11], are illustrated in Figure 11 (independent concepts) and Figure 12 (combined concepts). Evidently, the concepts synthesized by TI differ significantly from the input image, affirming the quantitative analysis in Sec. 4.2.

Detailed quantitative results on ten datasets. As shown in Tab. 1, our method consistently attains the highest image-alignment across most datasets while maintaining favorable text-alignment compared to the three baselines.

Attention map visualization of ablation studies. The attention maps for the component ablations are presented in Fig. 13, encompassing the following scenarios: (1) Removing the $\mathcal{L}_{\text{bind}}$ loss, (2) removing the $\mathcal{L}_{\text{s\&s}}$ loss, (3) using $\mathcal{L}_s$ (i.e., $\mathcal{L}_{\text{separate}}$ in Sec. 3.3) instead of $\mathcal{L}_{\text{s\&s}}$, (4) removing the suppression strategy, (5) applying twice suppression, (6) removing the Gaussian filter. Observing Fig. 13 reveals the following insights: (1) Without $\mathcal{L}_{\text{bind}}$, new modifiers tend to focus on incorrect classes or vague regions; (2) Absence of $\mathcal{L}_{\text{s\&s}}$ results in interdependence among learned class tokens, especially the "cat" token; (3) Sole reliance on $\mathcal{L}_s$ leads to tiny activation areas for crucial tokens; (4) Removal of the suppression strategy introduces unnecessary activations for new modifiers, apart from their corresponding class regions; (5) Applying twice suppression causes the loss of vital information for new modifiers, (e.g., the attention of V_2^* is obviously smaller than the "dog"); (6) The absence of the Gaussian filter may cause new modifiers to lack specific attributes related to the concepts, such as the attention on the mouth part for V_2^* in the specific dog instance. In summary, our full method generates independent and comprehensive attention maps for crucial tokens.

7. Implementation and Experiment Details

Datasets. We present each training image in Fig. 10.

Textual Inversion [11]. We utilized the implementation from [47] with 5000 training steps, a batch size of 4, and a learning rate of 0.0005. The input prompt, originally "A photo of V^*" in Textual Inversion, is modified to "A photo of V_1^* and V_2^*". The two new words (V_1^* and V_2^*) are initialized with the classes from the input image. For example, if the image contains a cat and a dog, V_1^* and V_2^* token embeddings are initialized as the pre-trained "cat" and "dog" token embeddings.

DreamBooth [40]. We employ the implementation from [47] with 250 training steps, a batch size of 2, and a learning rate of 5×10^{-6}. The input prompt is "V_1^* [class$_1$] and V_2^* [class$_2$]", consistent with our setting in Sec. 3.1. Additionally, we generate 1000 "a [class$_1$] and a [class$_2$]" images using the pre-trained model [40]. New modifiers are initialized as rare token embeddings.

Custom Diffusion [23]. We employ the official implementation with 250 training steps, a batch size of 8, and a learning rate of 8×10^{-5}. The input prompt is also "V_1^* [class$_1$] and V_2^* [class$_2$]", and modifiers are also initialized as rare token embeddings. For regularization, 200 images are selected using clip-retrieval [2] with the caption "a [class$_1$] and a [class$_2$]". We apply the default data augmentation in Custom Diffusion.

DisenDiff (ours). Implementation details are described in Sec. 4.1. For the total loss in Eq. (6), the weight of $\mathcal{L}_{\text{bind}}$ is set to 0.01 in all experiments. The weight of $\mathcal{L}_{\text{s\&s}}$ defaults to 0.01 and occasionally adjusts to 0.001 for specific cases.

