MotionStone: Decoupled Motion Intensity Modulation with Diffusion Transformer for Image-to-Video Generation

Supplementary Material

A. Implementation Details

We supplement more details of the training of motion estimator. For training the motion estimator, we utilize 8 A100 GPUs with batch size 64. The learning rate is set to 5×10^{-6}. To align with the training configuration of `MotionStone`, input videos are cropped to a resolution of 480×720 and sampled to 49 frames. The motion estimator is trained for 10,000 steps using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We set the weight of regression loss λ to 0.1.

B. Details on the Training Data for Motion Estimator

In this section, we provide more details on the training data for the motion estimator. We ask 15 annotators to participate in this annotation process. The annotators are asked to label video pairs from several aspects: First, they are asked to determine whether the two videos in a pair contain a moving object. A video is considered to have a moving object only if it features a foreground object in motion. Meanwhile, camera motion focuses on the global motion in the scene. If a video in the pair contains a moving object, it is labeled as 1; otherwise, it is labeled as 0. Note that comparisons of object motion between the two videos are only made when at least one video in the pair features a moving object. Next, annotators are tasked with labeling the relative magnitude of the object and camera motion in each video pair. If both videos contain object or camera motion, the corresponding item is annotated based on the annotators' subjective judgment. If only one video in the pair exhibits object or camera motion, the video with motion is considered significantly greater in the respective category. Specifically, we define the annotations as follows: if the first video shows significantly greater camera or object motion than the second one, it is labeled as 2; if it is only slightly greater, it is labeled as 1. Conversely, if the first video shows significantly or slightly less motion, it is labeled as -2 or -1, respectively. If neither video exhibits object or camera motion, the corresponding item is labeled as 0. During the training process using contrastive learning, this label is employed to amplify the motion differences between two videos. If a specific motion in the first video is significantly greater than that in the second, the corresponding loss is set to twice that of cases with a smaller difference.

After completing one round of annotation, we conduct a sampling check on 5,000 video pairs, reviewing 20% of them. The investigation achieves an accuracy rate of 95%, meeting the annotation standards. This demonstrates that the annotated data aligns well with human perception of the relative magnitude of object and camera motion in videos.

C. User Study on Comparisons with Existing Alternatives

Since the metrics in VBench [6] cannot fully evaluate the performance of the model, we conduct user studies. We ask 10 annotators to participate in this process. To ensure the generalization of the evaluation, we select a wide variety of real and animated images, including elements such as people, animals, camera movement, plants, and natural landscapes. Twenty image-text prompts are selected and processed by each compared method, including `MotionStone`, generating a total of 100 video clips. Each participant is presented with two videos generated by different methods for the same prompts and asked to choose the one that performed better in four aspects: *Text Consistency* evaluates if the motion and content follow the text prompt. *Image Consistency* assesses the ability to preserve the identity of the reference image. *Content Quality* determines the overall quality of video generation, including visual appeal, definition, and the logical coherence of the generated content. *Motion Quality* evaluates the plausibility and richness of the motion. The pairwise comparison is repeated for all combinations of videos, resulting in C_2^5 comparisons.

As shown in Tab. 1, our method demonstrates superior performance, particularly in terms of Text Consistency, Content Quality and Motion Quality. This highlights the effectiveness of our approach in text-based motion control and the generation of videos with content and motion that align more closely with human perception.

D. Evaluation Metrics

We select several metrics from VBench [6] for quantitative evaluation experiments, including *Background Consistency*, *Aesthetic Quality*, *Imaging Quality*, *Subject Consistency*, *Motion Smoothness*, *Dynamic Degree* and *Temporal Flickering*. It is important to note that, we utilize only its models and evaluation processes, excluding its prompt suite. Consequently, some metrics that strictly require the use of the prompt suite are omitted. The detailed information on each metric is introduced as follows.

Table 1. **Results of user study.** The best results for each column are **bold**. We ask annotators to rate videos based on four aspects: Text Consistency, which assesses how well the motion and content adhere to the textual descriptions; Image Consistency, which evaluates the ability to preserve the identity of the reference image; Content Quality, which focuses on inter-frame coherence and definition; and Motion Quality, which measures the plausibility and richness of the motion.

