The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

MoFu: Scale-Aware Modulation and Fourier Fusion for Multi-Subject Video Generation

Run Ling[1,2]*, Ke Cao[3]*, Jian Lu[4]*, Ao Ma[1]†‡, Haowei Liu[4], Runze He[5], Changwei Wang[5], Rongtao Xu[5], Yihua Shao[5], Zhanjie Zhang[1], Peng Wu[6], Guibing Guo[2]‡, Wei Feng[1], Zheng Zhang[1], Jingjing Lv[1], Junjie Shen[1], Ching Law[1], Xingwei Wang[2]

[1]JD.com, Inc.
[2]Northeastern University
[3]University of Science and Technology of China
[4]Chongqing University of Post and Telecommunications
[5]University of Chinese Academy of Sciences
[6]Northwestern Polytechnical University
{lingrun.1, maao.8}@jd.com, guogb@swc.neu.edu.cn

Abstract

Multi-subject video generation aims to synthesize videos from textual prompts and multiple reference images, ensuring that each subject preserves natural scale and visual fidelity. However, current methods face two challenges: scale inconsistency, where variations in subject size lead to unnatural generation, and permutation sensitivity, where the order of reference inputs causes subject distortion. In this paper, we propose MoFu, a unified framework that tackles both challenges. For scale inconsistency, we introduce Scale-Aware Modulation (SMO), an LLM-guided module that extracts implicit scale cues from the prompt and modulates features to ensure consistent subject sizes. To address permutation sensitivity, we present a simple yet effective Fourier Fusion strategy that processes the frequency information of reference features via the Fast Fourier Transform to produce a unified representation. Besides, we design a Scale-Permutation Stability Loss to jointly encourage scale-consistent and permutation-invariant generation. To further evaluate these challenges, we establish a dedicated benchmark with controlled variations in subject scale and reference permutation. Extensive experiments demonstrate that MoFu significantly outperforms existing methods in preserving natural scale, subject fidelity, and overall visual quality.

Introduction

Multi-subject video generation aims to synthesize videos from textual prompts and multiple reference images, producing temporally consistent results where each subject preserves visual fidelity, appears at a natural scale, and remains unaffected by permutations of reference images. Despite recent progress, multi-subject video generation faces two challenges. The first is **scale inconsistency**: reference images often vary significantly in subject scales due to different zoom levels, causing subjects in generated videos to appear unnaturally large or small. Another is **permutation sensitivity**: existing methods process images sequentially, inserting them one by one along the frame or channel dimension. This leads to subject distortion, disappearance, or physically implausible interactions that depend on the input order. Moreover, as the number of references increases, computational cost rises sharply and efficiency declines, making sequential processing increasingly impractical. Fig. 1 illustrates these challenges, where existing models fail to maintain natural scales or steadily handle permutations of reference images.

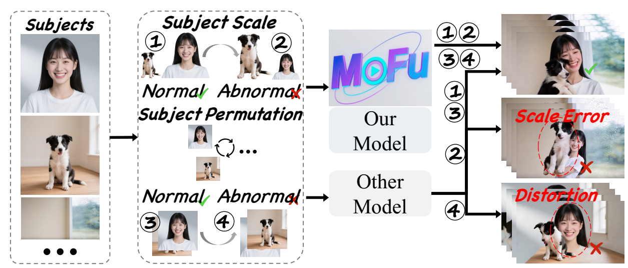

Figure 1: **Illustration of the challenges in multi-subject video generation.** Existing models often suffer from scale inconsistency and permutation sensitivity. Our MoFu framework addresses both issues, producing natural-scale and permutation-invariant videos.

Several recent works (Liu et al. 2025; Fei et al. 2025) attempt to address scale inconsistency by fusing textual prompts into the reference image representation, leveraging spatial relationships implicitly encoded in the prompt (e.g., a girl is in the room with a dog in her arms). However, without an explicit mechanism to interpret and enforce natural scale, these methods produce subjects with unnatural scales that contradict the prompt, especially when the reference images

*These authors contributed equally.
†Project leader.
‡Corresponding author.

themselves vary significantly in scale. To address permutation sensitivity, some methods (Wang et al. 2024; Deng et al. 2025b) concatenate multiple reference images onto a single blank canvas to form a unified visual input. While this avoids sequential processing, it introduces spatial biases (i.e., central or larger subjects may be prioritized) determined by subject placement in the composite image. Consequently, certain subjects may receive disproportionate attention or semantic weight, leading to distortion or disappearance when the spatial arrangement changes.

