Supplement of Mind the Generative Details: Direct Localized Detail Preference Optimization for Video Diffusion Models

Zitong Huang[1,*] Kaidong Zhang[2,*] Yukang Ding[2,*,†] Chao Gao[2] Rui Ding[2]
Ying Chen[2] Wangmeng Zuo[1,†]

[1]Harbin Institute of Technology
[2]Alibaba Group - Taobao & Tmall Group
zitonghuang99@gmail.com

1. 3D Mask Generation Algorithm for Negative Videos Generation

As described in the main text, the negative samples in our LocalDPO are obtained by applying localized corruption to real videos. To select the regions to corrupt, we propose a Bézier curve–based localized region corruption algorithm, which is shown in Alg .1.

Generally, our mask generation strategy is grounded in the principle of structured randomness: rather than using arbitrary pixel-level noise or simplistic geometric primitives (e.g., rectangles or ellipses), we generate temporally plausible occlusions by modeling them as smooth, closed contours with controllable irregularity. The core idea is to first construct a compact, non-convex shape through stochastic corruption of a circular template, then embed it at a random location within the video frame. This ensures that the resulting masks mimic real-world occluders—such as objects or foreground entities—that are typically compact, connected, and exhibit organic boundaries. By decoupling shape generation (via Bézier-spline-based contours) from spatial placement, our method offers both diversity and physical plausibility for region-aware video corruption. Specifically, k anchor points are sampled on a perturbed circle in polar coordinates, where the radial distance of each point is uniformly randomized within $[1-\rho, 1+\rho]$ to introduce shape irregularity. The resulting point set is then normalized by its axis-aligned bounding box and rescaled to a prescribed proposal region of size $h \times w$. This resized shape is randomly translated within a full video frame of size $H \times W$ by sampling a valid top-left offset. Then smoothness is enforced by connecting consecutive anchor points with cubic Bézier curves, where control points are placed along the chord directions with a fixed scaling factor α. Finally, the closed spline is rasterized onto the $H \times W$ grid to produce a binary mask $R \in \{0,1\}^{H \times W}$, where pixels inside or on the contour are set to 1 and others to 0. In practice, for each sample, k is randomly sampled from the range 6 to 8, ρ is randomly sampled from the interval $[0.6, 0.8]$, α is randomly set within $[0.2, 0.4]$, and h and w are randomly sampled from $[H/3, H]$ and $[W/3, W]$, respectively.

2. General Statistics of the Real Videos Dataset

2.1. Overview

Following the data-construction pipeline [3, 13, 14] and the filtering protocols [4, 8, 10–12, 15, 16], we curate a large dataset containing initial video clips from Pexels [1]. Subsequent content-tag filtering and human annotation yield 63K high-quality clips characterized by high aesthetic, high resolution, diverse scenes, and stable motion. Using a structured captioning schema [9, 13], we annotate each clip with Qwen2.5-VL [2].

2.2. Preprocessing Pipeline of Real-World Videos

To facilitate rigorous evaluation of video generation models, we construct a large-scale, high-quality video dataset from a real-world source. This section details the systematic pipeline for its collection, filtering, and annotation.

2.2.1. Data Source

Our primary data source is from Pexels [1], an extensive repository of royalty-free stock videos. We choose Pexels for its vast diversity in subjects, scenes, and motion patterns, as well as its high technical quality (HD, 4K formats). Our selection process aims to create a challenging and varied dataset using a keyword-based search strategy.