Figure 10. **Overview of ten datasets.**

	Method	Cat+Dog	Cow+Bird	Man+Woman	Chair+Vase	Chair+Lamp	Dog+Pig	Mother+Child	Woman+Dog	Horse+Dog	Baby+Toy	**Mean**
Image-alignment (Mean)	**Textual Inversion**	0.732	0.656	0.550	0.649	0.663	0.662	0.557	0.541	0.636	0.607	0.625
	DreamBooth	0.732	0.701	0.606	0.815	0.784	0.701	0.601	0.625	0.708	0.689	0.696
	Custom Diffusion	0.808	0.777	0.719	0.811	**0.798**	0.771	0.705	0.706	**0.747**	0.775	0.762
	Ours	**0.824**	**0.783**	**0.749**	**0.822**	0.795	**0.773**	**0.718**	**0.737**	0.744	**0.808**	**0.775**
Text-alignment (Mean)	**Textual Inversion**	0.802	0.815	**0.739**	**0.814**	**0.834**	**0.834**	0.764	0.776	0.811	0.767	**0.796**
	DreamBooth	**0.804**	0.816	0.738	0.732	0.811	0.830	**0.768**	**0.781**	0.817	0.778	0.788
	Custom Diffusion	0.773	0.843	0.731	0.759	0.794	0.793	0.740	0.754	0.818	**0.800**	0.780
	Ours	0.774	**0.847**	0.727	0.757	0.800	0.794	0.744	0.732	**0.826**	0.796	0.780
Image-alignment (Combined)	**Textual Inversion**	0.743	0.690	0.527	0.687	0.659	0.662	0.572	0.542	0.620	0.644	0.634
	DreamBooth	0.736	0.774	0.679	0.897	0.845	0.697	0.672	0.684	0.784	0.739	0.751
	Custom Diffusion	0.856	0.843	0.801	**0.914**	**0.903**	0.793	0.777	0.807	**0.801**	0.820	0.832
	Ours	**0.865**	**0.855**	**0.828**	0.909	0.883	**0.794**	**0.795**	**0.835**	0.792	**0.870**	**0.843**
Text-alignment (Combined)	**Textual Inversion**	**0.797**	0.805	0.738	**0.800**	**0.816**	**0.839**	0.797	0.777	0.823	0.811	**0.800**
	DreamBooth	0.780	0.823	**0.762**	0.705	0.799	0.824	**0.815**	**0.821**	**0.843**	**0.823**	0.799
	Custom Diffusion	0.736	0.882	0.719	0.698	0.747	0.749	0.735	0.729	0.826	0.792	0.761
	Ours	0.747	**0.896**	0.708	0.711	0.767	0.772	0.746	0.712	0.842	0.805	0.771
Image-alignment (Concept_1)	**Textual Inversion**	0.756	0.688	0.527	0.671	0.669	0.682	0.501	0.463	0.647	0.648	0.625
	DreamBooth	0.763	0.697	0.545	0.755	0.795	0.707	0.520	0.554	0.707	0.679	0.672
	Custom Diffusion	0.818	**0.768**	0.661	0.750	**0.803**	**0.779**	**0.661**	0.635	**0.761**	0.802	0.744
	Ours	**0.837**	0.766	**0.674**	**0.766**	0.804	0.771	0.651	**0.678**	0.752	**0.808**	**0.751**
Text-alignment (Concept_1)	**Textual Inversion**	0.798	0.845	0.732	**0.830**	**0.840**	**0.822**	**0.729**	**0.738**	0.815	0.771	**0.792**
	DreamBooth	**0.823**	0.800	0.722	0.768	0.792	0.818	0.724	0.715	0.819	0.775	0.779
	Custom Diffusion	0.776	**0.856**	**0.733**	0.812	0.805	0.777	0.688	0.718	0.819	**0.831**	0.781
	Ours	0.776	**0.856**	0.732	0.809	0.809	0.774	0.693	0.677	**0.822**	0.826	0.777
Image-alignment (Concept_2)	**Textual Inversion**	0.695	0.590	0.596	0.589	0.660	0.642	0.598	0.619	0.640	0.528	0.616
	DreamBooth	0.696	0.632	0.594	0.795	**0.711**	0.699	0.612	0.635	0.635	0.650	0.666
	Custom Diffusion	0.748	0.721	0.696	0.769	0.688	0.741	0.675	0.675	0.678	0.702	0.709
	Ours	**0.770**	**0.729**	**0.744**	**0.790**	0.699	**0.754**	**0.708**	**0.697**	**0.688**	**0.747**	**0.733**
Text-alignment (Concept_2)	**Textual Inversion**	**0.812**	**0.796**	**0.747**	**0.817**	**0.847**	0.842	0.766	0.812	0.794	0.719	0.795
	DreamBooth	0.809	0.794	0.729	0.725	0.843	0.848	0.765	0.808	0.790	0.737	0.785
	Custom Diffusion	0.808	0.792	0.742	0.767	0.830	**0.853**	**0.797**	**0.815**	0.808	**0.775**	**0.799**
	Ours	0.799	0.787	0.741	0.752	0.823	0.836	0.792	0.807	**0.814**	0.757	0.791

