Seeing What Matters: Visual Preference Policy Optimization for Visual Generation

Supplementary Material

Due to the page constraint of the main paper, additional methodological details, parameter settings, and extended qualitative and quantitative results are provided in the supplementary material. The content is organized into the following parts:

- The computation of features corresponding to different backbone choices within Perceptual Structruring Module (PSM), along with visualizations of the allocation maps and the results from different ViPO variants.
- More quantitative and qualitative comparisons.
- The hyperparameter configurations used during training.

A. Details of PSM

Computation Details. The Perceptual Structuring Module (PSM) enriches the scalar reward signal by modeling human visual preference through pretrained vision backbones. As detailed in Section 3.2, PSM extracts perceptual features and constructs allocation maps that guide the redistribution of advantages across spatio-temporal locations. In the experiments, DINOv2 [24], ResNet [7], and SAM [17] are respectively adopted as perceptual backbones for evaluating the effectiveness of PSM.

For Transformer-based perception models such as DINOv2 and SAM, $\mathbf{\Phi}$ outputs patch-level features $\mathbf{F} \in \mathbb{R}^{N\times D}$, where $N = H_p \times W_p$, $H_p = H/p$, $W_p = W/p$, p is the patch size, and D is the feature dimension. Because each patch embedding is semantically homogeneous, Principal Component Analysis (PCA) is applied to $\mathbf{F}$ to retain the top-K components:

$$\mathbf{Z} = \mathrm{PCA}(\mathbf{F}) \in \mathbb{R}^{N\times K}, \quad \boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_K), \tag{10}$$

where λ_j denotes the explained variance ratio of the j-th component. Each column $z_j \in \mathbb{R}^N$ represents the j-th principal component.

PCA decomposes the embedding space into semantic directions, providing a compact representation of visual importance. For each component z_j, values are normalized and inverted so that regions with lower PCA projections (often corresponding to salient content) receive higher importance:

$$z'_j = \frac{\max(z_j) - z_j}{\max(z_j) - \min(z_j)}, \quad j = 1, 2, \ldots, K. \tag{11}$$

The aggregated semantic map $\mathbf{S}$ is obtained by reshaping the PCA-projected features into a 2D map. A variance-weighted scheme is commonly used:

$$\mathbf{S} = \mathrm{Reshape}\left(\sum_{j=1}^{K} \lambda_j z'_j\right) \in \mathbb{R}^{H_p\times W_p} \tag{12}$$

For CNN-based models [7], intermediate feature maps $\mathbf{F} \in \mathbb{R}^{C\times H_f\times W_f}$ are extracted from a designated layer, where $\mathbf{F}_c \in \mathbb{R}^{H_f\times W_f}$ denotes the c-th channel. An activation map $\mathbf{S} \in \mathbb{R}^{H_f\times W_f}$ is obtained via channel-weighted aggregation:

$$\mathbf{S} = \sum_{c=1}^{C} \alpha_c \mathbf{F}_c, \tag{13}$$

where the weights $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_C] \in \mathbb{R}^C$ are derived from global average pooling followed by softmax:

$$\boldsymbol{\alpha} = \mathrm{softmax}\left(\frac{1}{H_f W_f}\sum_{i=1}^{H_f}\sum_{j=1}^{W_f} \mathbf{F}_{:,i,j}\right) \tag{14}$$

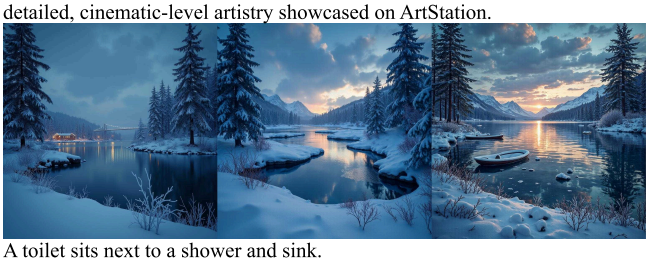
The resulting map $\mathbf{S}$ is further optionally smoothed and upsampled to the latent spatial resolution, yielding the final allocation map $\mathbf{M}$ used in PSM.

In our experiments, ResNet-50 are adopted as the CNN backbone and extract features from `layer4`, since this layer provides a good balance between semantic richness and spatial resolution, making it suitable for constructing allocation maps.

