

## A Implementation Details

This section provides comprehensive technical details about our MotionRAG framework implementation, covering network architectures, training procedures, and inference pipeline configurations. The code and model weights will be made publicly available to facilitate reproducibility and future research.

**Video and Image Encoders.** We employ VideoMAE-Base [1] pre-trained on Something-Something v2 [2] as our video encoder. We process 16 frames at 224×224 resolution and extract features from all tokens of the final layer. For image encoding, we utilize DINOv2-Large [3], which employs a ViT-L/14 architecture with a hidden dimension of 1024.

**Resamplers.** Our framework employs two separate resamplers for motion and appearance features. These resamplers compress the encoder outputs into a compact set of tokens for efficient processing. The configuration details are provided in Table 1.

Table 1: Configuration for Resamplers.

Configuration	Motion Resampler	Image Resampler
Architecture	Transformer	Transformer
Layers	4	4
Attention heads	12	12
Hidden dimension	768	768
Feed-forward dimension	4096	4096
Output tokens	25	25
Input feature dimension	768 (VideoMAE)	1024 (DINOv2)
Output feature dimension	1024	1024
Dropout rate	0.0	0.0
Trainable parameters	48.0M	48.3M
Initialization	Random	Random

Table 2: Configuration for Motion Context Transformer.

Configuration	Motion Context Transformer
Architecture	Causal Transformer
Layers	4
Attention heads	8
Hidden dimension	1024
Feed-forward dimension	4096
Maximum sequence length	500
Attention mask	Block Causal
Dropout rate	0.0
Position embedding	Sinusoid
Normalization	LayerNorm
Activation	GELU
Trainable parameters	50.4M

**Motion Context Transformer.** Our Context-Aware Motion Adaptation (CAMA) module uses a causal transformer architecture to facilitate in-context learning for motion transfer. The detailed specifications are provided in Table 2.

**Motion Adapters.** We implement separate Motion Adapters for SVD, DynamiCrafter, and CogVideoX-5b, inserting them after text cross-attention layers in the respective UNet architectures for SVD and DynamiCrafter, and all MMDiT layers for CogVideoX-5b. The configuration details for all adapters are provided in Table 3.

Table 3: Configurations for Motion Adapters across different video generation models.

Configuration	SVD Motion Adapter	DC Motion Adapter	CogVideoX Motion Adapter
Architecture	Cross-Attention	Cross-Attention	Cross-Attention
Insertion points	After text cross-attention layers	After text cross-attention layers	After MMDiT self-attention
Number of adapters	16	16	42
Attention heads	8	8	48
Key/Value dimension	1024	1024	3072
Scale factor	1.0	1.0	1.0
Trainable parameters	38M	38M	660M
Initialization	Random	Random	Random

**Training Protocol.** We employ a two-stage training approach for our MotionRAG framework across all three models (SVD, Dynamicrafter, and CogVideoX-5b). In the first stage, we train the Motion Adapter and Resampler modules, followed by training the Motion Context Transformer in the second stage. The training hyperparameters for all configurations are detailed in Table 4.

**Video Retrieval System.** Our text-based retrieval system uses GTE-base-1.5-en [5] to encode text queries and video captions into embedding vectors. For all experiments, we retrieve the top-9 most relevant videos based on cosine similarity between text embeddings.

**Video Generation.** We implement our approach on three state-of-the-art image-to-video generation models: Stable-Video-Diffusion-img2vid (SVD) [6], Dynamicrafter-1024 (DC) [7], and CogVideoX-5b-I2V [8]. The hyperparameters used for video generation are provided in Table 5.

Table 4: Training hyperparameters for the two-stage approach across different models.

Hyperparameter	Stage 1 (SVD)	Stage 1 (DC)	Stage 1 (Cog)	Stage 2 (Transformer)
Dataset	OpenVid-1M [4]	OpenVid-1M [4]	OpenVid-1M [4]	OpenVid-1M [4]
Optimizer	AdamW	AdamW	AdamW	AdamW
Learning rate	$5 \times 10^{-5}$	$5 \times 10^{-5}$	$5 \times 10^{-5}$	$1 \times 10^{-4}$
Resolution	$320 \times 576$	$576 \times 1024$	$480 \times 720$	$224 \times 224$
Batch size	16 (2 per GPU)	16 (2 per GPU)	8 (1 per GPU)	64 (8 per GPU)
Training steps	90K	60K	60K	50K
Loss function	MSE (denoised prediction)	MSE (denoised prediction)	MSE (denoised prediction)	MSE (motion features)
Hardware	8 NVIDIA RTX A6000 GPUs	8 NVIDIA RTX A6000 GPUs	8 NVIDIA RTX A6000 GPUs	8 NVIDIA RTX A6000 GPUs
Training time	48 hours	90 hours	108 hours	9 hours

Table 5: Generation hyperparameters for OpenVid-1K and SkillVid dataset.

