Adversarial Concept Distillation for One-Step Diffusion Personalization

Supplementary Material

6. Broader Impacts

Our method, *OPAD*, enables efficient and high-fidelity personalized image generation, which holds significant potential for a range of creative applications, including design, education, and virtual content creation. By reducing the need for extensive data and computation, *OPAD* democratizes access to advanced generative tools, empowering users with minimal resources to produce customized visual content. However, as with other powerful generative models, our approach also introduces potential risks. These include the unauthorized generation of content, impersonation, and the creation of misleading or harmful imagery. We acknowledge these risks and stress the importance of deploying appropriate safeguards, such as content moderation, usage auditing, and user authentication mechanisms, particularly in real-world applications. We advocate for the responsible use of this technology and encourage the research community and stakeholders to collaborate on developing ethical guidelines and technical solutions to mitigate potential misuse.

7. Limitations

Our proposed method, *OPAD*, marks an initial step toward enabling one-step diffusion models to learn novel concepts efficiently. While the experimental results are promising, several limitations remain and warrant further investigation: *(1) Training efficiency:* The current training pipeline is hindered by the computational overhead introduced by the multi-discriminator architecture. Optimizing or rethinking this component could significantly improve training speed. *(2) Limited one-shot personalization:* The discriminator's reliance on multiple reference samples to model the underlying data distribution makes true one-shot personalization challenging. Designing a more robust discriminator or alternative mechanisms to enable faithful learning from a single image remains an open problem. *(3) Training instability:* As with many GAN-based methods, our approach may exhibit instability across runs, particularly for challenging concepts, where achieving optimal results may require a few trials. Enhancing training stability remains a promising direction for future work. We leave these challenges as compelling avenues for future research, aiming to build upon this initial framework to support broader generalization, compositionality, and efficiency.

8. Ethical and LLM Statements

We acknowledge the potential ethical implications of deploying generative models, including concerns related to privacy, data misuse, and the propagation of biases. All models used in this paper are publicly available, and we will release the modified codes to enable reproduction of our results. We also emphasize the potential misuse of customization approaches in generating misinformation, and we strongly encourage and support their responsible usage. Regarding the use of LLMs, we clarify that in this work they were only minimally employed, specifically for correcting grammatical errors.

9. Overview of T2I Personalization Methods

In this section, we present a comprehensive comparison of representative text-to-image personalization methods, expanding upon the overview introduced in λ-Eclipse [64]. Table 4 provides an extended summary that systematically contrasts these approaches across several key dimensions, including support for single- or multi-subject personalization, training-free versus training-based paradigms, number of input images required, inference efficiency, etc.

10. Algorithmic Description of *OPAD* Training

Distillation Stage. The full training procedure of one-step personalization in *OPAD* is described in Sec. 3.3 of the main paper. For completeness, Algorithm 1 provides the detailed step-by-step implementation. Each iteration consists of three steps: *(1)* the teacher model and text encoder are jointly optimized via the noise prediction loss $\mathcal{L}_{\text{rec}}$ following the Custom Diffusion paradigm; *(2)* training of the student model through a combination of alignment losses with the teacher's outputs and adversarial losses against real image data; and *(3)* updating the discriminators to improve their capacity to differentiate between real and synthesized samples.

Collaborative Learning Stage. During the collaborative learning stage, the training remains the same as before, except that the training examples are replaced with one-step inference samples generated by the student model. This design is beneficial in multiple ways. *(1)* In the initial distillation stage, the availability of real images is limited, and this constrained data scale impedes the training of the teacher model. By contrast, the student model exhibits the ability to learn data distributions from few images. That is a capability endowed by the discriminator, as validated in few-shot GAN frameworks such as TransferGAN[98]

Table 4. We provide an overview of representative text-to-image personalization methods by extending the summary introduced in λ-Eclipse [64]. The base models listed correspond to those used in their original papers. For a fair comparison with the highlighted methods in our study, we re-implemented and adapted all approaches using the same base model configuration as described in our main paper. ChilloutMix is a community-contributed variant of the Stable Diffusion model [69]. Inifinity refers to a variant of the Text-to-Image VAR model [90], while LlamaGen denotes a text-to-image auto-regressive (AR) model [84].

