Let Your Image Move with Your Motion! – Implicit Multi-Object Multi-Motion Transfer

Supplementary Material

A. More Details

A.1. Full Text Prompts in Fig. 1

For the left-top image containing a single object with the motion <do yoga>, the full text prompt is: "A man is doing yoga."

For the right-top image containing 2 objects with motions <raise hands> and <exercise>, the full text prompt is: "A female cartoon character in armor raised his hands above his head while placing his right foot against his left knee. A shirtless man with a muscular build and a beard is performing a fitness exercise."

For the left-bottom image containing 3 objects with motions <exercise>, <do yoga>, and <raise hands>, the full text prompt is: "The left man is performing a fitness exercise. The middle man is doing yoga. The right man raised his hands above his head while placing his right foot against his left knee."

For the right-bottom image containing 3 objects with motions <do yoga>, <exercise>, and <raise hands>, the full text prompt is: "The left man is doing yoga. The middle man is performing a fitness exercise. The right man raised his hands above his head while placing his right foot against his left knee." Note that the order of motions is different from the above one, showing that our FlexiMMT is able to support flexible, compositional, and arbitrary motion–object rearrangements.

A.2. Details of Baseline Modifications

We evaluate five representative image-to-video (I2V) baselines, including FlexiAct [53], I2VEdit [29], AnyV2V [19], Go-with-the-Flow [4], and CogVideoX-5B-I2V [49]. To ensure a fair comparison and enable multi-object, multi-motion transfer across all methods without altering their core contributions, we introduce minimal adjustments to their implementations. The details are as follows:

AnyV2V [19]. We aggregate information from multiple source videos by averaging their inverted noise and intermediate feature representations to support multi-object multi-motion transfer.

FlexiAct [53]. FlexiAct trains an individual Freq-aware Embedding (Motion Tokens in our paper) for each motion from a single object. To enable multi-object, multi-motion transfer, we simply concatenate multiple motion tokens extracted from different reference videos.

I2VEdit [29]. I2VEdit trains a distinct LoRA module to represent each specific motion. Accordingly, during inference, we enable multi-object, multi-motion transfer by averaging and average the LoRA weights corresponding to different motions.

Go-with-the-Flow [4]. For Go-with-the-Flow, we use its first-frame editing (I2V) functionality. To achieve multi-object multi-motion transfer, we first align object features from the source video's initial frame to their corresponding locations in the target image. We then apply geometric transformations derived from the source masks to the target masks to localize each object, and use the extracted flow fields to guide the motion.

CogVideoX-5B-I2V [49]. CogVideoX-5B-I2V is designed solely for image-to-video generation and does not inherently support motion transfer. Therefore, for this baseline, we simply use the first frame and its corresponding caption to generate the video. In contrast, our FlexiMMT is built upon CogVideoX-5B-I2V and equips it with the capability for multi-object, multi-motion transfer.

A.3. Details of Human Evaluation

As shown in Fig. 8, we use the Gradio[1] toolbox to build a web interface that allows raters to perform annotation. In our human evaluation, we recruited 20 raters, each of whom was assigned 200 video sequences presented in random order. The interface first displays the reference motions, corresponding to the reference videos, followed by 6 video pairs generated by the five baseline models and our FlexiMMT using the same initial frame, caption, and reference motions. For fairness, we shuffled the order of the videos, and each is anonymized as the N-th candidate video. At the bottom of the interface, raters select the best video among the six candidates for each of the four evaluation indicators, including Appearance Consistency (AC), Temporal Consistency (TC), Text Similarity (TS), and Motion Fidelity (MF), through a single-choice scoring module. The definitions of these four indicators are provided in Fig. 9.

B. Additional Experiments

Impact of each mask part in MDMA. Our MDMA framework consists of two major components, *i.e.*, Motion-to-[X] (M2X) mask and Text-to-[X] (T2X) mask. The M2X masks ensure that motion-related tokens focus only on the specified object. This includes the Motion-to-Video and its sym-

[1] Abid, Abubakar, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. "Gradio: Hassle-free sharing and testing of ML models in the wild." arXiv preprint arXiv:1906.02569 (2019).