To address these challenges, we propose MoFu, a unified framework that tackles both scale inconsistency and permutation sensitivity in multi-subject video generation. For the former, we introduce Scale-Aware Modulation (SMO), an LLM-based module that extracts implicit scale relationships from the prompt and injects the condition into the model via modulation, ensuring subjects appear at natural scales. For the latter, to fuse the reference images, we draw inspiration from a theorem in high-dimensional probability: "the inner product between random vectors tends to zero, implying near-orthogonality (Vershynin 2018)". Building on this insight, we propose a simple yet effective Fourier Fusion strategy. Reference images are first segmented and encoded into fine-grained features, which are then transformed into the frequency domain using the Fast Fourier Transform (FFT). Their high- and low-frequency components are aggregated by direct summation, leveraging their near-orthogonality, and reconstructed via Inverse FFT (IFFT) to produce a permutation-invariant representation. To further enforce these objectives, we introduce the Scale-Permutation Stability Loss, which jointly enforces scale consistency and permutation-invariant generation, improving subject fidelity and visual quality.

We propose a high-quality dataset, MoFu-1M, to train our model. Besides, to further evaluate the challenges of multi-subject video generation, we construct a dedicated benchmark, MoFu-Bench, that systematically evaluates models under varying subject scales and reference permutation. Extensive experiments show that MoFu consistently outperforms existing methods by achieving superior scale consistency and permutation-invariant generation, while preserving subject fidelity and overall visual quality.

Overall, our key contributions are summarized as follows:

- We propose MoFu, a unified framework that simultaneously addresses scale inconsistency and permutation sensitivity in multi-subject video generation.
- We design Scale-Aware Modulation, which dynamically adjusts feature representations using scale cues derived from the prompt, ensuring natural subject appearance despite scale variations in the references.
- We introduce a simple yet effective Fourier Fusion strategy that aggregates reference features in the frequency domain, leveraging the near-orthogonality property to produce a permutation-invariant representation.
- We build a high-quality dataset, MoFu-1M, for training, and establish a dedicated benchmark, MoFu-Bench, for assessing scale consistency and permutation-invariance, demonstrating that MoFu significantly outperforms SOTA baselines.

Related Work

Video Generation Models

Recent advances in video generation commonly leverage Variational Autoencoders (VAEs) (Kingma, Welling et al. 2013) to encode videos into compact latent spaces, where large-scale generative pre-training is conducted using diffusion-based (Wan et al. 2025; Kong et al. 2024; HaCohen et al. 2024; Peng et al. 2025), auto-regressive (Teng et al. 2025; Deng et al. 2024; Xie et al. 2025), and other closed-source methods (Kling 2025; Kong et al. 2024; Pika 2025; Vidu 2025; OpenAI 2025; Runway 2025; Gao et al. 2025). These progressions have expanded controllable content generation to tasks such as text-to-video, image-to-video, motion-customized video (Zhang et al. 2025b; Bi et al. 2025; Huang et al. 2025a), and multi-subject video generation (Liu et al. 2025; Fei et al. 2025; Jiang et al. 2025; Deng et al. 2025b; Wang et al. 2024).

Multi-Subject Video Generation

Multi-subject video generation aims to synthesize videos containing multiple entities conditioned on prompts and reference images. Existing approaches tackle subject confusion through text–visual binding (Huang et al. 2025b; Chen et al. 2025; Xiao et al. 2025), LLM-guided layout reasoning (Deng et al. 2025a; Zhang et al. 2025a), or unified frameworks (Jiang et al. 2025). However, two challenges remain: scale inconsistency and permutation sensitivity. **Scale Inconsistency.** Although prior works fuse prompt and reference cues to model spatial relations (Liu et al. 2025; Fei et al. 2025), they lack explicit mechanisms to preserve realistic subject scales, leading to unnatural size variations or contradictions with prompt semantics, especially when reference scales differ greatly. **Permutation Sensitivity.** To reduce sequential bias, some methods concatenate reference images (Deng et al. 2025b; Wang et al. 2024) or apply cross-attention conditioning (Liang et al. 2025), yet both introduce order-dependent artifacts or lose fine-grained identity cues. As a result, subject appearance often becomes distorted or inconsistent when the input order changes.

Method

This section presents the implementation details of MoFu. We first review the foundation of our framework, including the Diffusion Transformer (DiT) backbone and its modulation mechanism. Next, we describe the construction of a high-quality multi-subject dataset, MoFu-1M, and benchmark, MoFu-Bench. Finally, we detail how Scale-Aware Modulation and Fourier Fusion are integrated into the video generation pipeline to explicitly address scale inconsistency and permutation sensitivity, and how the Scale-Permutation Stability Loss enforces these constraints during training.