Method	I2VGEN-XL	SVD	AnimateAnything	CogVideoX-5B	`MotionStone`
Text Consistency ↑	32.50%	39.38%	25.00%	63.13%	**90%**
Image Consistency ↑	27.50%	36.88%	56.25%	62.50%	**66.88%**
Content Quality ↑	31.25%	45.63%	33.13%	63.13%	**76.88%**
Motion Quality ↑	26.25%	48.13%	39.38%	61.25%	**75.00%**

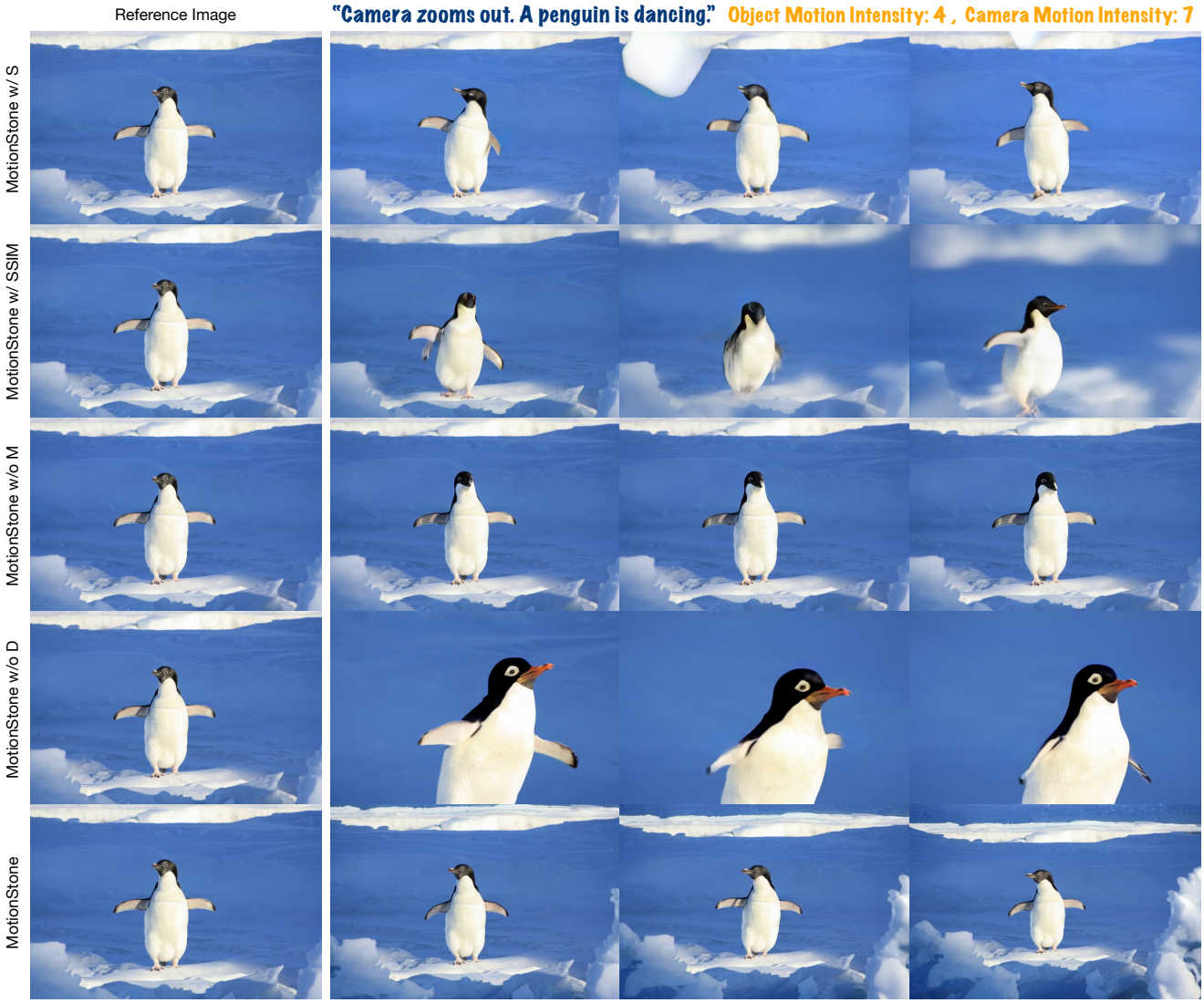

Figure 1. **Qualitative ablation for proposed modules.** Using inter-frame SSIM [3] and feature difference [4] (`MotionStone` *w/* SSIM and `MotionStone` *w/* S) causes varying degrees of unnatural background motion (In the first row, the snow block in the upper left corner of the third column appears. In the second row, background motion blur is observed.) and does not follow the camera motion described in the text prompt. Omitting the proposed motion estimator (`MotionStone` *w/o* M) and the decoupled injection method (`MotionStone` *w/o* D) results in issues such as generating static video and confusion or overlap between camera motion and object motion control, respectively. These approaches also fail to follow the camera motion described in the text prompt successfully.

Background Consistency. This metric measures the temporal consistency of the background scenes by calculating CLIP [10] feature similarity across frames.

Aesthetic Quality. This metric assesses the human-perceived artistic and aesthetic value of each video frame utilizing the LAION aesthetic predictor. This tool captures

various aesthetic dimensions, including composition, color richness and harmony, photorealism, naturalness, and the artistic quality of the video frames.

Imaging Quality. Imaging quality pertains to distortions such as over-exposure, noise, and blur observed in the generated frames. This metric measures this using the MUSIQ [8] image quality predictor, which is trained on the SPAQ [5] dataset.

Subject Consistency. This metric calculates the DINO [2] feature similarity across frames to evaluate the consistency of a subject's appearance throughout the video.

Motion Smoothness. Evaluating the smoothness of motion in generated videos and its adherence to real-world physical laws is crucial. To assess this, this metric leverages motion priors from the video frame interpolation model [9].

Dynamic Degree. As a completely static video might perform well in the previously mentioned temporal quality metrics, it is essential to assess the level of dynamics (i.e., the presence of significant motions) in the generated videos. To achieve this, this metric uses RAFT [11] to estimate the extent of dynamics in the synthesized outputs.