Method	Multi-Subject	Tuning-Free	Base Model	Input Images	Inference Steps	Note
Textual Inversion [21]	✗	✗	SDv1.4	Few-Shot	Multi-Step	Word-Inversion
P+ [93]	✗	✗	SDv1.4	Few-Shot	Multi-Step	Word-Inversion
ProsPect [115]	✗	✗	SDv1.4	1-Shot	Multi-Step	Word-Inversion
MATTE [3]	✗	✗	SDv1.4	1-Shot	Multi-Step	Word-Inversion
Cones 2 [49]	✓	✗	SDv2.1	Few-Shot	Multi-Step	Word-Inversion
DreamBooth [72]	✗	✗	SDv1.4	Few-Shot	Multi-Step	
ClassDiffusion [34]	✗	✗	SDv1.5	Few-Shot	Multi-Step	
DisenBooth [12]	✗	✗	SDv2.1	1-shot	Multi-Step	
CatVersion [118]	✗	✗	SDv1.5	Few-Shot	Multi-Step	
AttnDreamBooth [62]	✗	✗	SDv2.1	1-shot	Multi-Step	
ViCo [92]	✗	✗	SDv1.4	Few-Shot	Multi-Step	
TextBoost [63]	✗	✗	SDv1.5	1-shot	Multi-Step	
NeTI [4]	✗	✗	SDv1.4	Few-Shot	Multi-Step	
HyperDreamBooth [73]	✗	✗	SDv1.5	1-shot	Multi-Step	
E4T [22]	✗	✗	SD	1-shot	Multi-Step	
Hyper-E4T [5]	✗	✗	SD	1-shot	Multi-Step	
ARBooth [15]	✗	✗	Infinity [27]	Few-Shot	Multi-Step	
Proxy-Tuning [105]	✗	✗	LlamaGen [84]	Few-Shot	Multi-Step	
Continual Diffusion [81]	✓	✗	SD	Few-Shot	Multi-Step	
Perfusion [89]	✓	✗	SDv1.5	Few-Shot	Multi-Step	
Custom Diffusion [40]	✓	✗	SDv1.4	Few-Shot	Multi-Step	
Cones [48]	✓	✗	SDv1.4	1-shot	Multi-Step	
SVDiff [28]	✓	✗	SDv1.5	Few-Shot	Multi-Step	
FreeCustom [19]	✓	✗	SDv1.5	1-Shot	Multi-Step	
Mix-of-Show [25]	✓	✗	Chilloutmix	Few-Shot	Multi-Step	
LoRACLR [80]	✓	✗	Chilloutmix	Few-Shot	Multi-Step	
Orthogonal [65]	✓	✗	Chilloutmix	Few-Shot	Multi-Step	
OMG [38]	✓	✗	SDXL	Few-Shot	Multi-Step	
Zip-LoRA [78]	✓	✗	SDXL	Few-Shot	Multi-Step	
Break-A-Scene [7]	✓	✗	SDv2.1	1-shot	Multi-Step	
TokenVerse [24]	✓	✗	Flux [42]	1-shot	Multi-Step	
OPAD (Ours)	✗	✗	SDTurbo [77]	Few-Shot	1-step	
PhotoMaker [44]	✗	✓	SDXL	1-shot	Multi-Step	Human Face
ConsistentID [33]	✗	✓	SDv1.5	1-shot	Multi-Step	Human Face
InstantID [96]	✗	✓	SDXL	1-shot	Multi-Step	Human Face
Profusion [120]	✗	✓	SDv2	1-shot	Multi-Step	Human Face
PuLID [26]	✗	✓	SDXL	1-shot	Multi-Step	Human Face
Infinite-ID [104]	✗	✓	SDXL	1-shot	Multi-Step	Human Face
LCM-Lookahead [23]	✗	✓	SDXL	1-shot	Multi-Step	Human Face
InfiniteYou [37]	✗	✓	Flux [42]	1-shot	Multi-Step	Human Face
IP-Adapter [110]	✗	✓	SDv1.5	1-shot	Multi-Step	
ELITE [101]	✗	✓	SDv1.4	1-shot	Multi-Step	
UMM-Diffusion [56]	✗	✓	SDv1.5	1-shot	Multi-Step	
InstantBooth [79]	✗	✓	SDv1.4	Few-Shot	Multi-Step	
BLIP-Diffusion [43]	✗	✓	SDv1.5	1-shot	Multi-Step	
JeDi [112]	✗	✓	SDv1.4	Few-Shot	Multi-Step	
Re-Imagen [13]	✗	✓	Imagen [75]	1-shot	Multi-Step	
SuTi [14]	✗	✓	Imagen [75]	Few-Shot	Multi-Step	
Taming [36]	✗	✓	Imagen [75]	1-shot	Multi-Step	
Kosmos-G [61]	✓	✓	SDv1.5	1-shot	Multi-Step	
SSR-Encoder [116]	✓	✓	SDv1.5	1-shot	Multi-Step	
λ-Eclipse [64]	✓	✓	Kandinsky [6]	1-shot	Multi-Step	
FastComposer [107]	✓	✓	SDv1.5	1-shot	Multi-Step	
Subject-Diffusion [54]	✓	✓	SDv2.1	1-shot	Multi-Step	
RMCC [35]	✓	✓	SDXL	1-shot	Multi-Step	
Emu2 [85]	✓	✓	SDXL	1-shot	Multi-Step	
MS-Diffusion [97]	✓	✓	SDXL	1-shot	Multi-Step	

Algorithm 1 Training Pipeline of One-step Personalization in *OPAD*

Input: Real image dataset $\mathcal{D} = \{(x_0^r, y)\}$; teacher model ϵ_θ^{tc}; student model $\mathcal{G}^{st}$; discriminators $\{\mathcal{D}_k\}_{k=1}^K$; text encoder $\tau(\cdot)$; diffusion steps T; noise schedule functions $\{\alpha_t, \sigma_t\}$; weighting functions $c(t)$, $\lambda_{\text{id}}, \lambda_{\text{mse}}, \lambda_{\text{ms}}, \{\lambda_k\}_{k=1}^K$
for each training iteration **do**
  Sample real image and prompt: $(x_0^r, y) \sim \mathcal{D}$
  Random noise $\epsilon \sim \mathcal{N}(0, 1)$
  Encode prompt: $\mathcal{C} \leftarrow \tau(y)$
  Step 1: Teacher training
  Sample timestep $t \sim \mathcal{U}(1, T)$
  Generate noisy input: $x_t = \alpha_t x_0^r + \sigma_t \epsilon$
  Predict noise: $\hat{\epsilon} \leftarrow \epsilon_\theta^{tc}(x_t, t, \mathcal{C})$
  Compute loss: $\mathcal{L}_{\text{rec}} = \|\epsilon - \hat{\epsilon}\|_2^2$
  Update teacher model and the text encoder using $\mathcal{L}_{\text{rec}}$
  Step 2: Student training
  Sample latent: $x_T \sim \mathcal{N}(0, 1)$
  Generate image: $x_0^{\text{st}} \leftarrow \mathcal{G}^{st}(x_T, T, \text{stopgrad}(\mathcal{C}))$ ▷ stopgrad$(\cdot)$ denotes stop-gradient
  // Alignment loss
  Forward diffuse: $x_t^{\text{st}} = \alpha_t x_0^{\text{st}} + \sigma_t \epsilon$
  Denoise: $x_0^{\text{tc}} \leftarrow \text{stopgrad}(\epsilon_\theta^{tc}(x_t^{\text{st}}, t, \mathcal{C}))$
  Compute alignment loss:
  $\mathcal{L}_{\text{align}} = c(t) \cdot [\lambda_{\text{id}} \cdot \mathcal{L}_{\text{id}}(x_0^{st}, x_0^{tc}) + \lambda_{\text{mse}} \cdot \mathcal{L}_{\text{mse}}(x^s, x^t) + \lambda_{\text{ms}} \cdot \mathcal{L}_{\text{swd}}(x_0^{st}, x_0^{tc})]$
  // Adversarial loss
  Compute adversarial loss: $\mathcal{L}_{\text{GAN}}^G = \sum_{k=1}^K \lambda_k \cdot \mathbb{E}_{x_0^{st}} [-\log(D_k(x_0^{st}))]$
  Update student model using: $\mathcal{L}_{\text{st}} = \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{GAN}}^G$
  Step 3: Discriminator training
  for each discriminator $\mathcal{D}_k$ **do**
    Compute the discriminator loss:
    $\mathcal{L}_{\text{GAN}}^{D_k} = -\left[\mathbb{E}_{x_0^r} [\log D_k(x_0^r)] + \mathbb{E}_{x_0^{st}} [\log(1 - D_k(\text{stopgrad}(x_0^{st})))]\right]$
    Update $\mathcal{D}_k$ using: $\mathcal{L}_{\text{GAN}}^{D_k}$
  end for
end for
Output: Trained teacher model ϵ_θ^{tc}, student model $\mathcal{G}^{st}$, and text encoder τ

and MineGAN[99]. Consequently, during the collaborative learning stage, our objective is to sample from the data distribution learned by the student model, using these samples as training examples to enhance the teacher model's performance. *(2)* Notably, the 1-step diffusion student model learns distributions distinct from those acquired by the teacher model. Similar observation can be found in ADD[77], where the discriminator loss primarily shapes the data distribution of the student model, while the distillation loss facilitates convergence and enhances conceptual alignment with the teacher's outputs. Supporting evidence for this ablation study can be found in Table 9. *(3)* By shifting the training examples from few real image inputs to images generated by the 1-step student model, the teacher model is enabled to learn from the student's distribution, a distribution partially shaped by the discriminator's design and characterized by image features not inherently present in the teacher model, thereby yielding beneficial effects.