*Equal contribution
†Corresponding author

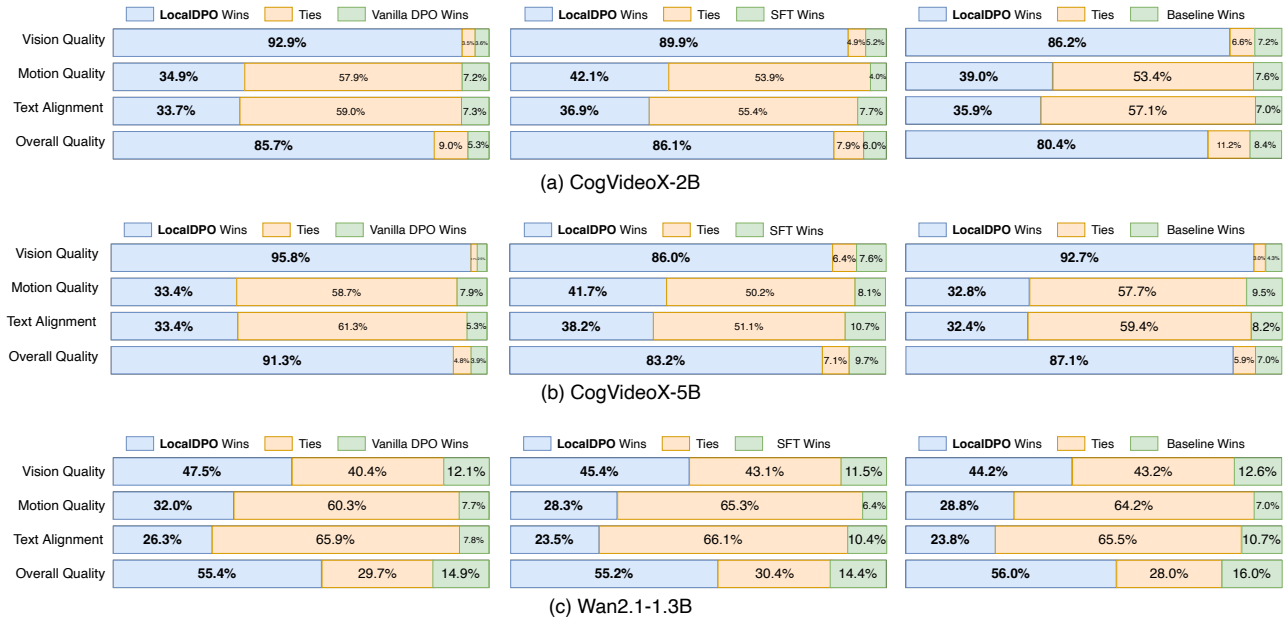

Figure 1. Human evaluation of LocalDPO *vs.* Baseline, SFT and Vanilla DPO on CogvideoX-2B [17], CogvideoX-5B [17] and Wan2.1-1.3B [13]. LocalDPO achieves the best results on all dimensions of human evaluation.

2.2.2. Video Selection Criteria

Our selection process is guided by the objective of creating a dataset that is both diverse and challenging. We employ a keyword-based search strategy with the following criteria:

Scene Diversity: A mix of environments, including keywords like *"indoor," "outdoor," "city,"* and *"nature."*

Motion Complexity: A spectrum from static shots to highly dynamic content, using keywords such as *"walking," "running,"* and *"slow motion."*

Subject Matter: A balance of subjects including *"people," "animals," "vehicles,"* and *"objects."*

Technical Quality: Only videos with a minimum resolution of 1080p and standard frame rates (24-60 FPS) are considered.

2.2.3. Data Filtering and Quality Assurance

To ensure a high standard of quality, every video is passed through a multi-stage automated filtering pipeline. Videos are discarded if they fail to meet predefined quality thresholds, assessed using the following state-of-the-art methods:

Technical Quality: The DOVER model [16] is used to assess a wide range of technical artifacts, providing a robust measure of overall fidelity.

Clarity: The MUSIQ model [4], a no-reference image quality assessor, is employed to ensure high sharpness and filter out blurry content.

Aesthetics: A pre-trained aesthetic scoring model [8] is utilized to evaluate the perceptual and artistic appeal of each frame.

Motion Smoothness: The "vmafmotion" filter from FFmpeg and [11] are applied to quantify motion, ensuring camera stability and removing clips with excessively shaky movements.

Text and Watermark Detection: An OCR-based approach combining SigLIP [12] for region proposal and GOT [15] for text recognition are used to detect and remove on-screen watermarks.

Shot Integrity: The TransNetV2 model [10] is utilized to identify and exclude videos containing scene transitions, ensuring each video clip contains a single, continuous shot.

