Training-free Motion Factorization for Compositional Video Generation

Supplementary Material

In this supplementary material, we provide additional information in the following aspects:

A. preliminaries of our motion guidance module;

B. illustration of benchmark construction for compositional video generation;

C. observations from additional ablation studies;

D. quantitative evaluation on the synthetic dataset.

A. Preliminary

Our motion guidance module is implemented by applying motion-specific masks to update attention maps. As illustrated in Fig. 1, the behavior of the query differs across motion types:

- **Motionlessness.** When an instance remains static, the query aggregates visual patches from the same spatial location, anchored by a reference frame.
- **Rigid Motion.** For rigid movements, the query gathers information from semantically corresponding regions displaced along the motion trajectory. The trajectory is estimated from cross-frame shifts of bounding-box centers.
- **Non-rigid Motion.** For non-rigid deformation, neighboring frames may deviate substantially, so patch correspondence is determined by the local deformation magnitude rather than strict geometric displacement.

B. CVG Benchmark Construction

From the perspective of video descriptions, CVG can be categorized into **four linguistic modes**:

- **Coordinating Structure (CS).** Diverse instances or actions are presented in parallel, often joined by conjunctions, *e.g.*, "A child drawing and a dog sitting nearby.";
- **Quantitative Expression (QE).** This expression specifies the number of each instance class in a scene, *e.g.*, "Five birds flying in the sky.";
- **Collective Noun (CN).** Unlike QE, CN describes a generalized number of instances, without indicating an exact quantity quantity. In grammar, multiple components are viewed as a single unit, causing the verb to be typically singular, *e.g.*, "A team of soccer players is practicing on the field.";
- **Interactive Verbs (IV).** IV mainly emphasizes actions involving various objects interacting with each other. These verbs inherently imply a reciprocal relationship, suggesting a mutual influence or effect between the components, *e.g.*, "A child is throwing a frisbee to a dog."

Following such rules, we employ LLaMA-70B [1] to automatically identify matched video descriptions from publicly available datasets.

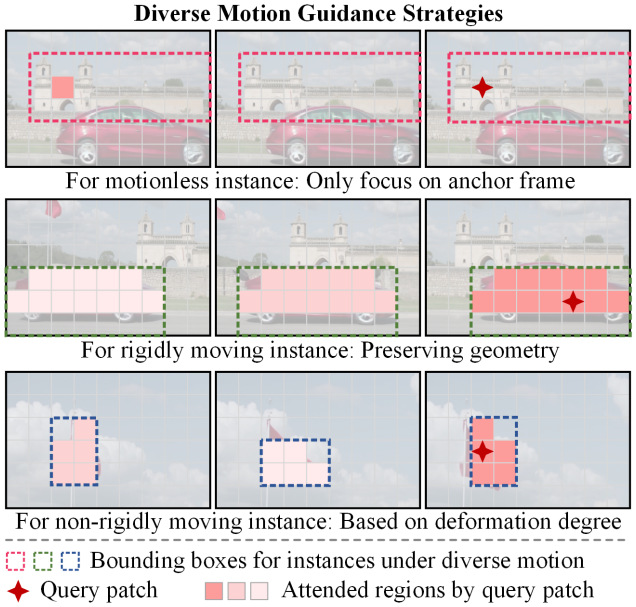

Figure 1. Case study of diverse motion guidance strategies. With spatiotemporal layouts as signals, motion is guided via cross-frame modulation between query patches and surrounding visual cues. For motionless instances, query patches attend only to spatially aligned regions in the anchor frame. For rigid motion, guidance preserves geometric invariance under displacement. For non-rigid motion, deformable regions highly sensitivity to spatial variations are emphasized.

Table 1. Analysis of motion guidance factor β based on CVGBench-p benchmark. Best/2nd best scores are **bolded**/underlined.

Param. β	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree
Baseline: VideoCrafter-v2.0 [3]					
$\beta = 100$	87.08%	92.34%	86.43%	87.72%	66.16%
$\beta = 10$	98.81%	**98.29%**	97.82%	98.79%	**78.24%**
$\beta = 1$	**98.83%**	98.26%	**97.84%**	**98.80%**	76.02%
$\beta = 0.1$	97.79%	97.20%	96.81%	98.32%	35.14%
Baseline: CogVideoX-2B [15]					
$\beta = 0.25$	98.41%	97.20%	98.36%	98.85%	85.77%
$\beta = 0.20$	98.44%	97.18%	**98.43%**	98.88%	86.49%
$\beta = 0.15$	**98.74%**	**98.23%**	98.38%	**98.94%**	87.06%
$\beta = 0.10$	98.45%	97.31%	98.42%	98.91%	**87.09%**

C. Additional Ablation Studies

Analysis of Guidance Factor. Using CVGBench-p benchmark, we analyze the hyperparameter β, which regulates motion guidance strength during denoising. As shown in Tab. 1, performance improves as β decreases from a large

Table 2. Ablation analysis of diverse backbones for motion reasoning based on CVGBench-p benchmark. Best scores are **bolded**.

Backbones	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree
Baseline: VideoCrafter-v2.0 [3]					
LLaMA-8B [4]	96.00%	96.60%	94.88%	96.23%	73.89%
LLaMA-70B [1]	**98.81%**	**98.29%**	**97.82%**	**98.79%**	**78.24%**
Baseline: CogVideoX-2B [15]					
LLaMA-8B [4]	97.49%	96.91%	97.53%	98.51%	83.27%
LLaMA-70B [1]	**98.74%**	**98.23%**	**98.38%**	**98.94%**	**87.06%**

Table 3. Ablation analysis of diverse motion guidance components based on CVGBench-p benchmark, including Reference Conditioned Guidance (RCG), Geometric Invariance Guidance (GIG), Spatial Deformation Guidance (SDG). Best scores are **bolded**.