11. Additional results on method comparison

In this section, we present additional quantitative and qualitative results to validate the effectiveness and efficiency of our proposed method. Table 5 extends the comparisons from the main paper by reporting both training and inference time (in seconds), along with the number of optimization iterations required by each method. These results underscore the efficiency of our approach, achieving image generation in just *0.18 seconds* per instance during inference. Figures 6 through 9 provide additional qualitative comparisons against representative baseline methods. Each figure presents a concept reference image (left) followed by results from various approaches. Our method, *OPAD*, consistently delivers superior visual fidelity while maintaining rapid inference, highlighting its practical advantages for resource-constrained applications in one-step diffusion-based image generation (*1-SDP*).

Table 5. Quantitative comparisons with existing text-to-image (T2I) personalization methods. NFEs indicates the number of function evaluations.

Methods	Model	Train NFEs	Inference NFEs	CLIP-T↑	CLIP-I↑	DINO↑	Train Time (s)	Inference Time (s)	iterations
Custom diffusion	SD 2.1	1000	25	0.264	0.761	0.555	345	2.73	1000
	SD Turbo	1000	1	0.207	0.530	0.097	541	0.22	1000
	SD Turbo	1000	4	0.257	0.597	0.235	541	0.54	1000
	SD Turbo	1000	25	0.276	0.647	0.337	541	1.57	1000
	SD Turbo	1	1	0.205	0.518	0.058	543	0.22	1000
	SD Turbo	1	4	0.246	0.556	0.109	543	0.53	1000
	SD Turbo	4	1	0.206	0.532	0.105	543	0.23	1000
	SD Turbo	4	4	0.258	0.600	0.244	543	0.54	1000
Textual Inversion	SD Turbo	1	1	0.252	0.564	0.166	2269	0.13	4000
Cones 2	SD Turbo	1	1	0.273	0.619	0.204	2446	0.51	4000
DreamBooth	SD Turbo	1	1	0.188	0.536	0.111	281	0.14	1000
TextBoost	SD Turbo	1	1	0.217	0.570	0.167	64	0.15	500
DisenBooth	SD Turbo	1	1	0.251	0.564	0.231	905	0.18	2000
Lora	SD Turbo	1	1	0.212	0.585	0.160	141	0.15	800
IP-Adapter	TCD + SDXL	/	1	0.204	0.628	0.325	/	0.39	/
OminiControl	Flux	/	1	**0.279**	0.727	0.455	/	2.48	/
OPAD	SD Turbo	1	1	0.252	**0.783**	**0.637**	3137	0.18	1000

Discussion on runtime cost. About the training time, this limitation is inherent to optimization-based customization approaches, which universally require additional runtime computation when encountering novel concepts. Conversely, existing optimization-free methods, including encoder-based frameworks (IP-Adatper [110], DreamO [58], Xverse [11], UNO [103], InfiniteYou [37], etc.) and unified models (BAGEL [18], GPT-4o [1], OmniGen2 [102], etc.), demand extensive datasets for training. Furthermore, no encoder-based or unified model to date fully supports few-step (or even one-step) diffusion models. This leaves integrating one-step speed with unified model versatility largely underexplored. In this work, we aim to be the first to investigate the realization of one-step personalization via an optimization-based approach, with optimization-free alternatives designated as future work.

Comparison with Flux+OminiControl. We further compare *OPAD* with the recent Flux [42] model combined with the OminiControl [86]. It is important to note that OminiControl is trained on large-scale datasets similar to IP-Adapter, which makes the evaluation against our method not entirely equitable. The evaluation is conducted on the DreamBooth dataset under the 1-step setup, and the results are summarized in the lower part of Table 5. As shown, Flux+OminiControl outperforms other baselines reported in the table; however, it remains significantly inferior to our proposed *OPAD*. This performance gap can be attributed to the weaker generation capability of Flux constrained to 1-step inference.

12. Discussions for alternative designs

We further compare our method with several alternative designs in order to clarify the motivation and validity of our proposed framework. All experiments in this section are performed on the DreamBooth dataset.

Teacher-first paradigm. In this design, the teacher is first trained, and the teacher-generated samples for the target concept (with varying text prompts) are directly used as supervised training data for the student. The identity loss (Eq. 2 in the main paper) is applied between the teacher-generated samples and the student outputs, allowing the student to learn identity-related features from the teacher model. As shown in Table 6, this design performs worse than our proposed approach. Moreover, it suffers from several inherent limitations: (1) Computational overhead: teacher inference requires multiple steps, which is inefficient; (2) Teacher irreliability: as discussed in the main paper, the teacher does not always successfully learn the target concepts; (3) Limited image diversity: the generated images consistently feature highly similar visual appearances; and (4) Performance ceiling: the student's performance is inherently bounded by the capabilities of the teacher.

In addition, we also attempted to directly apply VSD [100], SDS [67], and MSE losses to distill teacher-learned concepts into the student model under this paradigm. However, we observed that this approach was insufficient for transferring the teacher's personalization capabilities to the student.

Discussion on feed-forward customization methods. Beyond the teacher-first paradigm, an alternative direction

Table 6. Additional results for alternative designs on the DreamBooth dataset.

Method	CLIP-T ↑	CLIP-I ↑	DINO ↑
Teacher-first	**0.266**	0.725	0.503
Remove STE	0.242	0.742	0.551
Ours (*OPAD*)	0.252	**0.783**	**0.637**

Table 7. User study for diverse methods.

Method	Preference Rate (%)
Custom Diffusion [40]	32.26%
Cones 2 [48]	0.48%
DisenBooth [12]	0.36%
OPAD	66.90%

Table 8. One-shot performance of our *OPAD*.