2.2.4. Caption Annotation Pipeline

We generate descriptive captions for each video using a state-of-the-art Video Large Language Model (VLLM), Qwen2.5-VL-7B [2]. To elicit professional-grade descriptions, we design a detailed prompt that instructs the model to analyze key visual elements (subject, motion, scene) and adopt specific narrative constraints, such as describing camera work from a photographer's perspective and avoiding phrases like "the video shows." The prompt is presented as follows:

> *"Please describe the subject, motion, background, scene, camera motion, and style of this video in detail. Describe the camera motion as a professional photographer. If there are multi-*

Algorithm 1 Generate Binary Mask from Random Closed Contour

Require: Number of primary vertices $k \in \mathbb{Z}_+$, corruption ratio $\rho \in (0, 1)$, proposal region size (h, w), video frame size (H, W)

Ensure: Binary mask $R \in \{0, 1\}^{H \times W}$

// Step 1: Sample anchor points on a perturbed circle
for $j = 0$ to $k - 1$ **do**
  Compute base angle: $\phi_j \leftarrow \frac{2\pi j}{k}$
  Sample radial offset: $r_j \leftarrow 1 - \rho + 2\rho \cdot u_j$, where $u_j \sim \mathcal{U}(0, 1)$
  Set anchor point: $\mathbf{a}_j \leftarrow r_j \cdot \big(\cos\phi_j,\ \sin\phi_j\big)^\top$
end for
// Step 2: Compute axis-aligned bounding box and normalize to (h, w)
$x_{\min} \leftarrow \min_j a_j^{(x)}, \quad x_{\max} \leftarrow \max_j a_j^{(x)}$
$y_{\min} \leftarrow \min_j a_j^{(y)}, \quad y_{\max} \leftarrow \max_j a_j^{(y)}$
$w_{\text{bbox}} \leftarrow x_{\max} - x_{\min}, \quad h_{\text{bbox}} \leftarrow y_{\max} - y_{\min}$
for $j = 0$ to $k - 1$ **do**
  $a_j^{(x)} \leftarrow \frac{a_j^{(x)} - x_{\min}}{w_{\text{bbox}}} \cdot w$
  $a_j^{(y)} \leftarrow \frac{a_j^{(y)} - y_{\min}}{h_{\text{bbox}}} \cdot h$
end for
// Step 3: Randomly place the resized shape in the (H, W) canvas
Sample top-left corner: $x_0 \sim \mathcal{U}\big(0,\ H - h\big), \quad y_0 \sim \mathcal{U}\big(0,\ W - w\big)$
for $j = 0$ to $k - 1$ **do**
  $a_j^{(x)} \leftarrow a_j^{(x)} + y_0$ ▷ image x-axis is horizontal (column)
  $a_j^{(y)} \leftarrow a_j^{(y)} + x_0$ ▷ image y-axis is vertical (row)
end for
// Step 4: Construct cubic Bézier segments between consecutive anchors
Let $\mathbf{a}_k \equiv \mathbf{a}_0$ (cyclic indexing)
for $j = 0$ to $k - 1$ **do**
  Compute direction vectors: $\mathbf{d}_{j+1} = \mathbf{a}_{j+1} - \mathbf{a}_j$
  Place first control point near $\mathbf{a}_j$ along outgoing direction: $\mathbf{c}_j^{(1)} \leftarrow \mathbf{a}_j + \alpha \cdot \mathbf{d}_{j+1}$
  Place second control point near $\mathbf{a}_{j+1}$ along incoming direction: $\mathbf{c}_j^{(2)} \leftarrow \mathbf{a}_{j+1} - \alpha \cdot \mathbf{d}_{j+1}$
  // $\alpha > 0$ controls curve smoothness (e.g., $\alpha = 1/3$)
end for
// Step 5: Form closed spline and rasterize
Define closed contour $\mathcal{C}$ as the union of k cubic Bézier curves, each parameterized by $\big(\mathbf{a}_j,\ \mathbf{c}_j^{(1)},\ \mathbf{c}_j^{(2)},\ \mathbf{a}_{j+1}\big)$
Rasterize $\mathcal{C}$ onto a 2D grid of size (H, W): set pixel $(i, j) = 1$ if it lies inside or on $\mathcal{C}$, else 0
return binary mask R

ple subjects, clearly describe their spatial relationship. Do not use "the video" or "this video" as the subject of the sentence; directly start the sentence with the subject in the video. Keep the description clear and to the point, avoiding unnecessary details or repetition. Provide a coherent description without breaking it into sections or lists."