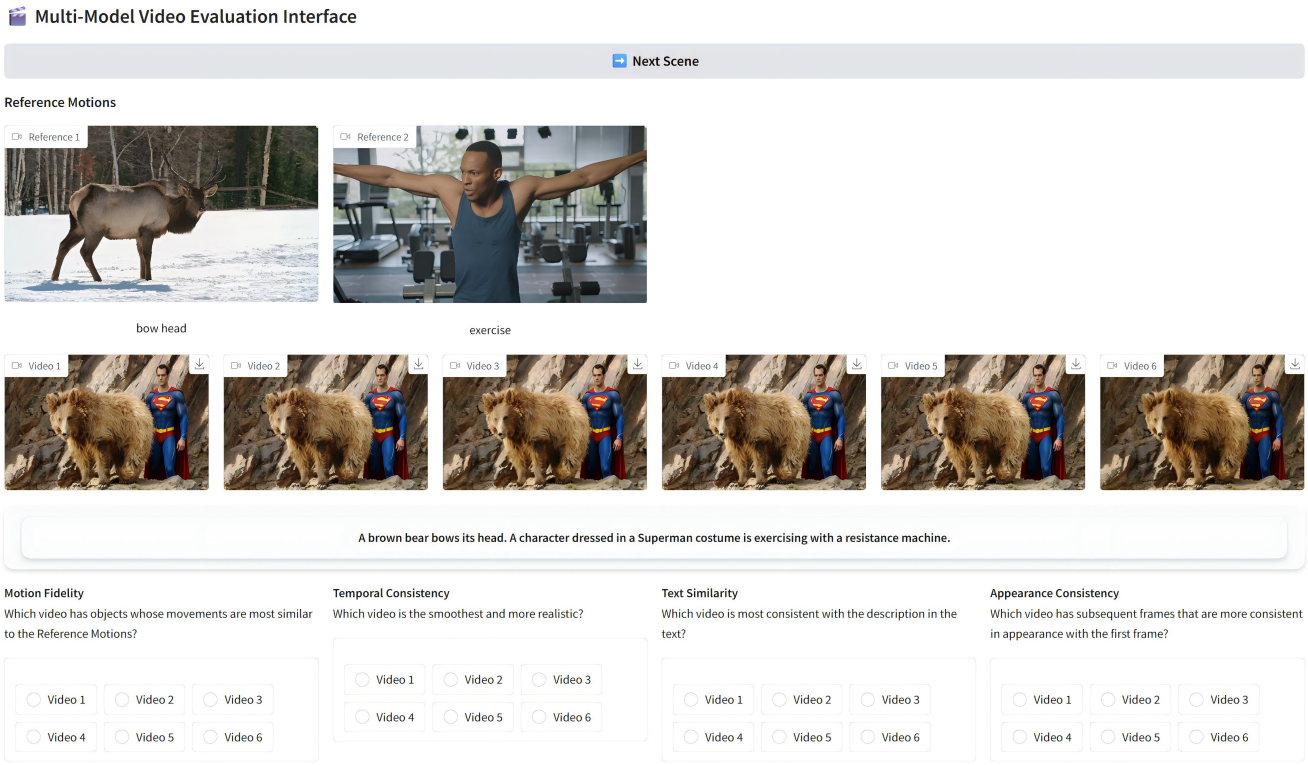

Figure 8. Visulization of human evaluation interface

- **Motion Fidelity (MF)**: Which video has objects whose movements are most similar to the Reference Motions?
- **Temporal Consistency (TC)**: Which video is the smoothest and more realistic?
- **Text Similarity (TS)**: Which video is most consistent with the description in the text?
- **Appearance Consistency (AC)**: Which video has subsequent frames that are more consistent in appearance with the first frame?

Figure 9. Detail of four indicators.

Table 4. Impact of each mask part in MDMA.

Method	AC ↑	TC ↑	TS ↑	TF ↑	FF ↑
w/o M2V	0.907	0.938	0.285	0.383	0.616
w/o M2M	0.902	0.931	0.286	0.576	0.723
w/o T2V	0.900	0.931	0.287	0.572	0.720
w/o T2T	0.903	0.931	0.286	0.573	0.722
w/o T2M	0.928	0.948	0.284	0.476	0.663
Ours	0.904	0.932	0.286	**0.577**	**0.723**

metric Video-to-Motion masks (collectively referred to as M2V), as well as the Motion-to-Motion (M2M) mask. The T2X masks guarantee that motion-related text tokens attend only to the video tokens corresponding to their associated object, thereby preventing cross-object motion interference. This category includes the Text-to-Video and Video-to-Text (T2V), Text-to-Text (T2T), and Text-to-Motion and Motion-to-Text (T2M) masks.

We conduct detailed ablation studies on each component of MDMA, as shown in Tab. 4. Using all mask components yields consistently strong performance across metrics, particularly in Trajectory Fidelity (TF) and Flow Fidelity (FF), demonstrating that every component contributes to accelerating multi-object, multi-motion transfer. There is a slight decrease in Text Similarity (TS) compared with the w/o T2V setting (0.286 vs. 0.287). This is because non-motion tokens in the text description may provide additional cues that enhance text–video alignment, but at the same time introduce motion entanglement that degrades TF and FF. The w/o M2V and w/o T2M settings yield higher Appearance Consistency (AC) and Temporal Consistency (TC); however, they severely impair TF and FF. This occurs because these settings mix motion tokens across dif-

ferent objects, and in some cases even produce static or low-motion videos—leading to artificially high AC and TC but extremely low TF and FF. Given the objective of multi-object, multi-motion transfer, we adopt all mask components as our final FlexiMMT.