Methods	**CLIP-T**	**CLIP-I**	**DINO**
One-shot	0.231	0.713	0.470
Few-shot	**0.252**	**0.783**	**0.637**

is to build on feed-forward customization methods such as SynCD [41] and JeDi [112], and then distill these models into a few-step diffusion framework. However, most existing distillation techniques for diffusion models are primarily designed to align the data distributions of few-step models with those of their teacher models. This emphasis stems from the inherent difficulty that few-step diffusion models face in replicating the full denoising trajectory of their teacher. As a result, subtle discrepancies in concept-specific details are often introduced, as observed in prior works such as AYF [74], ADD [77], and LCM [51].

Remove STE. We further investigate the effect of sharing the text encoder between the teacher and student models. Removing the shared text encoder (STE) results in a clear performance drop, demonstrating that STE not only simplifies the training framework but also improves learning efficiency.

13. User study

To evaluate alignment with human preferences, we conducted a user study involving 15 participants, yielding 840 preference annotations per method. In each trial, participants were presented with a set of generated images and instructed to "select the best image from each group, considering both text-image alignment and object identity consistency." The methods evaluated in our study include Custom Diffusion (under a multi-step setting), as well as Cones 2, DisenBooth, and our method (all under a single-step setting). As summarized in Table 7, our approach, *OPAD*, significantly outperformed other methods, receiving at least 34% more user votes than the second-best method. These findings underscore the strong alignment between *OPAD* 's outputs and human perceptual judgments.

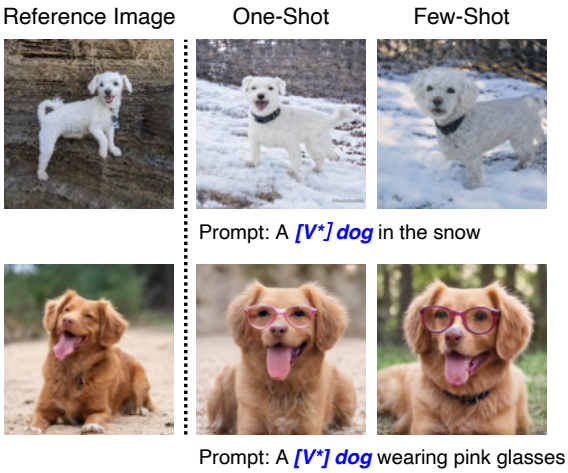

Figure 5. The qualitative results of the 1-shot performance.

14. Extended Ablation Study

In Sec. 4.3 of the main paper, we explore the contributions of key components in *OPAD*, namely the teacher model and discriminators. A more detailed ablation study is presented in Table 9, wherein individual loss terms and discriminators are systematically removed. The results indicate that omitting the ID loss, latent MSE loss, or MSSWD loss causes notable performance degradation, particularly reflected in reduced DINO scores, underscoring their critical role in maintaining alignment with the teacher model. Furthermore, removal of any single discriminator leads to more pronounced declines across all evaluation metrics. Collectively, these findings demonstrate the complementary nature of the various loss functions and discriminators in improving generation fidelity and semantic consistency. Qualitative comparisons provided in Fig. 10 further illustrate the visual impact of removing each component. Beyond the general decline in generation quality, we observe that omission of certain components can induce training instability or divergence for specific concepts.

15. 1-shot performance.

We further evaluate our method under a one-shot supervision setting, wherein only a single image is utilized for training. As summarized in Table 8, performance declines across all evaluation metrics relative to the few-shot scenario. This degradation is anticipated, given that our approach is not explicitly optimized for one-shot learning, and

Table 9. Ablation study of our method *OPAD*.

Methods	CLIP-T	CLIP-I	DINO
Full model	**0.252**	**0.783**	**0.637**
w/o the teacher	0.240	0.719	0.505
w/o all the discriminators	0.200	0.566	0.105
w/o Identity Feature Loss	0.248	0.769	0.618
w/o MSE loss	0.242	0.739	0.528
w/o MS-SWD loss	0.249	0.754	0.553
w/o the discriminator of the Dino v1	0.231	0.689	0.441
w/o the discriminator of the Dino v2	0.246	0.736	0.534
w/o the discriminator of the Clip	0.227	0.678	0.409

the scarcity of supervisory data increases the likelihood of training instability. Qualitative results illustrated in Fig. 5 demonstrate that, although one-shot training can yield visually plausible outputs, the generated images occasionally lack fine-grained details corresponding to the novel concept.

16. Results on the CustomConcept101 Dataset

We further evaluate our method on the CustomConcept101 dataset [40]. Qualitative results, presented in Fig. 11 and Fig. 12, demonstrate that our approach generalizes effectively across a diverse set of concepts and prompt types, consistently generating high-quality outputs.