Temporal Flickering. Generated videos may display imperfect temporal consistency, particularly in local and high-frequency details. To quantify this, this metric extracts static frames and calculates the mean absolute difference between them.

E. Limitation

Although `MotionStone` has made notable progress in I2V generation and motion intensity control, it still faces several limitations. First, `MotionStone` is built upon CogVideoX, and due to constraints in memory and computational resources, it can only generate videos of approximately 6 seconds in length at a specific resolution. We believe that as the computational demands of foundational video generation models decrease in the future, `MotionStone` will be able to generate longer videos with higher resolutions. Second, with reduced computational resource requirements, it will be feasible to design a larger motion estimator and leverage more extensive training datasets to develop a more powerful model. The enhanced motion estimator could better assist I2V generation, and such advancements will lead to superior performance. Third, since our motion estimator is trained by assigning a single motion intensity to the object and camera motion for each video, it may not be able to control different motion intensities for multiple objects within a video or for different parts of a single object. We believe that as the annotation of training video captions becomes more fine-grained and the labeling of video motion intensity expands to multiple dimensions, our model will be able to achieve more precise and diverse motion control.

F. More Experiments

In this section, we first present additional ablation studies, including more detailed qualitative and quantitative experiments, as well as an evaluation of the motion strength error of our proposed motion estimator compared to previous motion intensity estimation methods. Subsequently, we provide more specific quantitative comparison results. Finally, we provide additional cases to showcase the generative capabilities of `MotionStone`.

F.1. More Ablations

More Quantitative and Qualitative Results. We first supplement additional quantitative metrics on VBench [6] to demonstrate the superiority of `MotionStone`. As shown in Tab. 2, benefiting from the support of the motion estimator and the decoupled injection method, `MotionStone` outperforms other motion intensity modulation approaches and models without these strategies in terms of generated quality, inter-frame consistency of subjects and backgrounds, motion magnitude, and temporal quality.

