Appendix

This appendix includes the following sections:

- Additional Ablation Study (Sec. A):
- Additional Qualitative Results (Sec. C)
- Additional Implementation Details (Sec. D)

A. Additional Ablation Study

A.1. Motion Heatmap

Beyond using optical flow to generate the motion heatmap, we explore alternative approaches leveraging the Segment Anything Model 2 (SAM 2) [32], a state-of-the-art video segmentation model. SAM 2 produces object masks by accepting points or bounding boxes as prompts, generating an initial mask for the first frame, and propagating it consistently across subsequent frames. To target objects with significant motion, we compute optical flow intensity in the first frame and sample K high-intensity points as prompts for SAM 2. To ensure spatial diversity, the frame is divided into an $M \times N$ grid, and the top-K grids with the highest average intensity are selected. From each grid, the point with the highest intensity is chosen as the prompt. SAM 2 then generates masks for the corresponding objects across frames, producing a binary motion heatmap. In our experiments, we set $K = 5$ and use a 5×5 grid.

As shown in the last row of Table A1, the continuous heatmap generated via optical flow achieves better results than the binary SAM-based heatmap especially on text alignment and object motion while SAM-based heatmap improves the overall quality. We observe that although SAM 2 can generate high-quality instance masks across frames with appropriate prompting, it is less effective than optical flow at highlighting large motion regions. We believe that improving prompt sampling strategies could further enhance the performance of SAM 2-based heatmap.

TI2V Score ↑	Image Alignment ↑	Text Alignment ↑	Object Motion ↑	Overall Quality ↑
58.8/41.3	11.4/**12.1**	**34.0**/21.6	**31.2**/22.4	15.0/**20.6**

Table A1. **Ablation study comparing SAM heatmap to optical flow heatmap (MotiF).** Numbers on the left are for MotiF and right for the SAM heatmap.

A.2. Motion Focal Loss Weight

We investigate the impact of the motion focal loss weight λ on TI2V-Bench. As shown in Table A2, simply setting $\lambda = 1$ achieves the best overall results. Generally, reducing λ can improve the overall visual quality of the generated videos but results in lower TI2V Score specially on worse text alignment and object motion.

λ	TI2V Score ↑	Image Alignment ↑	Text Alignment ↑	Object Motion ↑	Overall Quality ↑
0.5	**66.6**/33.4	10.0/**10.8**	**37.3**/13.7	**39.8**/15.3	19.4/**19.7**
2	**63.8**/36.3	**12.4**/7.6	**33.0**/20.4	**31.7**/21.1	**27.5**/16.1
5	**65.6**/34.4	**15.2**/12.8	**39.7**/15.9	**36.8**/16.5	**23.1**/17.8

Table A2. **Ablation studies on the motion focal loss weight λ.** The numbers on the left is for MotiF and the right is for the comparing setting.

B. Additional Quantitative Evaluation

In the paper, we mainly rely on human evaluation for comparisons. Here, we provide additional automatic evaluation results for baseline models on VBench-I2V [20].

B.1. Evaluation on VBench-I2V

VBench-I2V [21] is another popular image-to-video (I2V) benchmark, consisting of 356 real-world images and 1,118 image-prompt pairs. Different from TI2V-Bench, VBench-I2V uses image captions as text conditions instead of action instructions and especially focus on controlling camera motion through text prompts. We evaluate MotiF alongside two strongest baseline models, DynamiCrafter and Cinemo, as well as the static video baseline, on VBench-I2V. Results are presented in Table A3.

MotiF achieves comparable performance to DynamiCrafter and Cinemo in consistency, temporal flickering, motion smoothness, and video quality, while the static video baseline significantly outperforms all models in most metrics except dynamic degree and camera motion. This underscores a key limitation of automatic evaluation: the trade-off between video dynamics and overall quality makes it challenging to provide a holistic assessment of model performance. Additionally, we observe that existing metrics for video dynamics, often based on optical flow, tend to favor videos with significant camera or background motion over object motion. This highlights the importance of conducting human evaluations for the TI2V task to address these shortcomings in automatic evaluation.

C. Additional Visualization

C.1. Comparison with Baseline Models

Figure A1 shows more qualitative results comparing to prior methods. MotiF can generate videos that better align with the input text prompts, which validates the effectiveness of the proposed motion focal loss. We also include the video samples in the supplementary folder.