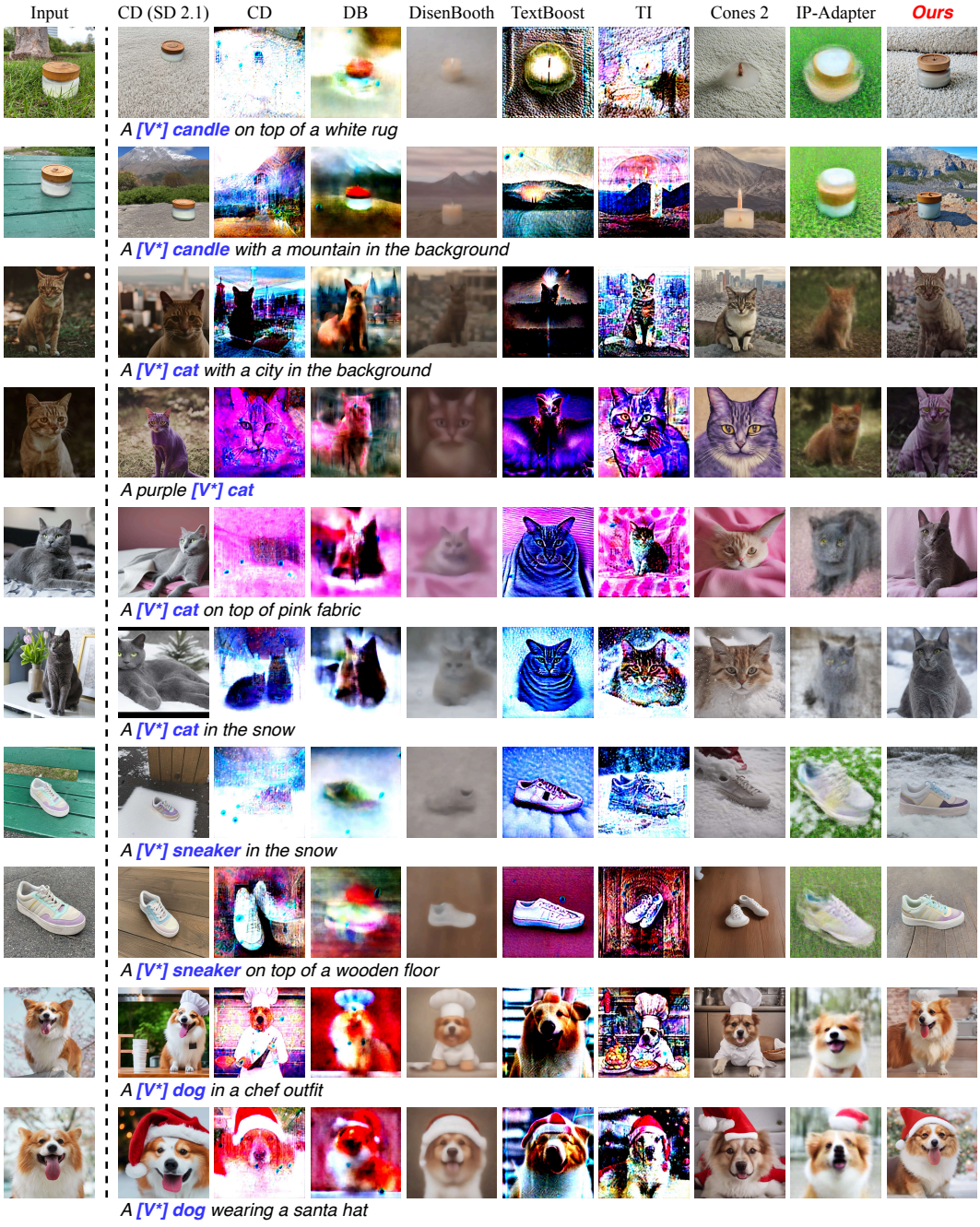

Figure 6. Our method *OPAD* (last column) compared with existing methods applied to the *1-SDP* setup with SDTurbo [77] as the one-step diffusion backbone. One representive concept image is shown on the left-most column. (Part 1)

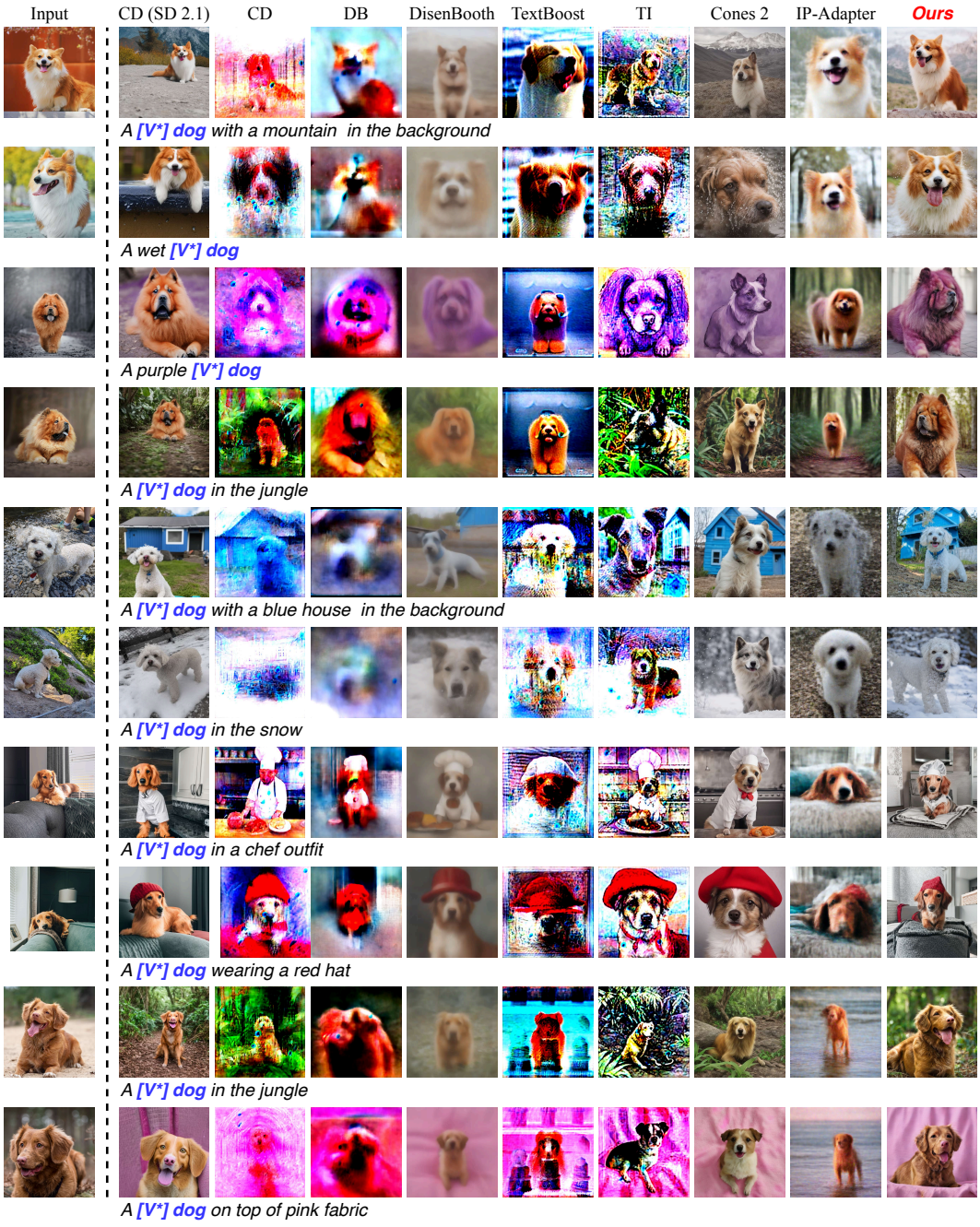

Figure 7. Our method *OPAD* (last column) compared with existing methods applied to the *1-SDP* setup with SDTurbo [77] as the one-step diffusion backbone. One representive concept image is shown on the left-most column. (Part 2)