2.2.5. Dataset Statistics

Our pipeline results in a dataset including **63K diverse video clips**. The technical specifications and thematic distribution are presented below. Tab. 1 summarizes the key metrics of the dataset, while Fig. 2 visualizes the category distribution, confirming a well-balanced composition for robust evaluation.

Table 1. Statistics of the curated data on key attributes.

Metric	Value / Range
Total Videos	63K
Resolution	1080p, 4K
Frame Rate (FPS)	24-60
Average Duration (s)	9.5

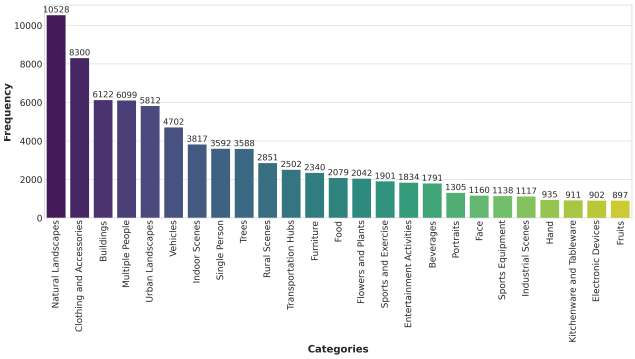

Figure 2. Category Distribution of the constructed video dataset.

3. Additional Human Evaluation

We present additional human evaluation results for CogVideoX-2B [17], CogVideoX-5B [17], and Wan2.1-1.3B [13] across four evaluation dimensions: Visual Quality (VQ), Motion Quality (MQ), Text Alignment (TA), and Overall Quality in Fig. 1. We compare our method with the baseline model, Supervised Fine-Tuning (SFT) and Vanilla DPO for comprehensive human evaluation.. Generally, the voting distributions consistently indicate that our method is preferred by a larger proportion of participants than either method in all four dimensions, further corroborating the superiority of our approach in human perceptual evaluation.