C.2. Additional Examples

We also provide additional examples that goes beyond TI2V-Bench for complex scenarios including occlusions

Method	Subject Consistency↑	Background Consistency↑	Temporal Flickering ↑	Motion Smoothness↑	Dynamic Degree ↑	Aesthetic Quality↑	Image Quality↑	I2V Subject↑	I2V Background↑	Camera Motion ↑
Baseline: static	100.00	100.0	100.0	99.84	0	65.54	71.61	98.77	97.24	14.29
DynamiCrafter	94.70	97.55	95.17	97.39	39.51	60.40	68.16	96.89	96.68	30.88
Cinemo	96.80	99.04	98.67	98.95	17.32	59.92	64.37	97.43	98.14	15.83
MotiF (ours)	95.27	98.37	97.27	98.16	30.98	58.70	66.95	96.89	97.00	24.35

Table A3. **Results on VBench-I2V**.

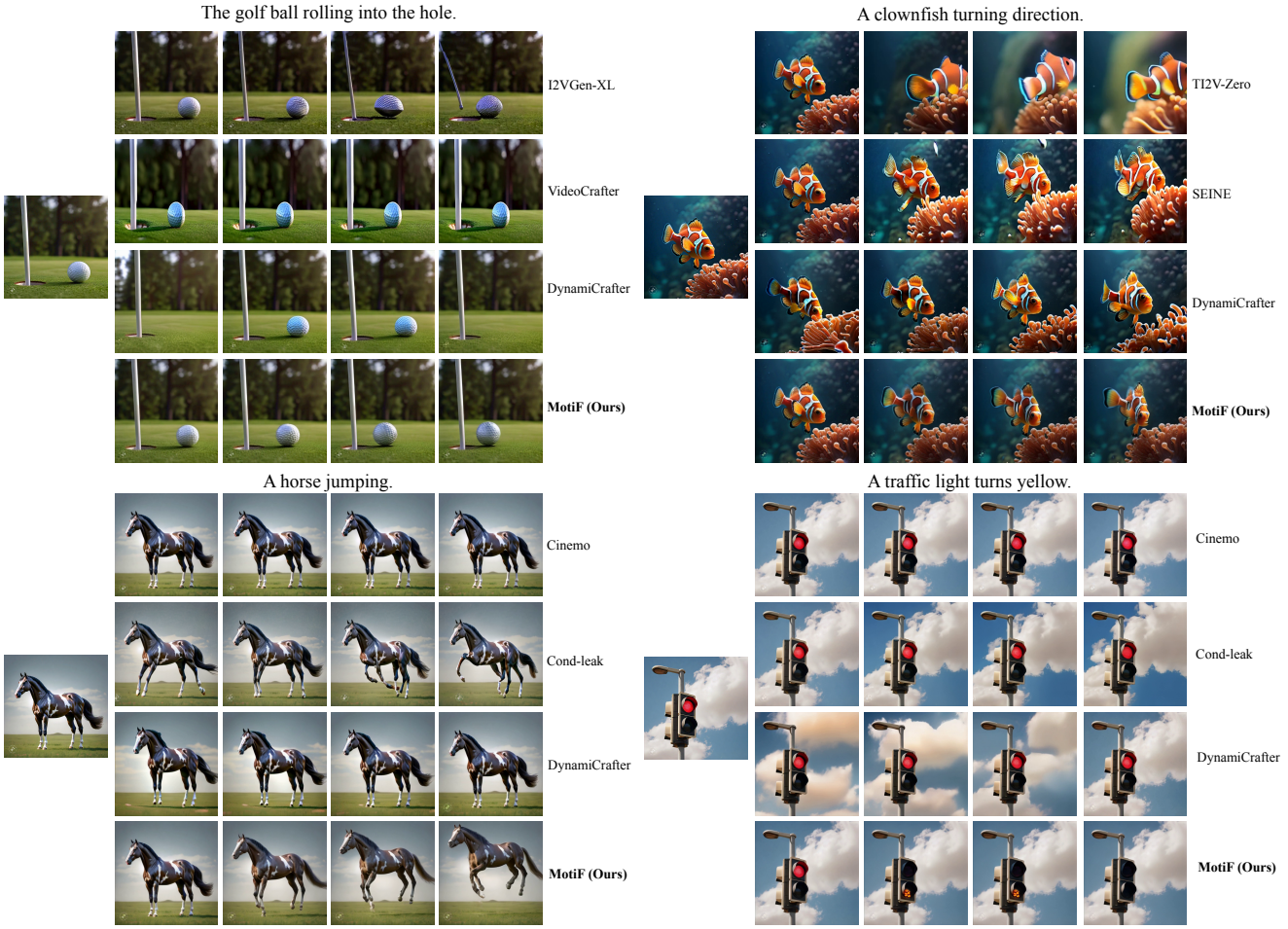

Figure A1. **More qualitative comparison to prior works on TI2V-Bench.** MotiF can generate videos that align better with the text prompts. More video samples are available in the project website.

and multi-object interaction. The results are shown in Figure A2.