RCG	GIG	SDG	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree
Baseline: VideoCrafter-v2.0 [3]							
✗	✗	✗	97.90%	97.10%	96.83%	98.47%	31.44%
✓	✗	✗	98.42%	97.75%	97.21%	98.55%	36.21%
✗	✓	✗	98.63%	98.08%	97.54%	98.69%	42.90%
✗	✗	✓	98.44%	97.83%	97.24%	98.56%	70.69%
✓	✓	✗	98.68%	98.14%	97.61%	98.71%	43.40%
✓	✓	✓	**98.81%**	**98.29%**	**97.82%**	**98.79%**	**78.24%**
Baseline: CogVideoX-2B [15]							
✗	✗	✗	93.01%	94.26%	96.12%	97.95%	81.52%
✓	✗	✗	97.47%	96.68%	97.53%	98.52%	82.30%
✗	✓	✗	97.57%	96.95%	97.57%	98.51%	82.65%
✗	✗	✓	97.62%	96.92%	97.51%	98.47%	84.62%
✓	✓	✗	98.11%	97.60%	97.92%	98.72%	84.00%
✓	✓	✓	**98.74%**	**98.23%**	**98.38%**	**98.94%**	**87.06%**

value, followed by degradation when β becomes overly small. Empirically, $\beta = 10$ and $\beta = 0.15$ are optimal values, when using VideoCrafter-v2.0 [3] and CogVideoX-2B [15] as baselines, respectively. These observations indicate that moderate strength yields the best performance, while extreme high or low values break the balance between consistency and dynamic. Accordingly, we set $\beta = 10$ for VideoCrafter-v2.0 and $\beta = 0.15$ for CogVideoX-2B.

Analysis on CVGBench-p benchmark. We conduct ablation studies on CVGBench-p benchmark (main-paper results are on CVGBench-m benchmark).

- As shown in Tab. 2, LLaMA-70B consistently achieves better results than the 8B variant, indicating the advantage of adopting larger backbones for motion reasoning.
- As shown in Tab. 3, combining all guidance components yields the best performance, showing that they provide complementary contributions to motion generation.

In summary, our key modules remain effective on the CVGBench-p benchmark.

D. Evaluation on Synthetic Dataset

We evaluate our framework on the synthetic benchmark T2V-CompBench [11], where each video description is generated by GPT-4 [2]. Specifically, we use "Spatial", "Motion", "Action", and "Interaction" categories. We evaluate "Spatial" by a detection approach [9]; assess "Motion" and "Action" with a multi-modal large language model [5], and measure "Interaction" leveraging a Tracking approach [6]. As shown in Tab. 4, compared to other generation models, our framework achieves the highest scores on "Motion" (0.2762), "Action" (0.6304), and "Interaction" (0.8048). This demonstrates a strong capability in modeling fine-grained dynamics and complex relationships. However, our framework yields worse performance on "Spatial" than LVD [8] and Vico [14], likely because these models employ explicit layout constraints.

Table 4. Quantitative Comparison on T2V-CompBench. Best/2nd best scores are **bolded**/underlined. † indicates compositional generation models.

Models	Spatial	Motion	Action	Interaction
modelScope [13]	0.4118	0.2408	0.3639	0.4613
LATTE [10]	0.4340	0.2155	0.4146	0.4146
Show-1 [16]	0.4544	0.2291	0.3881	0.6244
CogVideoX-5B [15]	0.5172	0.2658	0.5333	0.6069
Open-Sora-v1.2 [17]	0.5053	0.2468	0.4833	0.5039
T2V-Turbo-v2 [7]	0.5025	0.2556	0.6087	0.6439
LVD† [8]	**0.5469**	0.2699	0.4960	0.6100
VideoTetris† [12]	0.5148	0.2204	0.5280	0.7600
Vico† [14]	0.5432	0.2412	0.6020	0.7800
VideoCrafter-v2.0 [3]	0.4838	0.2259	0.5030	0.6365
+ Ours	0.5255	**0.2762**	**0.6304**	**0.8048**

References

[1] Llama 3.3. 2024. 1, 2

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[3] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7310–7320, 2024. 1, 2

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024. 2

[5] Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a vlm. *IEEE Access*, 2024. 2

[6] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. Dense optical tracking: Connecting the dots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19187–19197, 2024. 2

[7] Jiachen Li, Qian Long, Jian Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhu Chen, and William Yang Wang. T2v-

turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design. *arXiv preprint arXiv:2410.05677*, 2024. 2

[8] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[9] 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 *European Conference on Computer Vision (ECCV)*, pages 38–55. Springer, 2024. 2

[10] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 2

[11] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8406–8416, 2025. 2

[12] Ye Tian, Ling Yang, Haotian Yang, Yuan Gao, Yufan Deng, Xintao Wang, Zhaochen Yu, Xin Tao, Pengfei Wan, Di ZHANG, et al. Videotetris: Towards compositional text-to-video generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 29489–29513, 2024. 2

[13] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2

[14] Xingyi Yang and Xinchao Wang. Compositional video generation as flow equalization. *arXiv preprint arXiv:2407.06182*, 2024. 2

[15] 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. *arXiv preprint arXiv:2408.06072*, 2024. 1, 2

[16] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision (IJCV)*, pages 1–15, 2024. 2

[17] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 2