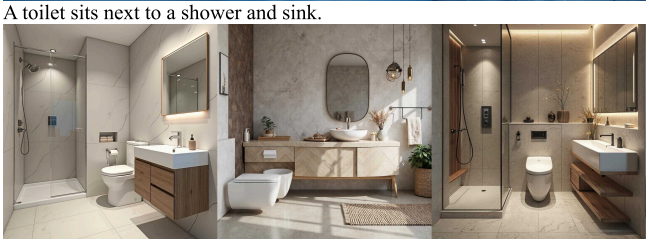
Visualization Examples. Additional visualizations are provided to illustrate the behavior of PSM under different perception backbones. Figure 8 (a) shows the final allocation maps generated by DINOv2, ResNet, and SAM, respectively. Figure 8 (b) further visualizes the first three principal components extracted from DINOv2 embeddings. In both subfigures, brighter colors indicate higher allocation weights.

It can be observed that DINOv2 produces maps that closely align with human visual preference, often assigning higher weights to semantically salient regions that typically attract immediate human attention, such as primary objects or dominant visual entities. In contrast, ResNet tends to respond to texture and spatial structure, yielding more distributed activations. SAM emphasizes boundary-aware regions, focusing on contours and segmentable areas as shown in the third and forth rows of Figure 8 (a). Moreover, for maps derived by SAM, high allocation weights do

A vividly realistic depiction of a snowy Swedish lake at night with hyper-detailed, cinematic-level artistry showcased on ArtStation.

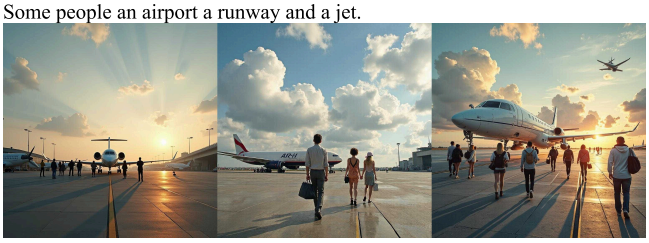


A toilet sits next to a shower and sink.



A close-up portrait of a cute anime girl with extremely detailed eyes, featured as a key visual in official media.

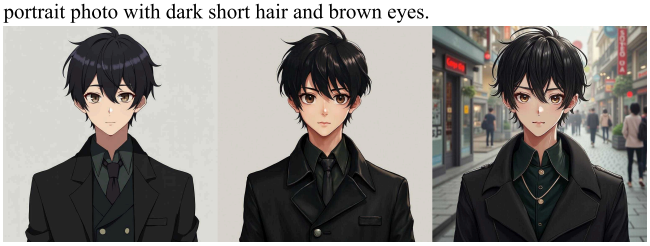
Some people an airport a runway and a jet.



A fire-type Pokemon is depicted in concept art found on ArtStation.

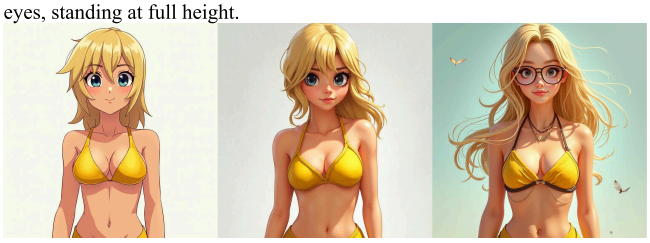
A still image of Johnny Bravo in Twin Peaks (1990) television show.

A male teen wearing a dark formal overcoat and anime style, depicted in a portrait photo with dark short hair and brown eyes.



Anime illustration of Gundam mech suit on Pixiv.

Tracer game character wearing a yellow bikini with blonde hair and black eyes, standing at full height.



Fine art exhibit in a white cube by Marcel Broodthaers.

A close-up anime portrait of Sailor Moon against a grey background.

A stop sign out in the middle of nowhere.

Figure 6. More qualitative comparison results for Flux. Each group of images, from left to right, shows the output from Flux, DanceGRPO, and our ViPO. As observed, DanceGRPO optimization generally introduces richer details and improved composition. In contrast, ViPO achieves more refined enhancements, delivering superior visual quality and finer improvements.