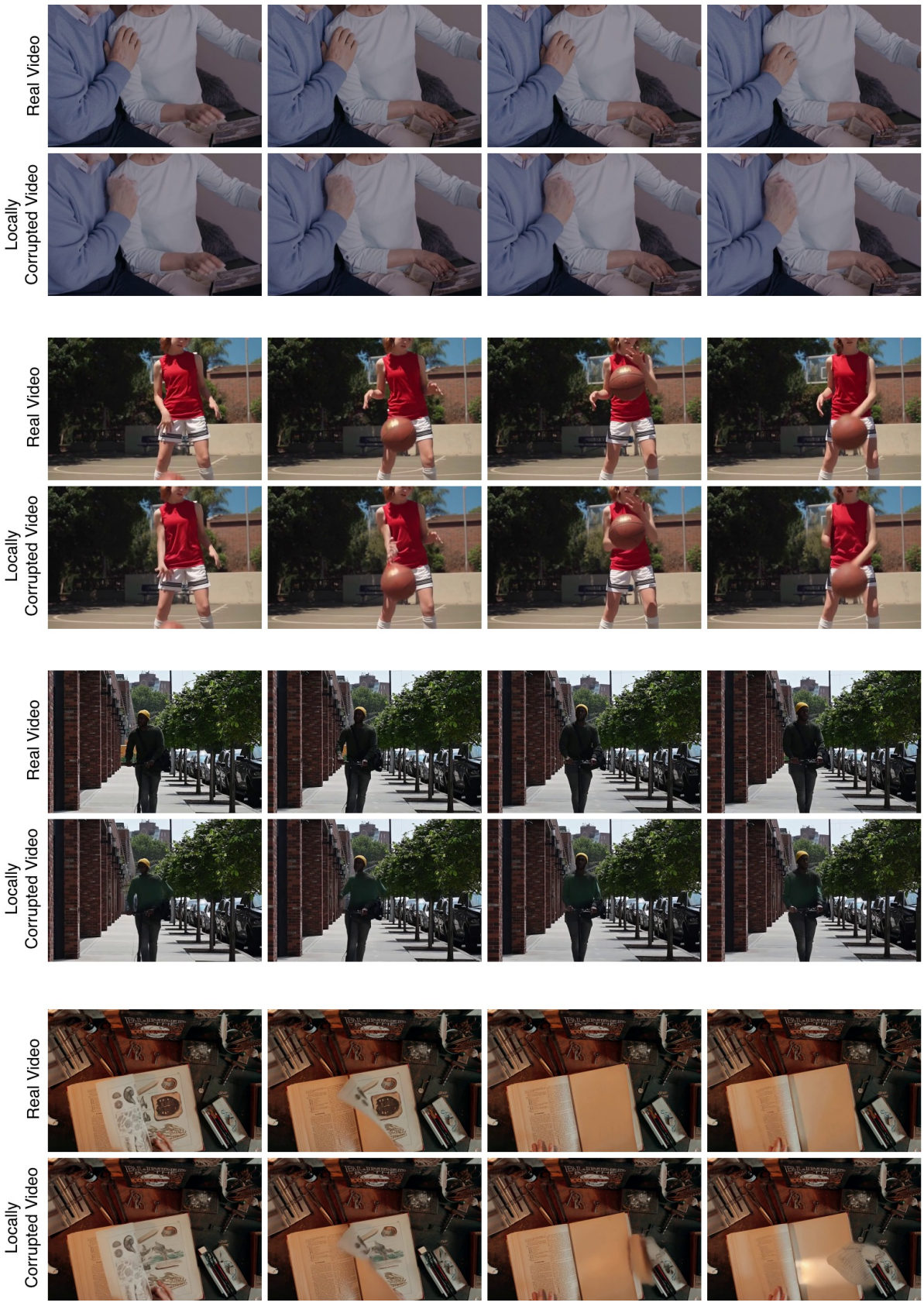

Figure 3. Visualization of generated locally corrupted videos.