C.3. Failing Cases Analysis

Although MotiF shows clear advantages over prior work, it is still far from perfect for TI2V generation especially on our proposed challenging benchmark TI2V-Bench. As shown in Figure A3, we observe two main types of failure cases. First, sometimes the generated motion is not very natural. Second, the generated video may not follow the text prompt. In the second case, there are two challenging scenarios of TI2V-Bench: 1) when the text prompt describes a new object that needs to appear in the scene, the generated video may not be very coherent; 2) when there are multiple objects and the text prompt only refers to one of them, it will be hard for the model to generate precise motion.

We hope MotiF and TI2V-Bench will help the research community to tackle this challenging problem. MotiF is generic and complementary to existing techniques for TI2V generation. We believe that improving the motion heatmap accuracy can potentially boost the performance. Moreover, MotiF is potentially applicable to text-to-video generation.

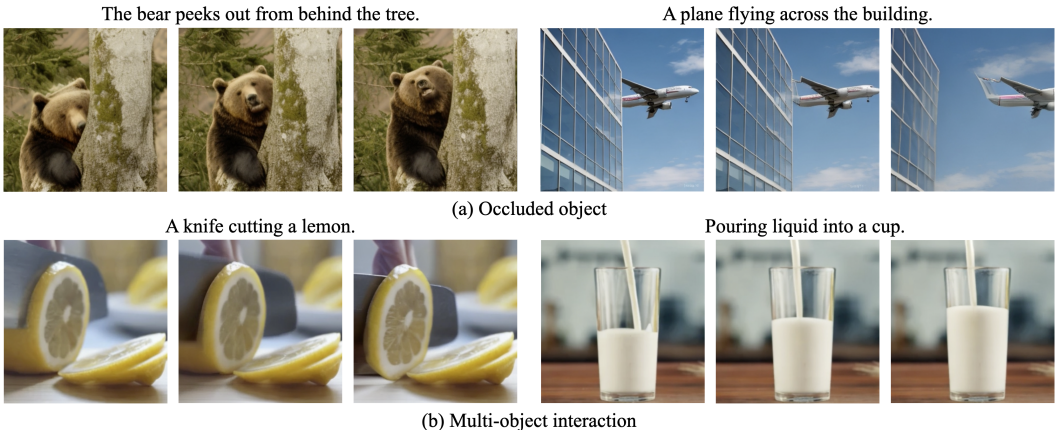

Figure A2. **Results on complex scenarios.** MotiF generates faithful videos for (1) object occlusion and (b) multiple object interaction.

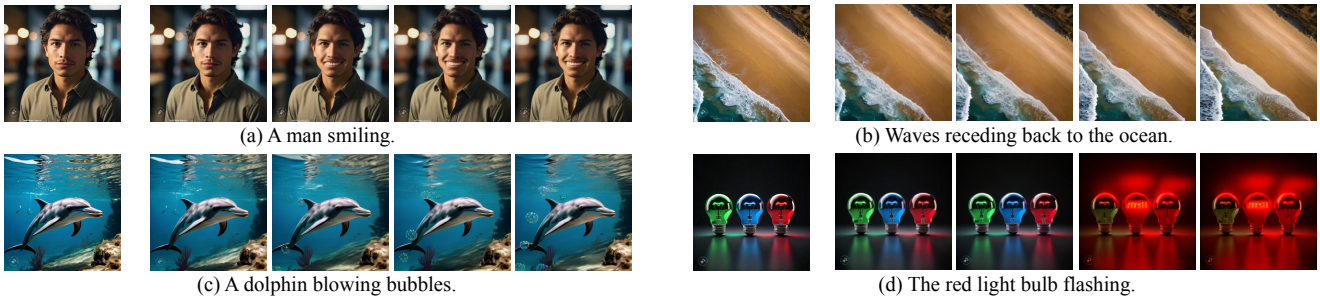

Figure A3. **Typical failure and challenging cases of MotiF on TI2V-Bench.** We observe two typical cases that the model fail: 1) the generated videos may have unnatural motion ((a)); 2) the generated videos do not align well with the prompts ((b), (c), (d)). For 2), there are two specific scenarios when following the text is challenging including novel object ((c)) or multiple objects ((d)). We also include more video samples in the project website.

D. Additional Implementation Details

D.1. Computational Costs of Optical Flow

The average speed of optical flow generation is 2.1 videos per second on a single Nvidia A100 GPU. The speed remains the same for videos of higher resolution, as the model resizes all video frames to a fixed resolution of 960×520.