Table 1. **Quantitative comparison on each dataset.** Evaluation metrics are outlined in Section 4.1 (higher is better for both metrics). We report four types of scores (Mean, Combined, Concept_1, Concept_2), and the averaged results across ten datasets are illustrated in Figure 6. The term "Cat+Dog" signifies the presence of both "Cat" and "Dog" concepts within the dataset.

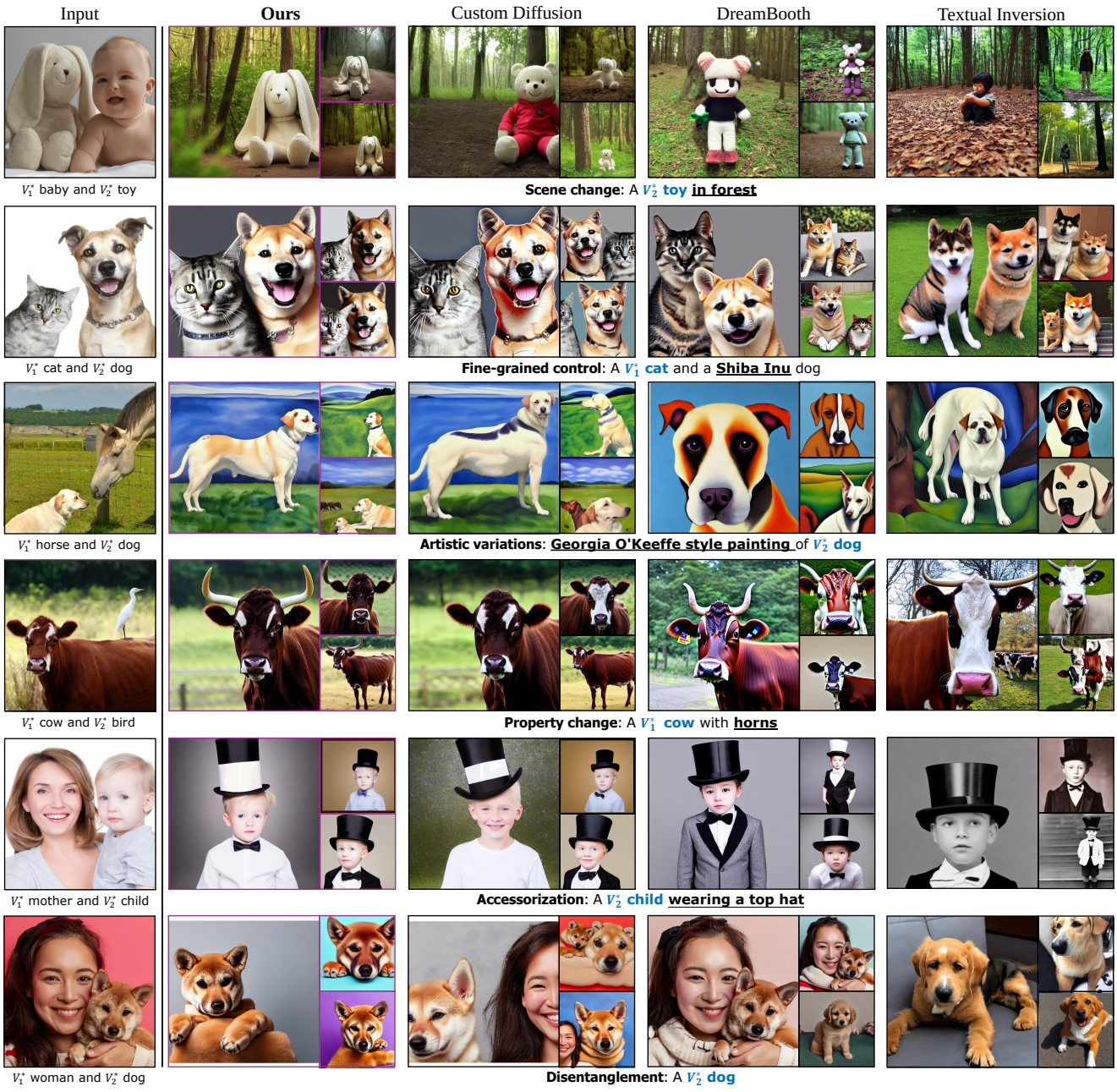

Figure 11. **Qualitative comparison on independent concepts including Textual Inversion.**