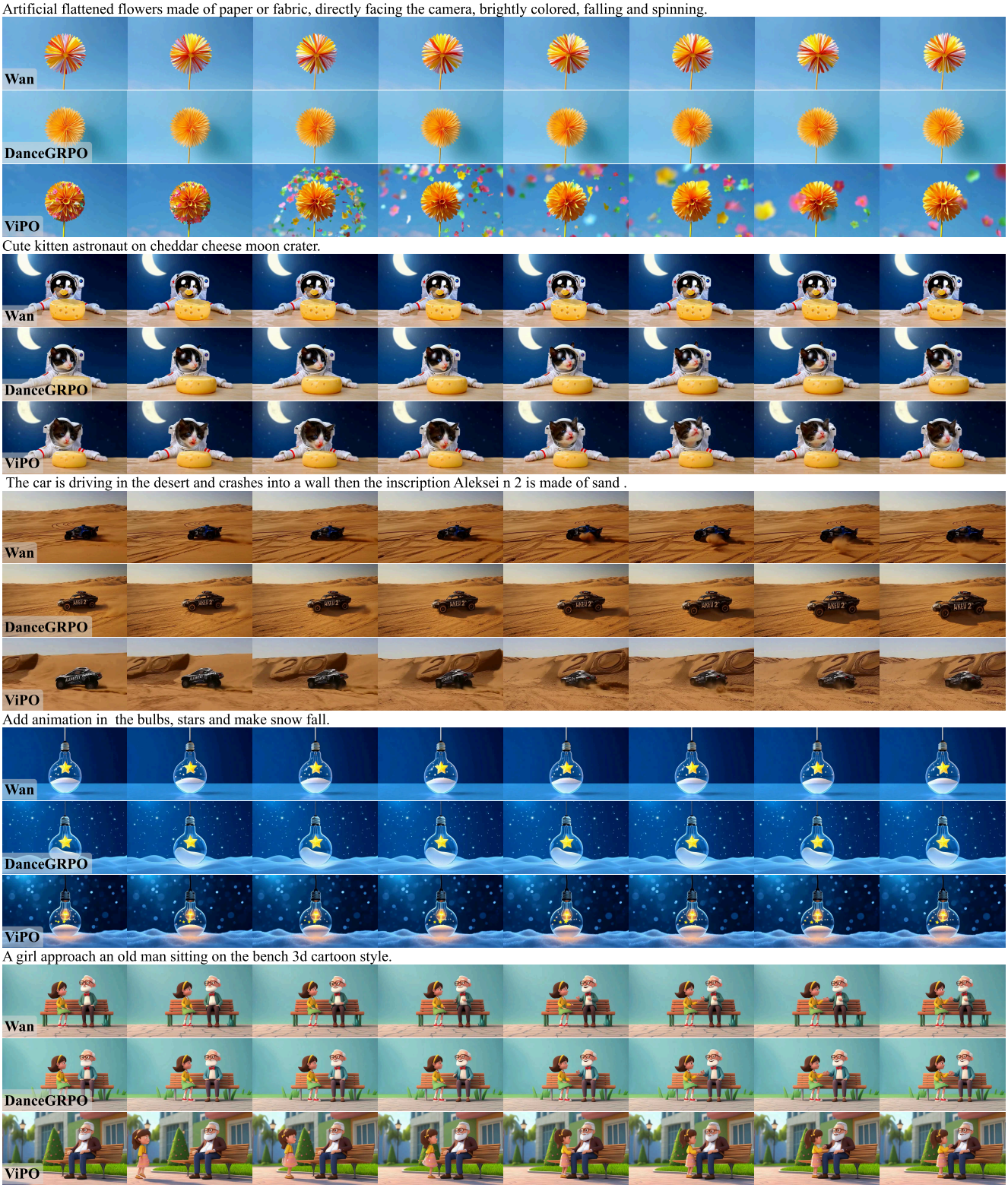
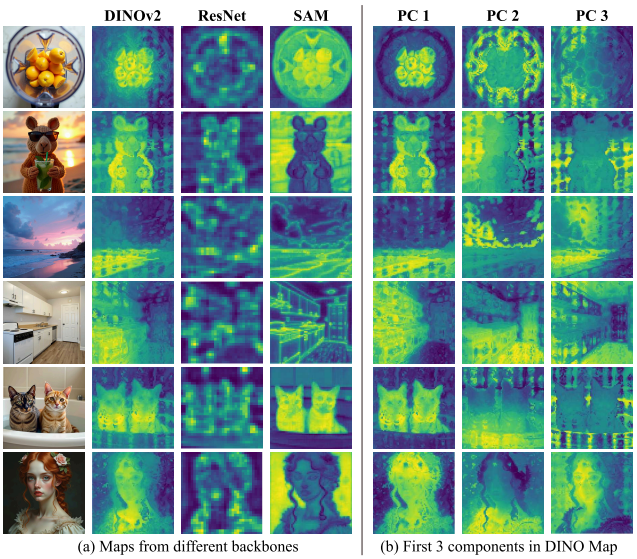


Figure 7. More qualitative comparison of video generation. For each group of sequences, the rows correspond to outputs from Wan2.1, DanceGRPO, and ViPO, respectively. As shown, DanceGRPO tends to enhance visual detail and yields moderate improvements in dynamic fidelity. In contrast, ViPO achieves more substantial gains in motion quality and visual realism, while further strengthening semantic alignment with the prompts.