Furthermore, we conduct qualitative ablation studies. As shown in Fig. 1, we generate videos using prompts containing both camera and object motions. We observe that `MotionStone` *w/* S and `MotionStone` *w/* SSIM fail to follow the camera motion described in the text prompt. Additionally, `MotionStone` *w/* S exhibits unnatural motion in background objects (e.g., the snow block in the upper left corner of the third column), while `MotionStone` *w/* SSIM displays motion blur issues. These problems are common to non-decoupled motion intensity modulation methods, as they inadvertently cause undesirable background motion while animating the subject. We observe that the `MotionStone` *w/o* M model, which does not utilize the motion estimator, generates static frames without responding to the specified motion intensity. This issue arises because, during training, the model does not receive varying signals corresponding to different motion intensities but rather a constant signal. As a result, the model fails to interpret the provided intensity control signals and is unable to model motion intensity accordingly. `MotionStone` *w/o* D exhibits excessive motion, affecting both the object and the camera motion. Moreover, it fails to follow the text prompt to perform a zoom-out motion, instead generating an opposite camera motion. This issue stems from the lack of decoupled injection of camera and object motion intensity signals. Without clear separation, the model struggles to associate the signals with the specific motion components they are meant to control, leading to unpredictable overlap or confusion. Consequently, the generated video lacks coherent and orderly control. In contrast, `MotionStone` accurately follows the object and camera motion descriptions provided in the text prompt and generates visually appealing and motion-consistent videos

Table 2. **More quantitative ablation results on VBench [6].** The best results for each column are **bold**. Motion Estimator (M), Decoupled injection strategy (D). SSIM and S mean previous motion modeling methods: inter-frame SSIM [3] and feature difference [4] respectively.

Method	Background Consistency	Aesthetic Quality	Imaging Quality	Subject Consistency	Motion Smoothness	Dynamic Degree	Temporal Flickering
MotionStone *w/o* M	95.13%	45.61%	60.15%	93.34%	98.51%	43%	96.51%
MotionStone *w/o* S	94.97%	46.13%	60.73%	92.99%	98.48%	42%	96.42%
MotionStone *w/* SSIM	92.99%	45.72%	54.75%	88.96%	97.51%	47%	93.54%
MotionStone *w/o* D	94.03%	46.27%	58.73%	92.54%	97.59%	48%	95.20%
MotionStone	**95.76%**	**46.78%**	**62.29%**	**94.56%**	**98.96%**	**48%**	**97.41%**

Figure 2. **Illustrations of object and camera motion intensity guidance.** MotionStone can decouple and independently control camera motion and object motion intensities. When either camera motion or object motion is increased, the generated videos exhibit excellent adherence to the respective motion changes.

based on the specified motion intensities. This demonstrates the effectiveness of the proposed modules.

Motion Intensity Guidance. We provide an additional example to demonstrate the decoupled control capabilities of MotionStone for object motion and camera motion intensities. As shown in Fig. 2, in the first two rows, the text prompt does not specify camera motion, so the camera motion intensity is set to the minimum. By increasing the control of object motion intensity, it is evident that the camel moves faster. In contrast, in the last two rows, we introduce camera motion descriptions in the text prompt and adjust the camera motion intensity while reducing the control of object motion intensity. It is observable that as the object motion intensity decreases from 7 to 4, the camel slows down. Meanwhile, as the camera motion intensity increases from 3 to 9, the camera pans to the right more