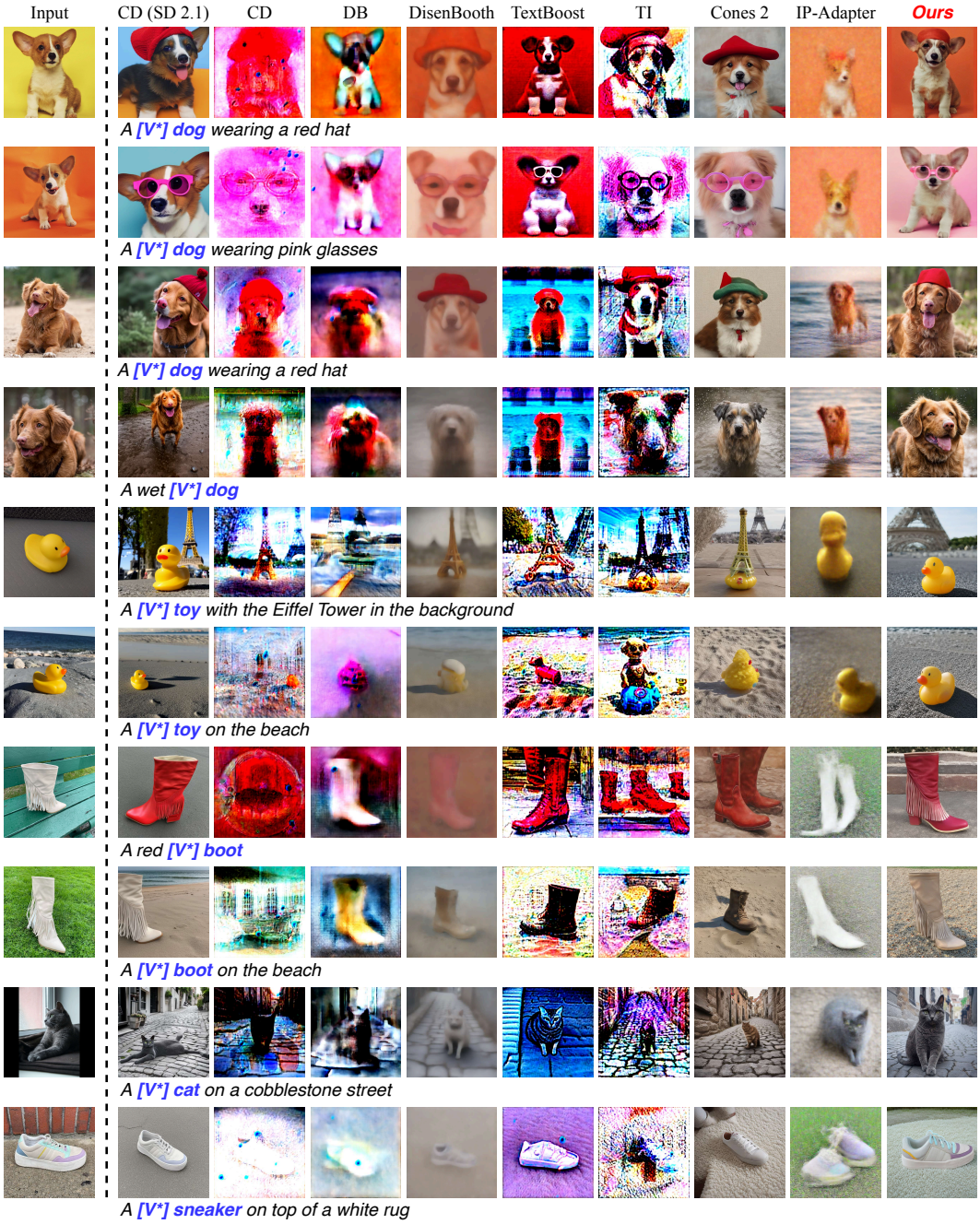

Figure 8. Our method *OPAD* (last column) compared with existing methods applied to the *1-SDP* setup with SDTurbo [77] as the one-step diffusion backbone. One representive concept image is shown on the left-most column. (Part 3)