Hyperparameter	SVD	DC	CogVideoX
Resolution	$576 \times 1024$	$576 \times 1024$	$480 \times 720$
Frame count	16	16	17
Sampler	EDM	DDIM	DPM
Steps	25	30	25
CFG scale	1.0-3.0	2.0	3
FPS/Motion Strength	7	15	-
Inference time (A6000)	44s	90s	60s

For video preprocessing during training, we extract 16-frame clips at 8 FPS with random temporal crops during training, and for each video, we use the first frame as the reference image.

## B Extended Visualization Results

This section presents additional qualitative results generated by our MotionRAG framework across a diverse range of scenarios.

### B.1 Retrieval Visualization

To illustrate how our retrieval mechanism influences motion generation, Figure 1 shows examples of the retrieval process and resulting generated videos. For each query prompt, our system retrieves semantically relevant videos that contain similar motion patterns, which then guide the generation process.

These examples demonstrate how our approach transfers motion characteristics across visual domains:

**Physics-based motion:** For "metal balls suspended in the air" the system retrieves videos of Newton’s cradles, magnetic balls, and physics experiments. The generated video exhibits realistic pendulum-like oscillations derived from these references.

**Fluid dynamics:** For "a person pouring water into a teacup" retrieved videos show various pouring actions with different teapots and cups. The generated video captures the natural flow of liquid and the subtle hand movements during pouring.

**Human locomotion:** For "a man running on a dirt road" the system retrieves videos of people jogging in various environments. The generated video reproduces natural running gait and body mechanics.

**Animal-human interaction:** For "a person riding on the back of a horse led by another person" retrieved videos show various horse-riding scenarios. The generated video captures the coordinated movement between horse and riders.

Despite differences in background, lighting, and specific object arrangements, the retrieved videos provide valuable motion priors that guide the generation process. The resulting videos exhibit realistic motion while maintaining the visual appearance specified in the input images.

### B.2 Additional Generation Results

Figure 2 showcases video sequences generated using our Dynamicafter+RAG and CogVideoX+RAG models, demonstrating their ability to produce realistic motion patterns across various domains.