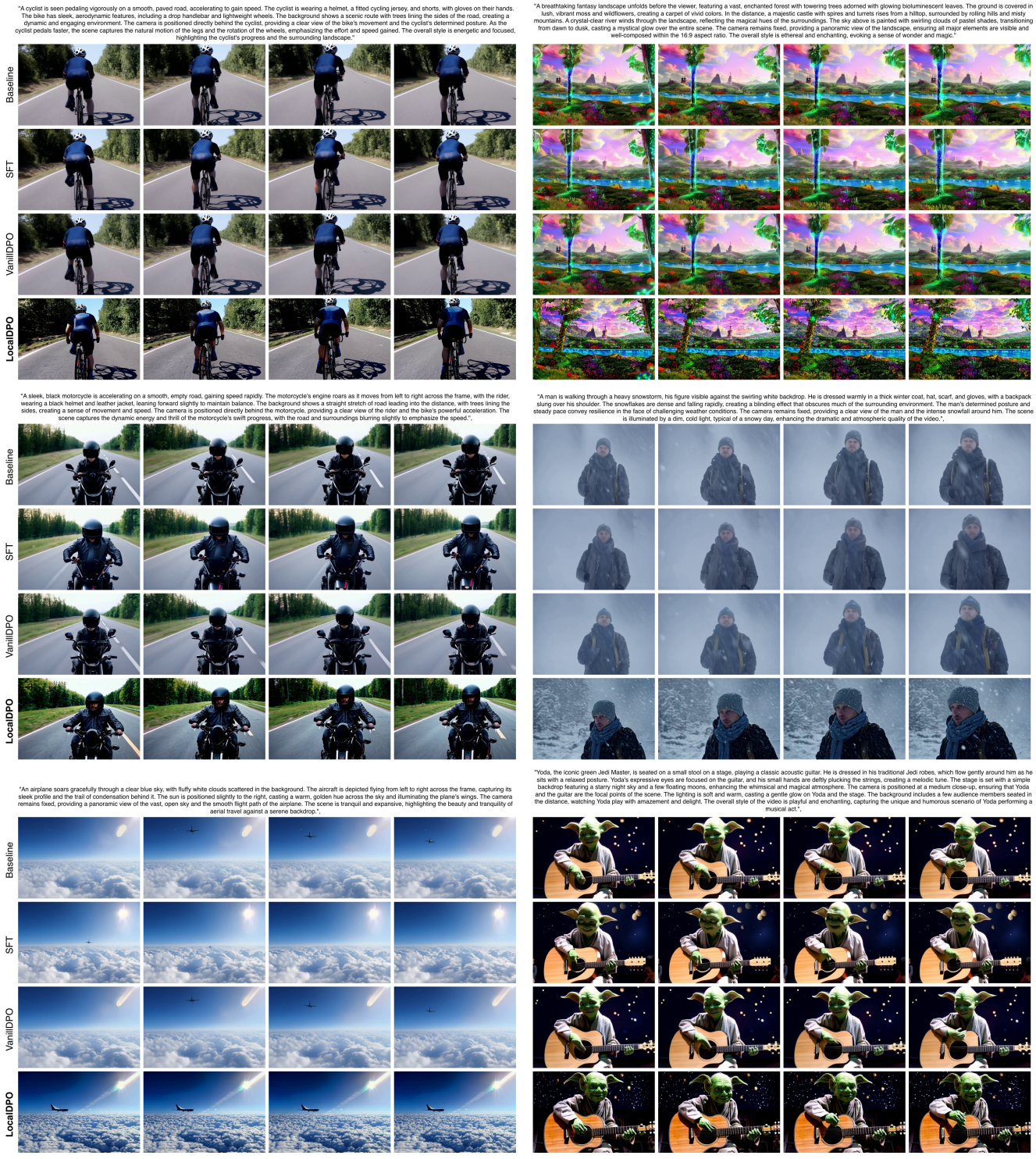

Figure 4. Visualization of LocalDPO *vs.* Baseline, SFT and VanillaDPO on CogvideoX-2B.

4. Visualization of the LocalDPO training pairs

In our LocalDPO, negative samples are constructed by applying localized corruption to the positive samples (i.e., real videos). In this subsection, we visualize the perturbed negative samples alongside their corresponding original videos (positive samples), as shown in Fig. 3. It is