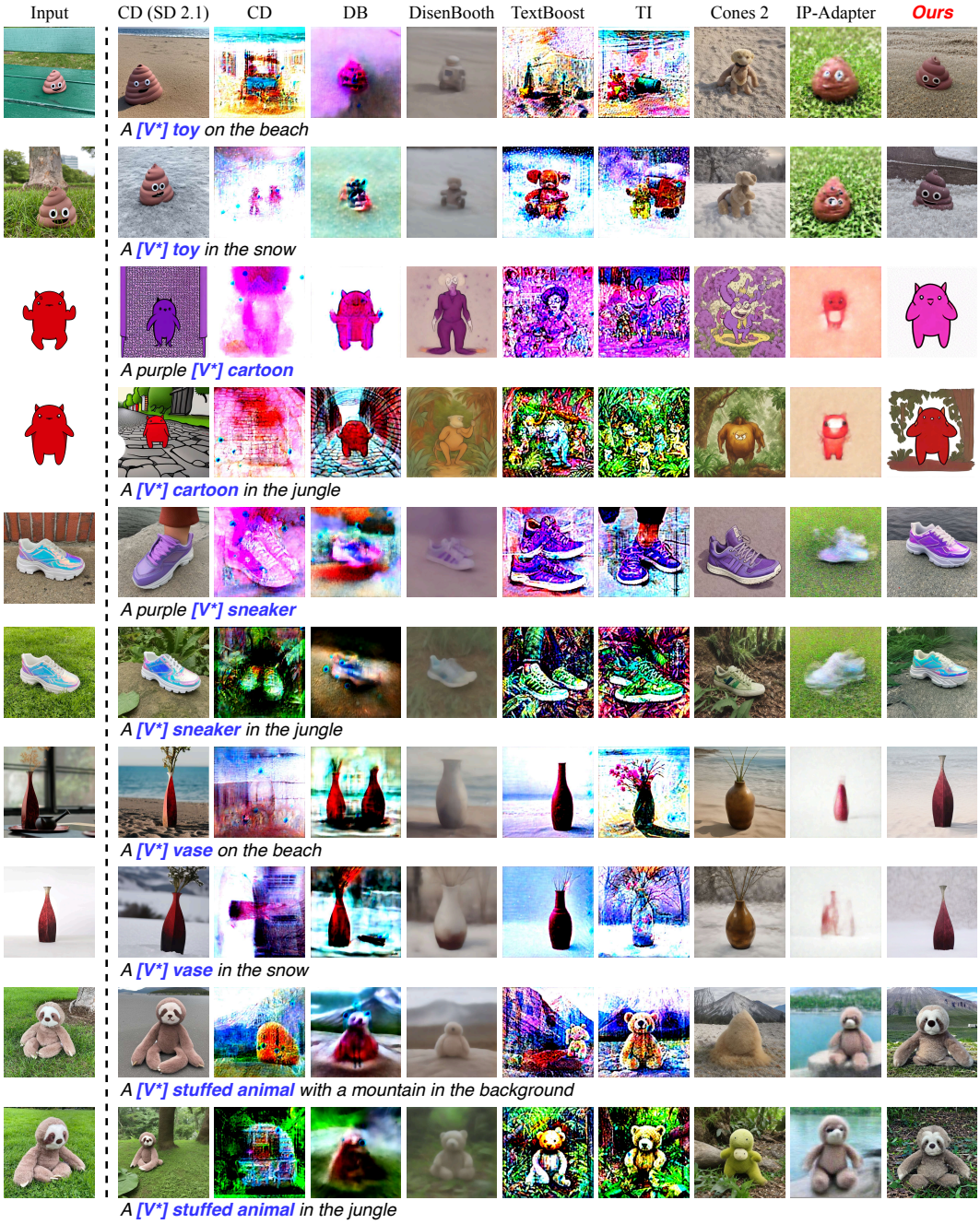

Figure 9. Our method *OPAD* (last column) compared with existing methods applied to the *1-SDP* setup with SDTurbo [77] as the one-step diffusion backbone. One representive concept image is shown on the left-most column. (Part 4)

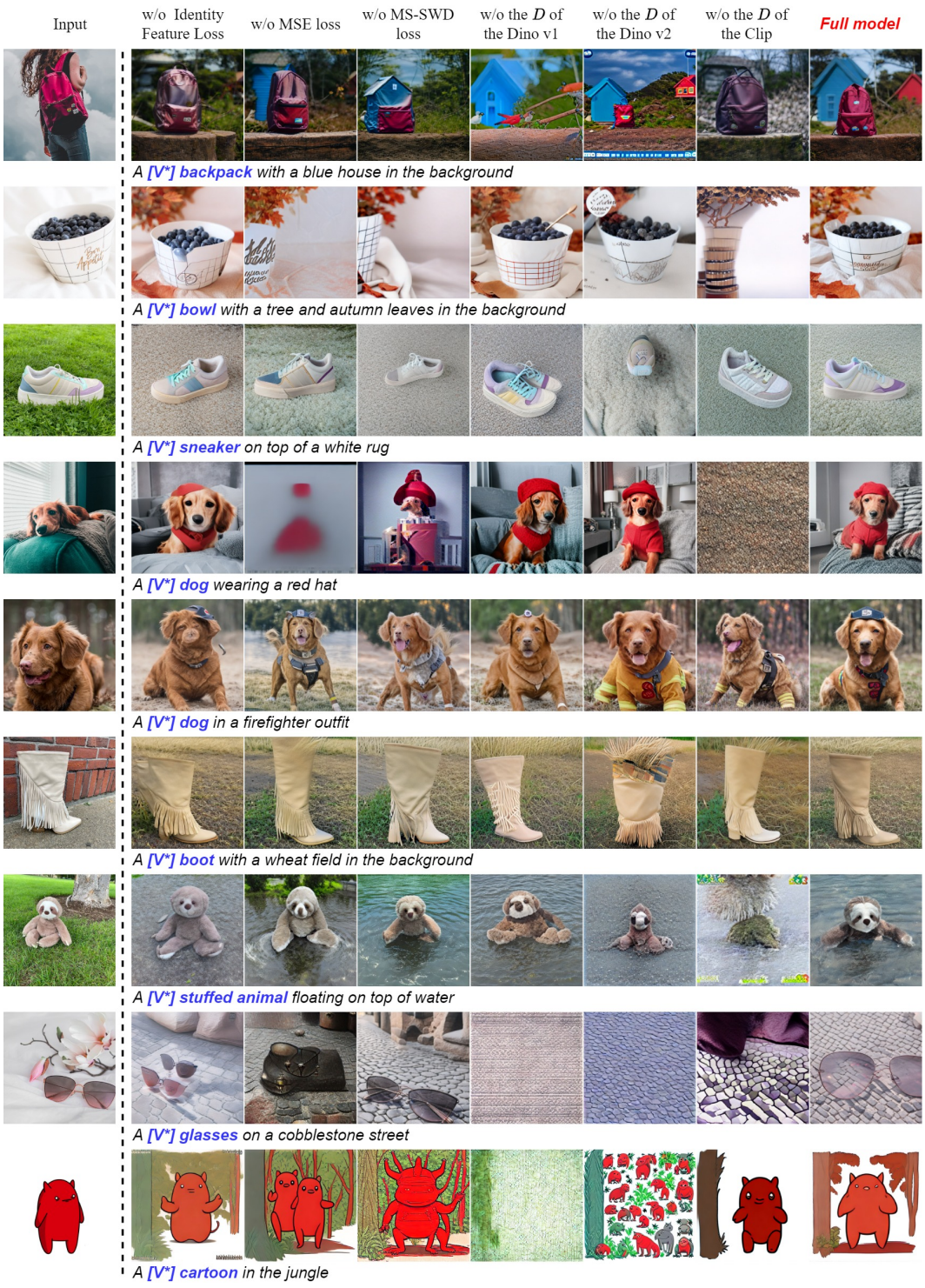

Figure 10. Qualitative results of the extended ablation study. D denotes the discriminator.