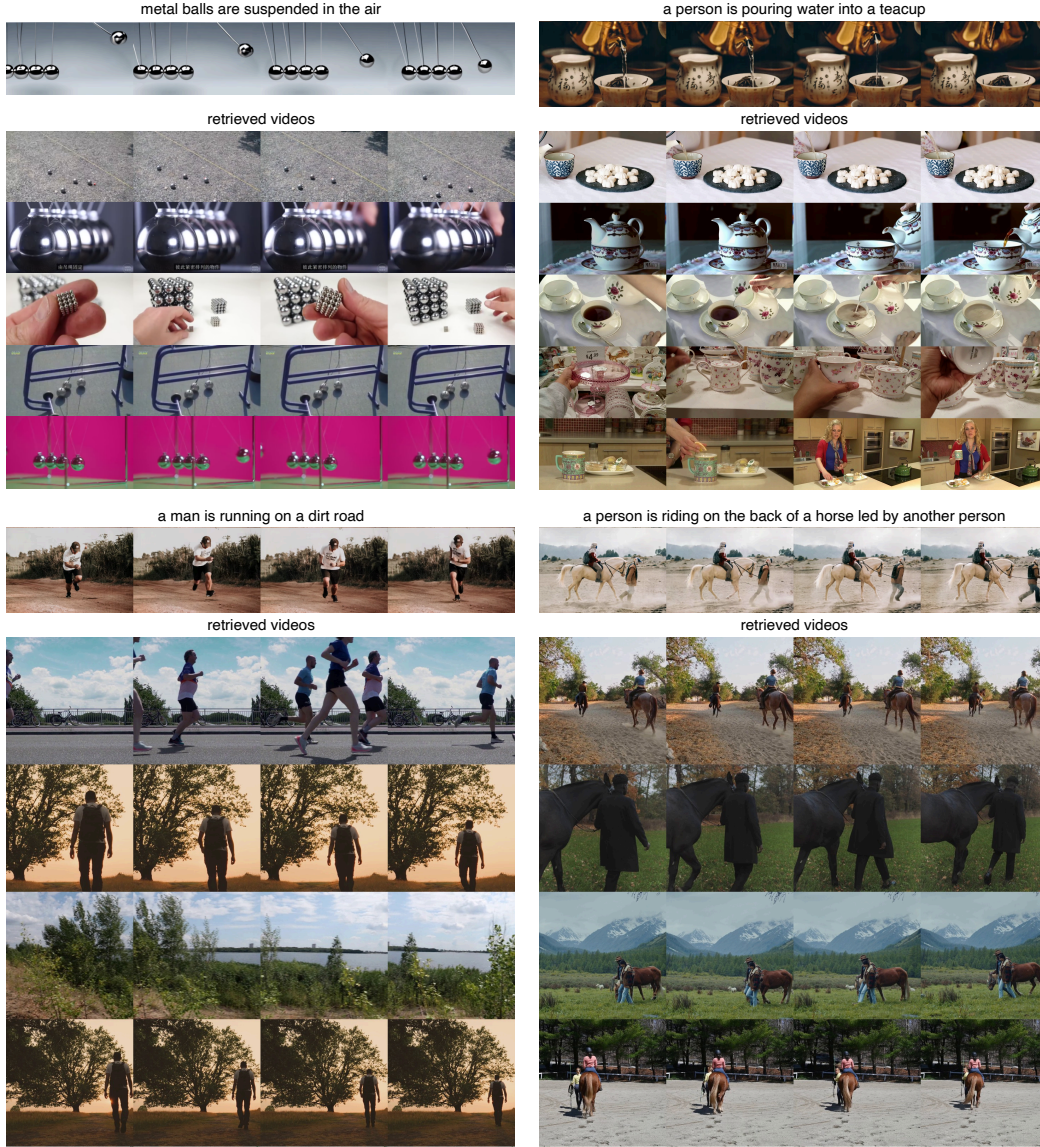
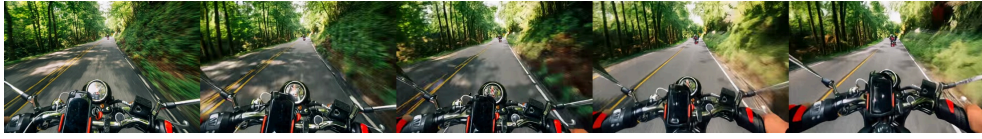


Figure 1: **Retrieval and generation examples.** Each panel shows a different scenario: (top-left) metal balls suspended in air with pendulum-like motion, (top-right) a person pouring water into a teacup, (bottom-left) a man running on a dirt road, and (bottom-right) a person riding on a horse led by another person. For each example, the top row displays frames from our generated video, while the rows below show frames from retrieved reference videos. Note how our system extracts relevant motion patterns from visually different but semantically similar videos.

58 These results highlight our methods' ability to transfer motion patterns across visual domains  
59 while maintaining physical plausibility and semantic consistency. The generated videos preserve  
60 the appearance specifications while introducing temporally coherent motion that aligns with the  
61 described actions. For the best view, please refer to the videos in the supplementary materials.



a motorcycle driving down a road



a bunch of cars are driving on a highway



a penguin walking on a beach near the water



a red double decker bus driving down a street



a city bus driving down a snowy street at night



an older man jogging by the water



a red panda eating bamboo in a zoo



A yellow boat is cruising in front of a bridge



a zebra walking across a dirt road near a field



Figure 2: **Additional video generation results.** Each row displays five frames from a generated video sequence. The first four rows show results from CogVideoX+RAG, while the remaining rows present Dynamicrafter+RAG outputs. Our approach successfully captures motion characteristics across these diverse scenarios.



## References

- [1] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.
- [2] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.
- [3] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [4] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.
- [5] Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*, 2024.
- [6] 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.
- [7] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024.
- [8] 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.