When Numbers Speak: Aligning Textual Numerals and Visual Instances in Text-to-Video Diffusion Models

Supplementary Material

S1. Additional Results

Compatibility with CogVideoX [5]. To substantiate the generalizability and robustness of our method beyond a single model architecture, we evaluate our method on CogVideoX-5B, which employs a Multi-Modal Diffusion Transformer (MMDiT). Unlike vanilla DiTs in Wan models [4], MMDiT employs a unified global attention mechanism over concatenated visual-textual tokens without a dedicated cross-attention module. To bridge this gap, we adapt our strategy in Sec. 4.1 of the manuscript by decomposing the unified attention into distinct components. The video-to-video attention is treated as self-attention, while the text-to-video attention sub-matrix is extracted as cross-attention.

As shown in Tab. 1, quantitative results demonstrate a consistent and significant improvement in numerical accuracy when our method is applied to CogVideoX-5B. Specifically, CogVideoX-5B achieves only 40.2% accuracy under minimal settings, while Seed search and Prompt enhancement provide limited gains of only 2.5% and 2.3%, respectively. In contrast, NUMINA substantially elevates the performance to 44.4% using simple prompts and a single generation pass. Furthermore, our method improves overall generation quality, improving the TC and CLIP scores to 80.2% and 35.4%, respectively. This successful extension to MMDiT further confirms the general applicability of our training-free approach across different implementations of the architecture.

Integration with enhancement strategies. As shown in Tab. 1 of the manuscript, our method alone achieves substantial improvements on CountBench. We further demonstrate that NUMINA is fully compatible with prompt enhancement and seed search, which represent the most accessible techniques for boosting counting accuracy. By integrating our method with these enhancement strategies, we achieve the best performance with 54.2% counting accuracy, reported in Tab. 2. This combined approach significantly surpasses all compared methods, including our standalone NUMINA (49.7%), prompt enhancement (47.2%), and seed search (45.5%). In particular, it also enables the 1.3B model to outperform larger baseline models, including Wan2.2-5B at 47.8% and Wan2.1-14B at 53.6%. These results establish our approach as a superior alternative to existing workflows, providing a more effective solution for the challenging counting alignment in video generation.

Evaluation on VBench [2] metric. To assess the temporal stability of the generated object instances, we adopt the Subject-Consistency metric from VBench. For each instance, we extract DINO [1] features in all frames and compute the cosine similarity with both the first frame and the preceding frame. The two similarities are averaged, and the final video-level score is obtained by averaging over all non-initial frames. We report the mean score across instances. As shown in Tab. 3, our method achieves competitive performance on this metric, indicating that the edited instances remain temporally stable and visually coherent. This result further validates the reliability of our TC metric, as both measurements capture complementary aspects of temporal coherence. In addition, our counting accuracy follows the Generative Numeracy evaluation protocol in T2V-CompBench [3], ensuring that our overall evaluation framework is both consistent and reliable.

Analysis on no-reference addition. We analyze the effectiveness of adding missing instances when no reference instances are available. This presents a particularly challenging setting where baseline models typically fail to generate the required objects. As shown in Tab. 4, the no-intervention baseline achieves only 48.8% accuracy without layout refinements in such cases. To address this limitation, we compare two geometric priors for layout refinement: a circular template and a rectangular alternative of equivalent

Table 1. Evaluation results on CogVideoX [5].

Models	CountAcc (%)	TC (%)	CLIP Score
CogVideoX-5B [5] (81 frames, 1360×768)			
CogVideoX-5B	40.2	78.1	34.8
+ Seed search	42.7(+2.5)	78.3(+0.2)	34.8(-0.0)
+ Prompt enhancement	42.5(+2.3)	79.0(+0.9)	34.5(-0.3)
+ NUMINA (**ours**)	**44.4**(+4.2)	**80.2**(+2.1)	**35.4**(+0.6)

Table 2. Ablation on combined methods.

Models	CountAcc (%)	TC (%)	CLIP Score
Wan2.1-1.3B [4] (81 frames, 832×480)			
Wan2.1-1.3B	42.3	81.2	33.9
+ Seed search	45.5(+3.2)	82.3(+1.1)	34.6(+0.7)
+ Prompt enhancement	47.2(+4.9)	82.1(+0.9)	33.7(-0.2)
+ NUMINA (**ours**)	49.7(+7.4)	83.4(+2.2)	**35.6**(+1.7)
+ Combined method (**ours**)	**54.2**(+11.9)	**83.6**(+2.4)	35.5(+1.6)

Table 3. VBench [2] Subject-Consistency scores.

Models	Baseline	+ NUMINA (**ours**)
Wan2.1-1.3B	83.1	**83.6**(+0.5)
Wan2.1-14B	84.3	**84.7**(+0.4)
Wan2.2-5B	83.4	**83.5**(+0.1)
CogVideoX-5B	84.6	**84.6**(+0.0)

Table 4. Ablation on strategy for no-reference addition.

Method	CountAcc (%)	TC (%)
Baseline	42.3	81.2
No intervention	48.8(+6.5)	83.0(+1.8)
Rectangle	49.5(+7.2)	83.3(+2.1)
Circle	**49.7**(+7.4)	**83.4**(+2.2)

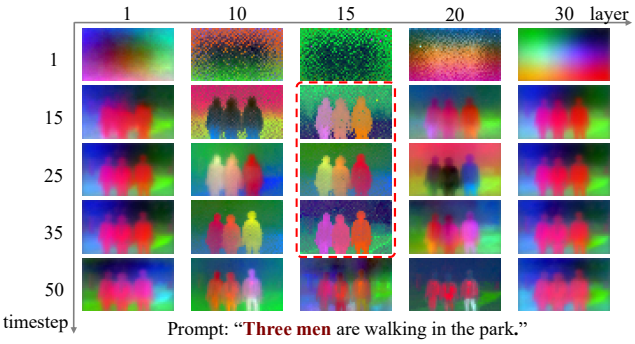


Figure 1. PCA visualization across timesteps and layers.

area. Experimental results demonstrate the effectiveness of both strategies, with the rectangular prior reaching 49.5% accuracy and the circular prior achieving 49.7%. In practice, we employ the circular prior as described in Sec. 4.2 of the manuscript. This design minimizes structural assumptions, granting T2V models the flexibility to interpret and form the most contextually plausible objects.