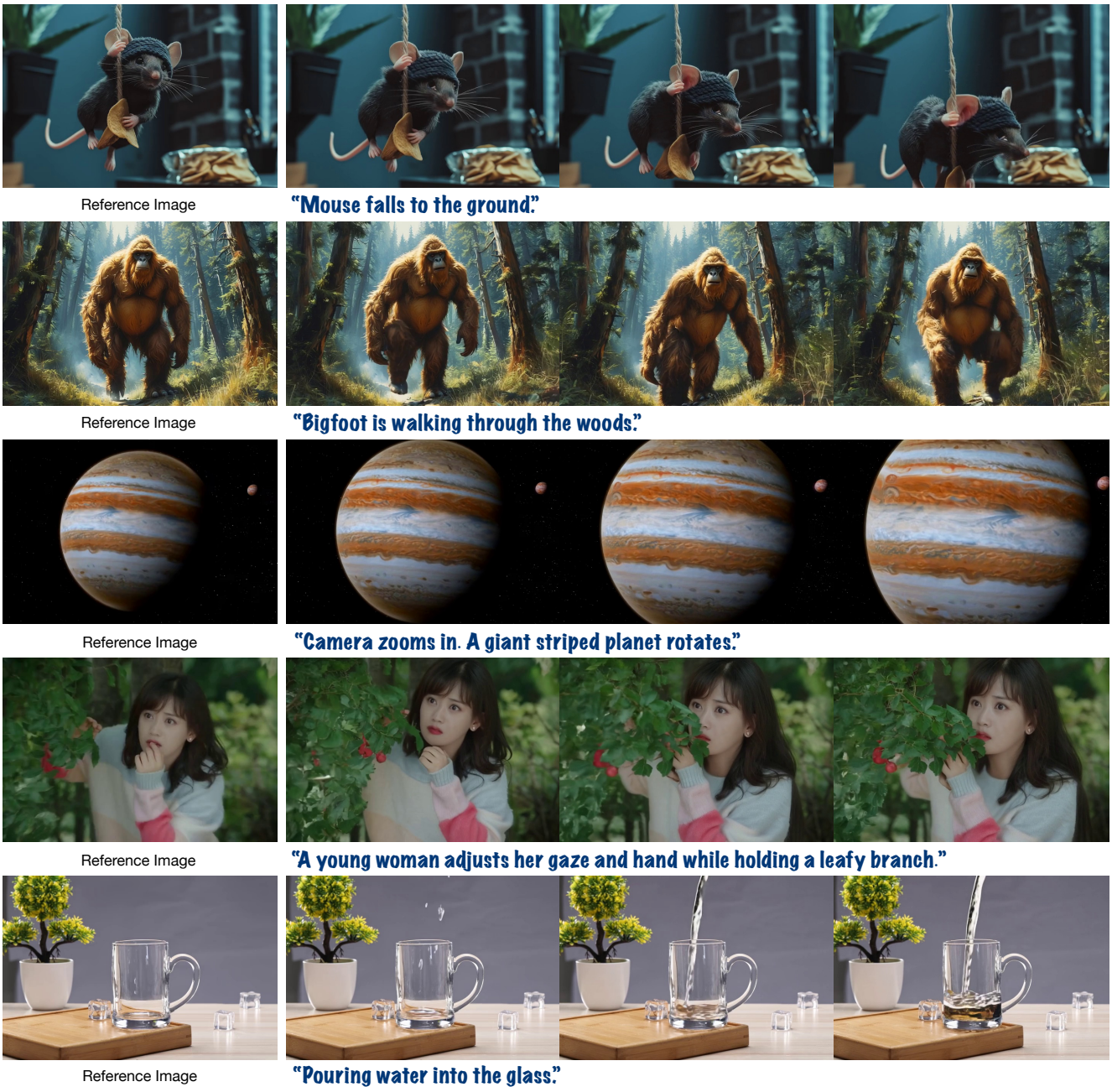

Figure 3. **More cases generated by `MotionStone`.** `MotionStone` demonstrates impressive generation quality across various scenarios. Here, the default object motion intensity or camera motion intensity (if applicable) is set to 5.

Table 3. **Ablation on motion intensity guidance.** Compared to previous methods, our motion estimator achieves more precise control over motion intensity, generating videos with camera or object motion that better aligns with user requirements.

Method	Motion Strength Error
Feature Difference (S) [4]	11.55
SSIM [3]	11.27
Ours	**2.52**

quickly. These examples strongly demonstrate the ability of the `MotionStone` to decouple and independently control camera and object motions in generated videos.

Furthermore, we compare the performance of different motion intensity guidance methods. Using predefined motion intensity values, we generate videos and subsequently apply a motion estimator to obtain the corresponding motion intensities. The mean squared error (MSE) between the generated video intensities and the input values is then cal-

Table 4. **More quantitative comparison results on VBench [6]**. The best results for each column are **bold**.