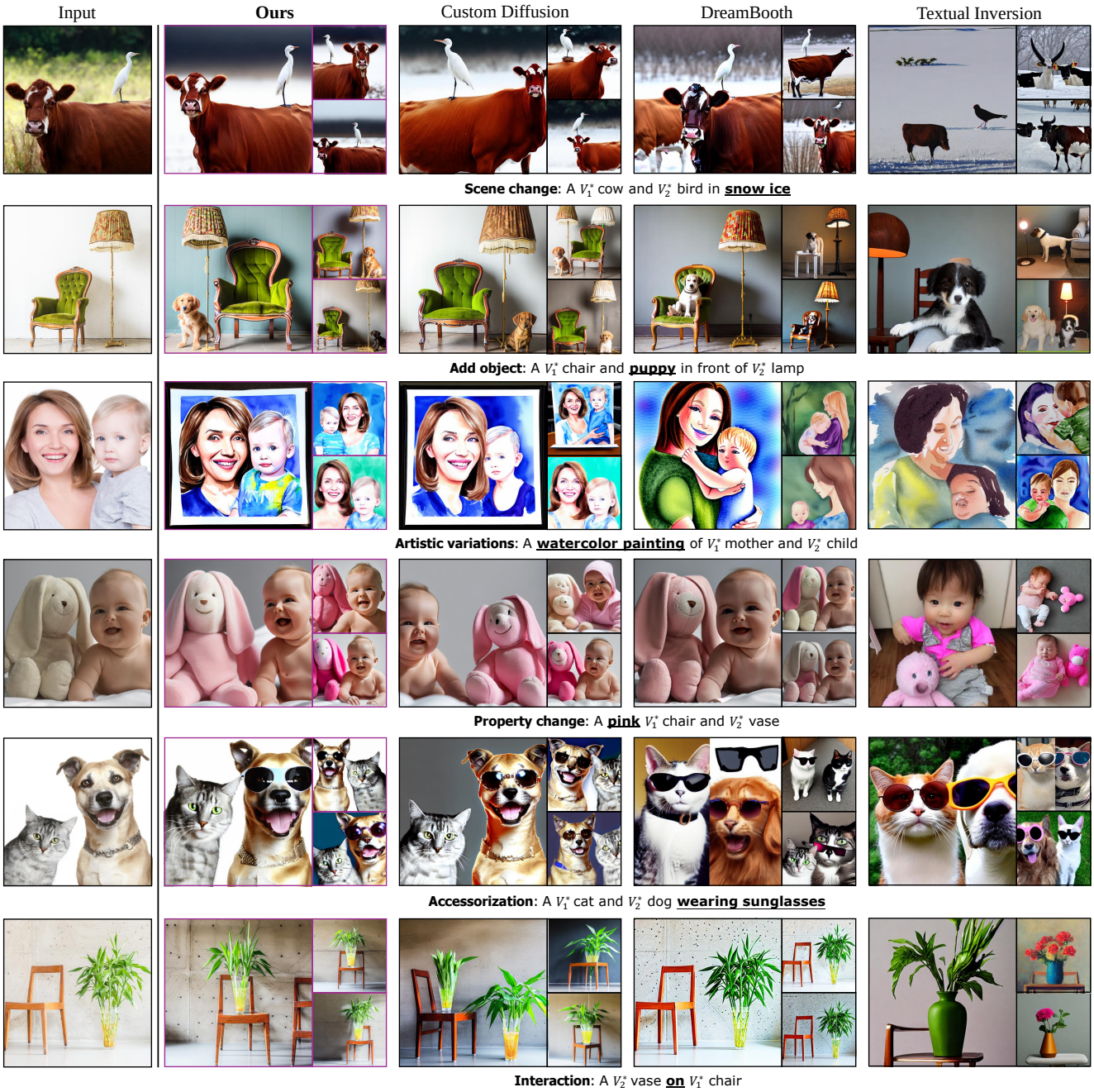

Figure 12. **Qualitative comparison on combined concepts including Textual Inversion.**