(a) Maps from different backbones (b) First 3 components in DINO Map

Figure 8. Visualization of allocation maps. (a) Allocation maps produced by the PSM with different vision backbones. From left to right: the original image generated by Flux, followed by maps obtained using DINOv2, ResNet, and SAM. (b) Visualization of the top three principal components that compose the allocation map derived from DINOv2.

not always correspond to semantic foreground. For example, in the second and sixth rows, background regions receive stronger emphasis, while in the fifth row, the cat is clearly highlighted.

Figure 8 (b) visualizes the first three principal directions extracted from DINOv2 embeddings. Each component corresponds to a distinct semantic region, and their ordering generally reflects the progression of visual saliency, beginning with primary objects and extending to secondary structures or contextual cues. These components are subsequently integrated through a weighted aggregation to construct the final allocation map, enabling PSM to highlight perceptually meaningful regions in a manner aligned with human visual preference.

Figure 9 presents results from ViPO variants trained on Flux with different perception backbones. The comparisons show that backbone choice influences how advantages are adaptively allocated according to visual content, which in turn affects generation quality and further supports the effectiveness of perceptual structuring in preference-aware optimization.

Figure 9. Visualization of results obtained with different ViPO variants. From left to right: ViPO with DINOv2 as the PSM backbone, ViPO with ResNet, and ViPO with SAM.

B. More Results

In this section, more detailed evaluation results on Wan2.1 are provided. As shown in Table 5, ViPO achieves a substantial improvement on *Dynamic Degree*, and also yields gains in *Imaging Quality*, which is consistent with the qualitative results. Furthermore, several semantics-related dimensions, including *Multiple Objects*, *Spatial Relationship*, and *Temporal Style*, are improved. The enhancement in *Overall Consistency* further demonstrates the superiority of our approach.

Additional qualitative visualizations are also provided. Figure 6 illustrates extended results on Flux, where ViPO

Table 5. Quantitative comparison across detailed evaluation dimensions in VBench. ViPO consistently achieves superior performance across most dimensions.

VBench	Wan2.1	DanceGRPO	ViPO
Dynamic Degree	52.77	45.83	**63.89**
Imaging Quality	67.90	68.31	**68.88**
Multiple Objects	69.96	63.18	**74.70**
Color	89.20	86.51	88.97
Spatial Relationship	72.94	71.18	**81.44**
Temporal Style	24.12	24.14	**24.25**
Appearance Style	21.51	20.91	21.39
Scene	32.48	29.14	31.90
Overall Consistency	25.06	25.02	**25.32**

consistently produces outputs with richer details and improved aesthetics. Figure 7 shows extended qualitative comparisons on video generation, where ViPO achieves noticeable improvements in motion fidelity, visual quality, and semantic alignment. Interestingly, Although the Text Alignment score is not explicitly included as a reward signal, semantic alignment still shows improvement. likely as a side effect of optimizing for different regions of visual preference, which indirectly enhances semantic consistency. More qualitative comparisons for both video and image results are included in the supplementary MP4 file.

In addition, further qualitative visualizations based on the redness reward are provided in Figure 10. The results show that ViPO better preserves the semantic content of the images compared to baselines, maintaining object identity and visual coherence under preference optimization. These results further confirm the effectiveness of our approach in aligning visual generation with human visual preference.

C. Training Details

The parameter η controls the level of randomness in SDE sampling. In the reverse-time SDE formulation (Equation 3), the stochastic term is instantiated as $\varepsilon_t = \eta\sqrt{\Delta t}$, where Δt denotes the step size in the noise schedule. For Flux, η is set to 0.3, while for Wan2.1 it is set to 0.25. The learning rate was configured as 1×10^{-5} for Flux and 5×10^{-6} for Wan2.1. During training, backpropagation is not performed through all sampling steps; instead, a timestep fraction of 0.6 is used, meaning that only 60% of the timesteps contribute to gradient updates. All samples within a group are generated from the same initialization noise to ensure consistency. In the training objective, the clipping range for the importance ratio ρ^p is set to 1×10^{-4}. For Wan2.1, videos are sampled with 53 frames at 16 FPS, while the reward model processes them at 2 FPS.

Figure 10. More comparison of results using the redness reward. The left column shows outputs from Flux without RL fine-tuning, the middle column presents results from ViPO, and the right column displays results from DanceGRPO. The comparisons indicate that ViPO better preserves the semantic content of the images, maintaining object identity and visual coherence under preference optimization.