Method	Background Consistency	Aesthetic Quality	Imaging Quality	Subject Consistency	Motion Smoothness	Dynamic Degree	Temporal Flickering	Camera Motion
I2VGen-XL [13]	90.93%	40.14%	58.35%	86.97%	97.02%	44%	95.24%	18.87%
SVD [1]	93.17%	42.38%	59.61%	93.23%	97.39%	40%	94.70%	22.67%
AnimateAnything [4]	93.89%	46.04%	61.69%	93.72%	97.58%	4%	95.48%	12.19%
CogVideoX-5B [12]	94.91%	45.88%	61.99%	94.39%	98.76%	36%	96.73%	73.26%
`MotionStone`	**95.76%**	**46.78%**	**62.29%**	**94.56%**	**98.96%**	**48%**	**97.41%**	**81.52%**

culated. As shown in Tab. 3, the motion estimator proposed in this work provides more stable motion guidance and ensures that the motion intensities in the generated videos align more closely with the user-specified values.

F.2. More Results

We supplement additional quantitative comparison results across more evaluation dimensions on VBench [6] and VBench++ [7], as shown in Tab. 4. `MotionStone` demonstrates superior performance in terms of temporal quality and motion magnitude of the generated videos compared to previous methods.

We also provide additional examples generated by `MotionStone`, as shown in Fig. 3. These include real human figures, anime-style characters, animals, and natural scenes. `MotionStone` demonstrates remarkable capabilities in conjuring entirely new content out of thin air.

We provide the original video cases showcased in the paper within the supplementary materials. The detailed video effects can be found in the designated folder.

Since there is currently no perfect metric for evaluating a model's ability to disentangle motion intensity, we further propose evaluation metrics that assess motion disentanglement capability from two perspectives: absolute values and linear correlation. As shown in Tab. 5, we propose a new benchmark for motion intensity control using newly designed **MSES** (Motion Strength Error Score) and **M-SRCC** (Motion Spearman Rank-Ordered Correlation Coefficient) metrics. Each prompt combines camera and object motion descriptions to evaluate the model's decoupling of motion intensity control. We construct a benchmark consisting of 30 text prompts and their corresponding images. For evaluating object and camera motion, we randomly sample motion intensities from 1 to 10 for the targeted aspect while selecting from three fixed values (2, 6, and 10) for the other aspect, generating 10 motion intensity combinations per video. MSES computes the normalized ([0,1]) mean squared error between generated and input motion intensity, while M-SRCC quantifies motion consistency via Spearman rank correlation across varying motion inputs. Both independently assess object and camera motion, denoted by subscripts $_o$ and $_c$. The final score is obtained through averaging. As shown in Tab. 5, `MotionStone` outperforms other methods in all dimensions, proving its effectiveness.

Table 5. New designed metrics for motion intensity control.

Metric	Motion Type	SVD	AnimateAnything	`MotionStone`
$MSES_o$	Object	44.18%	45.07%	**83.95%**
$M\text{-}SRCC_o$	Object	60.00%	66.50%	**80.50%**
$MSES_c$	Camera	47.57%	29.41%	**85.64%**
$M\text{-}SRCC_c$	Camera	68.00%	58.00%	**82.00%**
Final Score	All	54.94%	49.74%	**83.02%**

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 6

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[3] Xi Chen, Zhiheng Liu, Mengting Chen, Yutong Feng, Yu Liu, Yujun Shen, and Hengshuang Zhao. Livephoto: Real image animation with text-guided motion control. In *European Conference on Computer Vision*, pages 475–491. Springer, 2025. 2, 4, 5

[4] Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Animateanything: Fine-grained open domain image animation with motion guidance. *arXiv e-prints*, pages arXiv–2311, 2023. 2, 4, 5, 6

[5] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3677–3686, 2020. 3

[6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 1, 3, 4, 6

[7] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024. 6

[8] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 3

[9] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023. 3

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[11] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 3

[12] 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. 6

[13] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 6