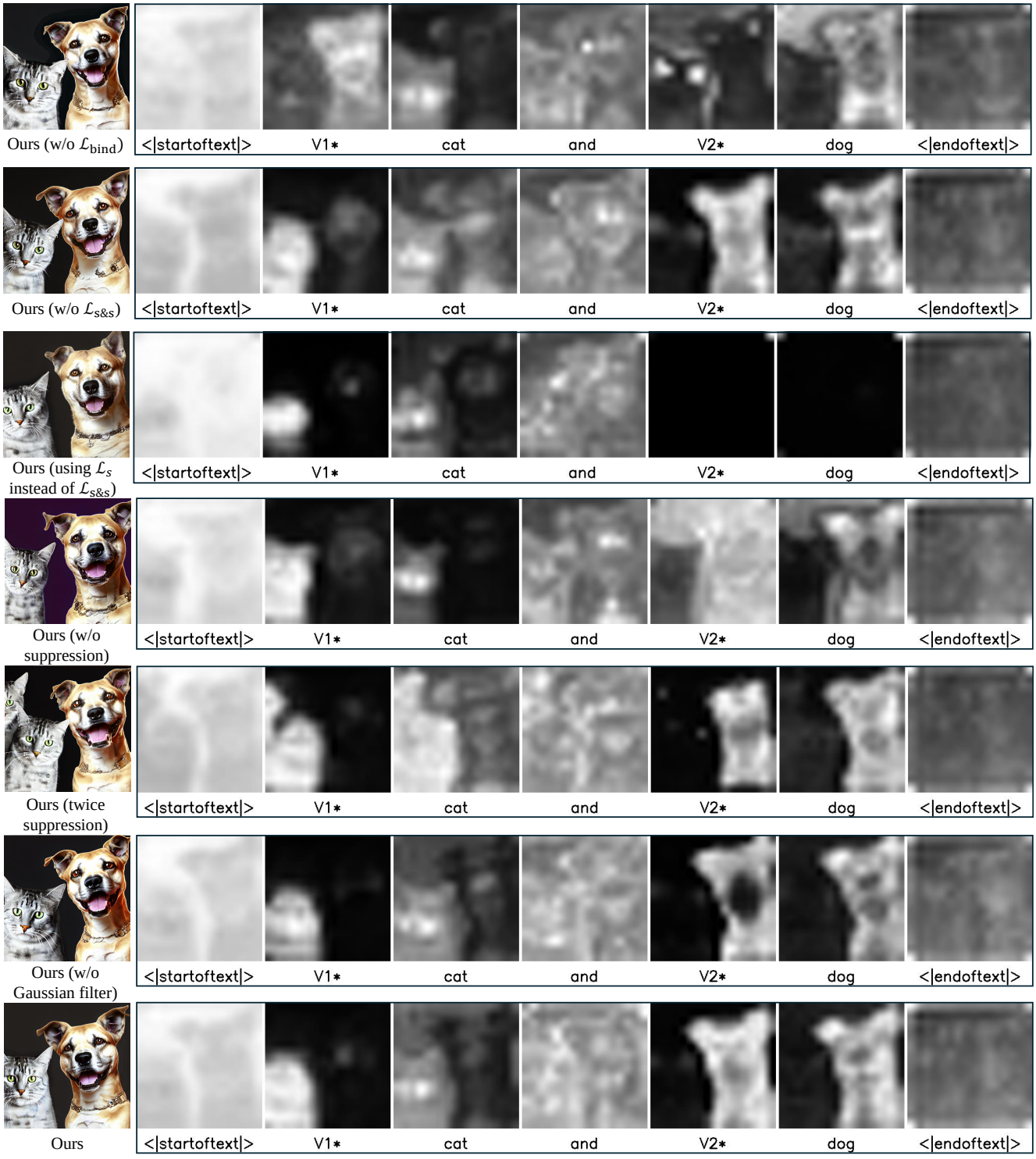

Figure 13. **Attention map visualization of ablations.** Each row represents the generated image and attention maps for all input tokens by ablation methods.