Analysis on layout localization. We next analyze the feasibility of layout localization based on Wan2.1-1.3B [4]. As visualized in Fig. 1, our analysis reveals clear instance-separable attention patterns during denoising. These discriminative layouts emerge most distinctly at middle denoising steps, with intermediate layers providing the sharpest spatial separation of object instances. We accordingly set $t^\star = 20$ and $\ell^\star = 15$ to balance efficiency and accuracy. By performing layout localization at this point and early stopping, we reduce the denoising steps for pre-generation by approximately 60% without significantly sacrificing accuracy, as quantified in Fig. 7 of the manuscript. This early termination delivers significant computational savings, particularly for larger models. The same relative proportions can be directly applied to other model architectures through straightforward scaling.

Analysis on hyperparameters. We emphasize that our hyperparameters are generic and are largely set without exhaustive tuning. Selections of layer and timestep vary solely due to intrinsic model differences (e.g., the total number of inference steps) rather than specific heuristic design. We uniformly set $t^\star = 20$ and $\ell^\star = 15$ in this section for a fair ablation study. As detailed in Tab. 5, our method maintains stable performance across a wide range of hyperparameter values.

Analysis on the object addition/removal. We finally analyze the effect of layout-guided generation operations, i.e., object addition and removal. Tab. 6 shows that addition alone significantly boosts accuracy by 5.4%, while removal yields a smaller 1.5% gain. This suggests that the baseline model primarily struggles with object omission, making addition the more impactful correction. Furthermore, combining both operations achieves the highest accuracy, slightly exceeding the sum of individual gains, proving a synergistic effect between the two complementary guidance methods.

Table 5. Ablation results for different hyperparameter values.

λ / CountAcc (%)	τ / CountAcc (%)	k / CountAcc (%)
4 / 49.3	0.1 / 48.4	0.5 / 48.2
8 / 49.7	0.2 / 49.7	0.8 / 49.7
16 / 49.5	0.3 / 49.2	1.0 / 49.2

Table 6. Ablation on object addition or removal.

Addition	Removal	CountAcc (%)	TC (%)
Baseline		42.3	81.2
✓		47.7(+5.4)	83.0(+1.8)
	✓	43.8(+1.5)	82.4(+1.2)
✓	✓	**49.7**(+7.4)	**83.4**(+2.2)

Table 7. VBench Aesthetic & Imaging Quality scores.

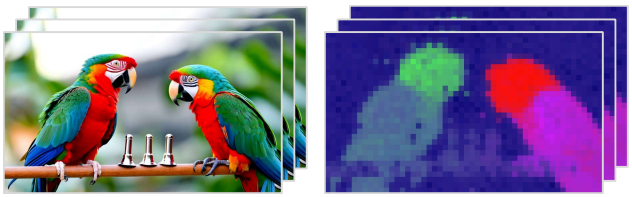
Method	Imaging↑	Aesthetic↑
Wan2.1-1.3B	71.3%	61.5%
+NUMINA	70.9%	63.5%

Evaluation of visual quality. We evaluate visual generation quality using VBench (Aesthetic & Imaging Quality). As shown in Tab. 7, our method maintains comparable or even superior metric scores, introducing no degradation in video generation quality while significantly enhancing numerical alignment, which is further confirmed by the user study, demonstrating the quality of our approach.

User study. We conduct a blind user study involving 10 participants (balanced gender ratio) using 100 pairs of randomly sampled videos. Participants are asked to evaluate both visual quality and instruction following. The results show a 61% preference for our method versus 39% for the baseline. This clear preference confirms that our method delivers not only better objective metric performance but also a superior user experience.

S2. More Visualization

Additional demos. We provide more comprehensive qualitative comparisons in Fig. 3, showcasing our method's effectiveness across different model architectures. The consolidated visualization presents successful numerical alignment cases on Wan2.1 [4] and CogVideoX [5], demonstrating consistent improvement in generating accurate object counts. These cross-architecture validations collectively confirm our method's strong generalizability and practical utility for enhancing numerical accuracy in text-to-video generation systems. More video demos can be found on



Prompt: " **Three parrots** mimicking **three whistles**."

Figure 2. A failure case of NUMINA. The parrots' heads become decoupled from their bodies in layout construction.

our project page.

Failure cases. A characteristic failure mode of our method occurs when instance-separable attention heads focus excessively on the most salient parts of an object (e.g., an animal's head) rather than its entirety, as demonstrated by the representative failure case in Fig. 2. This leads to an over-segmented layout where parts of a single instance are mistaken for multiple objects, ultimately propagating an irrecoverable error into the final video output. This limitation underscores the challenge of defining instances solely via raw attention and suggests the need for future work to incorporate more holistic perceptual grouping cues.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 9650–9660, 2021. 1

[2] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 1

[3] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 8406–8416, 2025. 1

[4] 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

[5] 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 *Proc. of Intl. Conf. on Learning Representations*, 2025. 1, 2

Figure 3. More representative examples where our method faithfully generates the specified number of objects.