Table 5. Anchor frames W in RMPM. Time represents the time required for each step of denoising.

W	AC ↑	TC ↑	TS ↑	TF ↑	FF ↑	Time
1	0.901	0.931	0.288	0.575	0.718	17.39s
3	**0.904**	**0.932**	**0.286**	**0.577**	**0.723**	**17.81s**
5	0.903	0.932	0.287	0.582	0.724	18.60s

Effect of W in RMPM. In RMPM, we maintain a small set of anchor features and corresponding anchor masks to calculate the correlation matrix. We use a local temporal window size W to control the number of anchor frames included. In this subsection, we evaluate the impact of varying the number of anchor frames. The results are shown in Tab. 5. When $W = 1$, our method drops to only using the first frame as the anchor frame, *i.e.*, the anchor set contains only the initial frame. Increasing W to 3, *i.e.*, using the two nearest frames in addition to the first frame, improves performance across all metrics, although training efficiency decreases slightly. Further increasing W yields marginal additional performance gains but leads to higher inference cost. Considering the trade-off between effectiveness and efficiency, we set W to 2 in all experiments by default.

Table 6. Effect of α in Dynamic RMPM.

α	AC ↑	TC ↑	TS ↑	TF ↑	FF ↑	Time
w/o	0.902	0.931	0.286	0.579	0.725	977s
5%	**0.904**	**0.932**	**0.286**	**0.577**	**0.723**	**564s**
10%	0.902	0.930	0.286	0.577	0.722	530s

Effect of α in Dynamic RMPM. In Dynamic RMPM, α is a pre-defined threshold used to determine mask stability. If the difference between the current mask and the mask from the previous step falls below α, further mask updates are terminated, and the most recent stable mask is reused for all subsequent steps.

In Tab. 6, we evaluate the impact of different α values. The results show that $\alpha = 5\%$ yields the best overall performance while maintaining balanced computational efficiency. Thus, we set α to 5 in all experiments by default.

More results. Fig. 11 presents additional qualitative results. The results demonstrate that our method can support flexible, compositional, and arbitrary motion–object rearrangements, effectively handling arbitrary multi-object, multi-motion transfer scenarios.

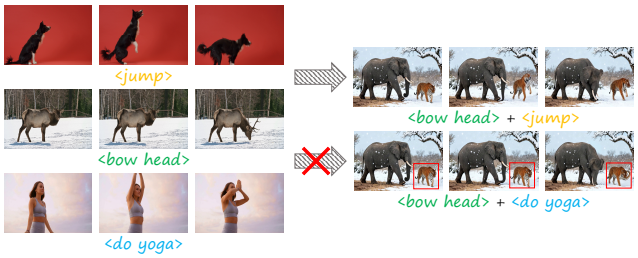

Figure 10. **Limitation.** Visualization of failure cases in our FlexiMMT. The caption for the correctly generated video is: "An elephant bows its head. A tiger jumps up while barking." The caption for the incorrectly generated video is: "An elephant bows its head. A tiger is doing yoga."

Full text prompts in Fig. 11 are:

- <bow head> + <exercise> : "A brown bear bows its head. A character dressed in a Superman costume is exercising with a resistance machine."
- <exercise> + <stand up> : "A man is exercising with a resistance machine. A dog stands up."
- <stand up> + <do yoga> : "A dog stands up. A girl is doing yoga."
- <exercise> + <walk> : "A girl is exercising with a resistance machine. A dog walks on the street."
- <walk> + <do yoga> : "A tiger walks on the snow ground with rocks. A woman wearing a light beige blouse and a delicate necklace is doing yoga."
- <crouch> + <exercise> : "A female cartoon character in armor is exercising with a resistance machine. A shirtless man wearing black shorts is crouching."
- <crouch> + <exercise> + <do yoga> : "The left man is crouching. The middle man is performing a fitness exercise. The right man is doing yoga."
- <exercise> + <do yoga> + <stand up> : "A elderly man with gray hair and a beard is performing a fitness exercise. A man wearing blue suit is doing yoga. A capybara stands up."

C. Limitation

Although our method can transfer multiple motions to images with multiple objects, due to the limitation of the model's capabilities, it is difficult to complete the transfer of motion when the structural differences between objects are too large. As shown in Fig. 10, our FlexiMMT failed to transfer the human motion <do yoga> to the tiger. This deficiency is common in existing models. Therefore, the future goal is to explore better ways to solve the problem of structural adaptability, such as using models with stronger structural compatibility.

Figure 11. **More qualitative comparisons.** The left row shows reference videos. The right row shows generated videos.