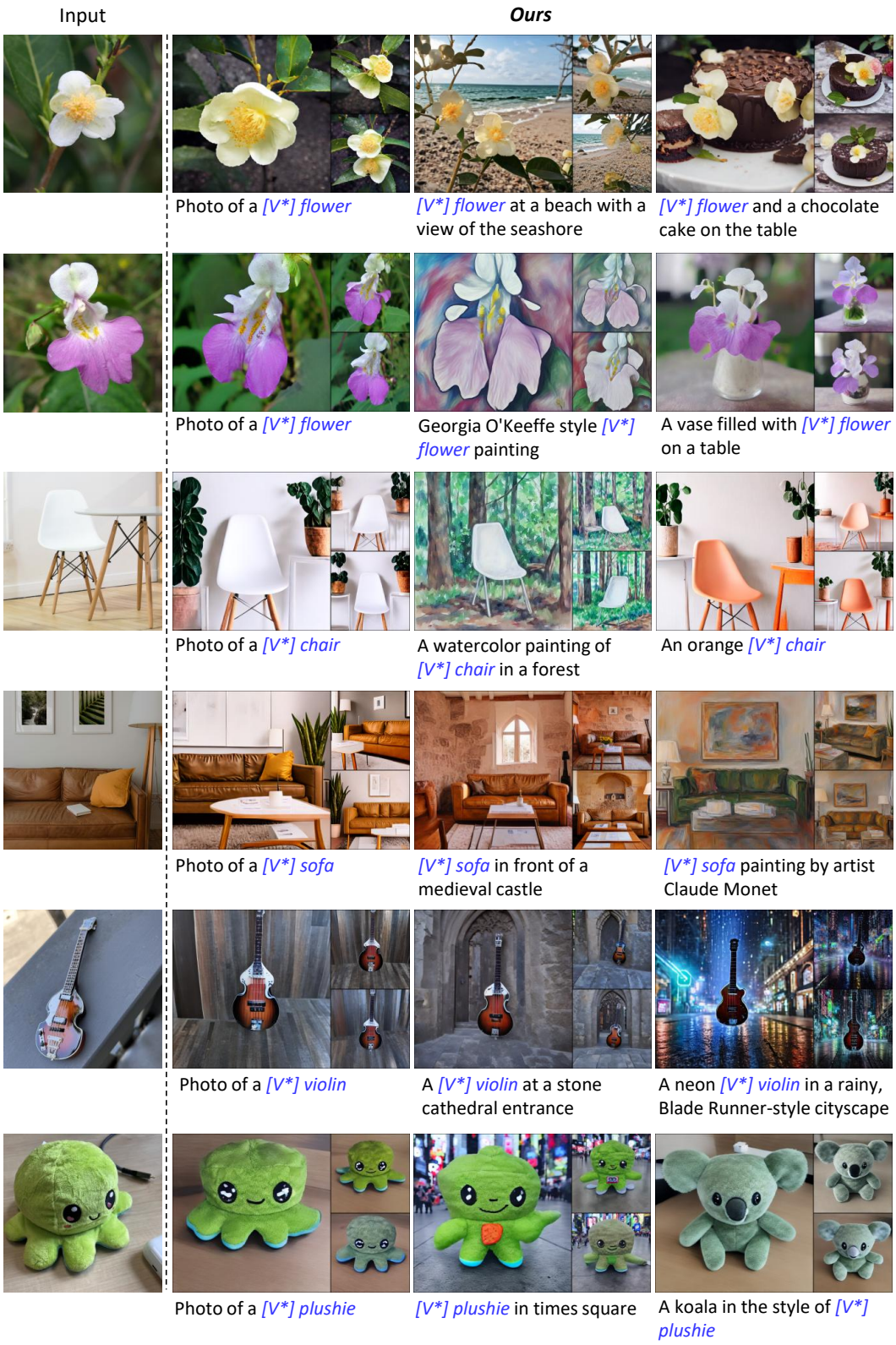

Figure 11. Qualitative results of *OPAD* on the CustomConcept101 dataset. Our method demonstrates strong generalization across a variety of concepts and prompt styles. (Part 1)

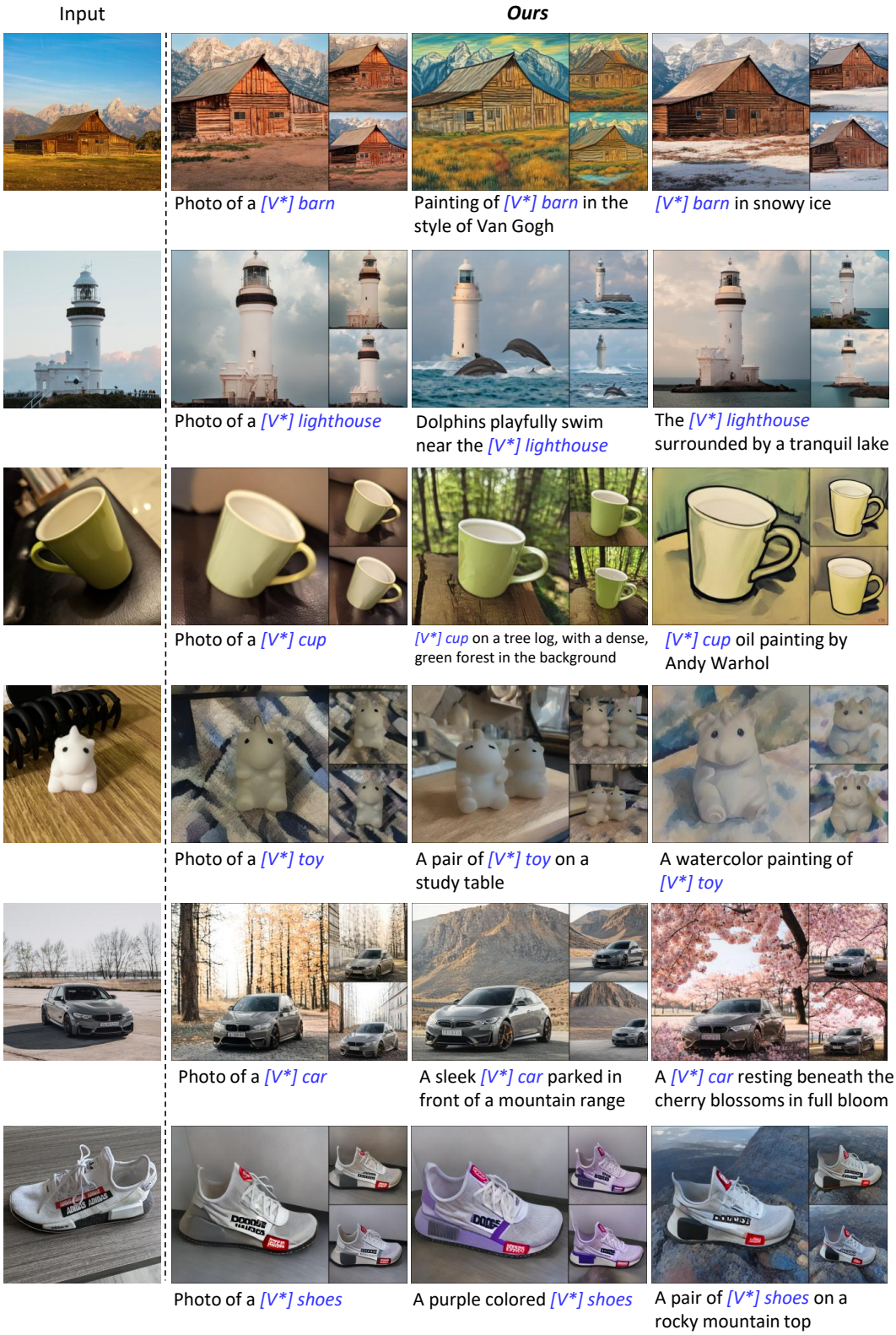

Figure 12. Qualitative results of *OPAD* on the CustomConcept101 dataset. Our method demonstrates strong generalization across a variety of concepts and prompt styles. (Part 2).