D.2. Scenes in TI2V-Bench

TI2V-Bench contains a total of 22 diverse scenes, designed to cover a wide range of scenarios. The detailed scenes include: *car on the road, balance scale, (multiple) balloons, bird, (multiple) bulbs, butterfly, candle, child in a playground, dog, rubber duck on a pool, fish, flower, golf ball, horse, animal on a meadow, human face, human body, sun, tide, traffic light, tree, and volcano.*

D.3. Human Evaluation

We show the human evaluation interface in A4. During the evaluation process, the annotators are required to read the instructions first and then answer the two questions based on the specified criteria.

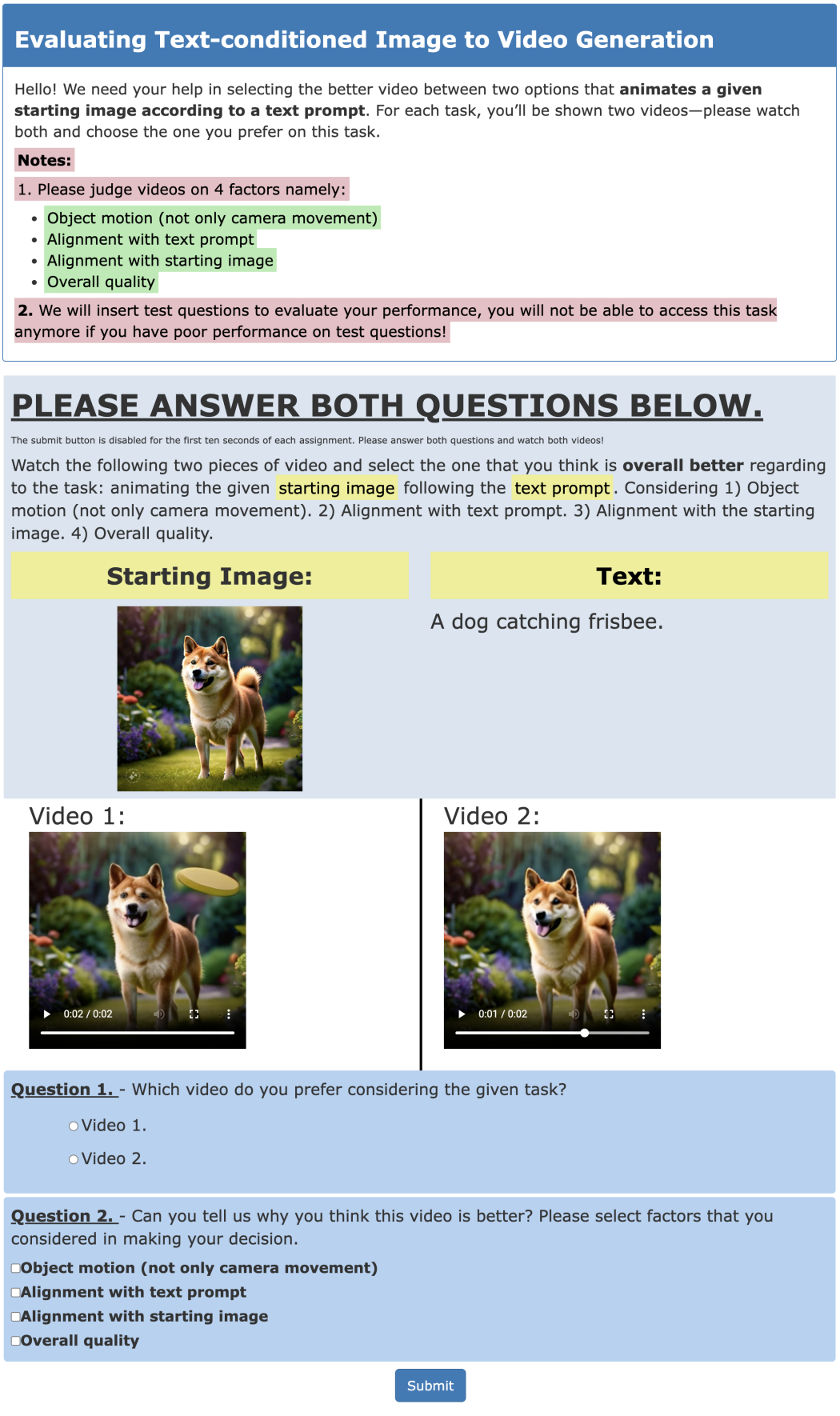

Figure A4. **Illustration of human evaluation interface on the Amazon Mechanical Turk platform.**