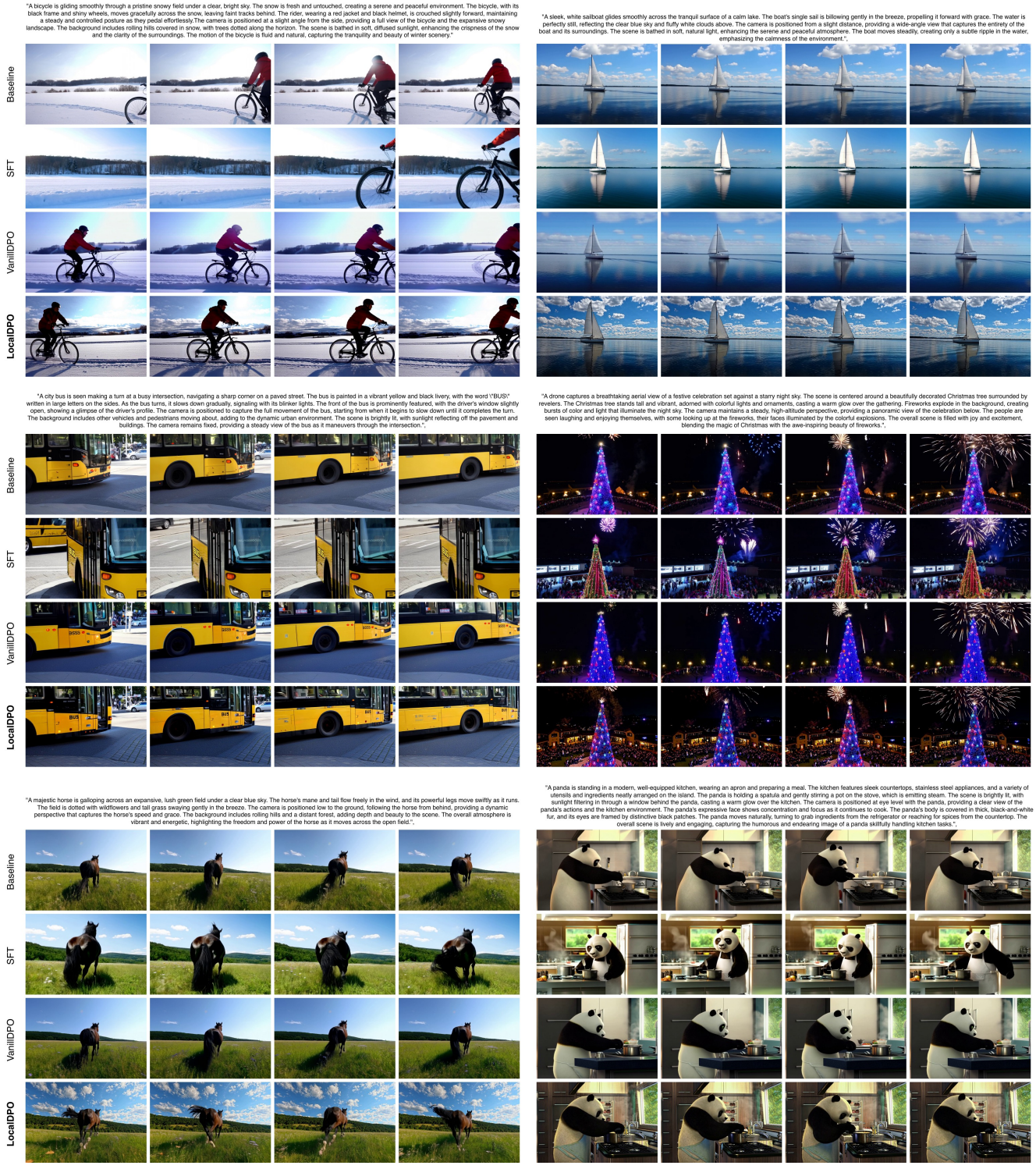

Figure 5. Visualization of LocalDPO *vs.* Baseline, SFT and VanillaDPO on CogvideoX-5B.

clearly observable that the perturbed regions often exhibit artifacts, distortions, or blurriness compared to the authentic video content, thereby forming reasonable training pairs that encode fine-grained, local-level preferences. Moreover, these imperfections precisely reflect the current limitations of pre-trained video generation models; consequently, train-

Figure 6. Visualization of LocalDPO *vs.* Baseline, SFT and VanillaDPO on Wan2.1-1.3B.

ing with such negative samples provides explicit feedback that effectively guides the model toward gradual improvement.

5. Limitations and Future Work

Our current approach generates spatio-temporal masks via random Bézier curves, which ensures diversity in cor-

rupted regions but may lacks semantic awareness. Specifically, the corruptions are not tailored to particular object categories or semantic parts (e.g., faces, hands, or vehicles), potentially overlooking critical regions where quality degradation most affects user perception. As a result, the preference signal may be less effective for improving generation fidelity of specific object classes.

In future work, we will incorporate vision foundation models, such as Grounding DINO [6] for object detection and SAM [5, 7] for segmentation, to guide mask placement towards semantically meaningful regions. This would enable targeted refinement of object-level realism and controllability in text-to-video generation.

6. More Qualitative Comparisons

We present additional visual comparisons between our method and other methods, including the baseline, SFT, and vanilla DPO. Fig 4, Fig 5, and Fig 6 show comparisons based on CogVideoX-2B, CogVideoX-5B, and Wan2.1-1.3B, respectively. Clearly, our LocalDPO generates videos with higher visual quality, better captures fine-grained details of the subject, and more faithfully adheres to the appearance. These consistency results strongly demonstrate the effectiveness of our LocalDPO, particularly in enhancing video quality and preserving subject details.

References

[1] Pexels. https://www.pexels.com/, 2025.10. accessed: 2025-11-01. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2

[3] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *NeurIPS*, 37:48955–48970, 2024. 1

[4] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pages 5148–5157, 2021. 1, 2

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 8

[6] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, pages 38–55. Springer, 2024. 8

[7] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *ICLR*, 2025. 8

[8] 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. 1, 2

[9] Team Seawead, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. Seaweed-7b: Cost-effective training of video generation foundation model. *arXiv preprint arXiv:2504.08685*, 2025. 1

[10] Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *ACMMM*, pages 11218–11221, 2024. 1, 2

[11] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419. Springer, 2020. 2

[12] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 1, 2

[13] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 2, 3

[14] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *CVPR*, pages 8428–8437, 2025. 1

[15] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024. 1, 2

[16] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, pages 20144–20154, 2023. 1, 2

[17] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025. 2, 3