Preliminary

Diffusion Transformer (DiT) and the Modulation DiT (Peebles and Xie 2023) is a diffusion-based generative model that adopts a vanilla Transformer architecture,

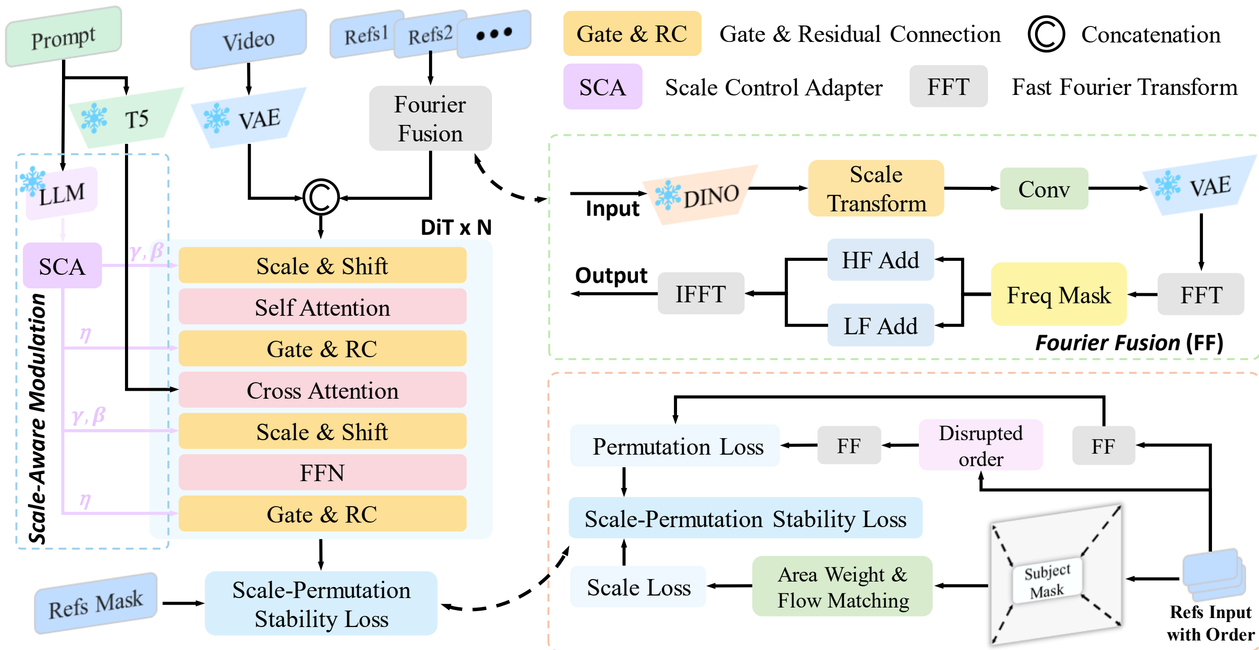

Figure 2: **Overview of the MoFu framework.** MoFu integrates Scale-Aware Modulation (SMO), Fourier Fusion, and the Scale-Permutation Stability Loss (SPSL) into a DiT backbone. SMO extracts scale cues from the prompt via an LLM and adaptively modulates scale features to maintain natural subject scales. Fourier Fusion aggregates reference features in the frequency domain to form a permutation-invariant representation. Furthermore, SPSL jointly enforces scale consistency and permutation-invariant generation.

replacing the U-Net structures commonly used in latent diffusion models. This design enhances the ability to model long-range spatial dependencies and improves scalability. A key strength of DiT lies in its modulation mechanism, which integrates external conditioning signals in a structured manner. Instead of directly concatenating or adding conditioning vectors, DiT applies adaptive scaling and shifting through normalization layers. Given a hidden representation h and a modulation vector m, the modulation is defined as:

$$\mathrm{Mod}(h) = \gamma \cdot \mathrm{LayerNorm}(h) + \beta,$$

where the scaling and shifting parameters γ and β are derived from m via lightweight MLPs. This mechanism enables the model to dynamically adjust internal representations according to conditioning inputs, effectively guiding the generative process with external conditions.

Dataset and Benchmark

To train and evaluate MoFu, we construct a high-quality multi-subject video dataset, **MoFu-1M**, and a dedicated benchmark, **MoFu-Bench**, specifically designed to assess scale consistency and permutation invariance (see Fig. 3).

- **Video Preprocessing and Captioning.** Large-scale raw videos are segmented into coherent clips using scene detection (Castellano and contributors 2025). Low-quality clips are filtered via aesthetic (Schuhmann and contributors 2025) and motion metrics, and the remaining clips are automatically captioned with Qwen2.5-VL (Bai et al. 2025).

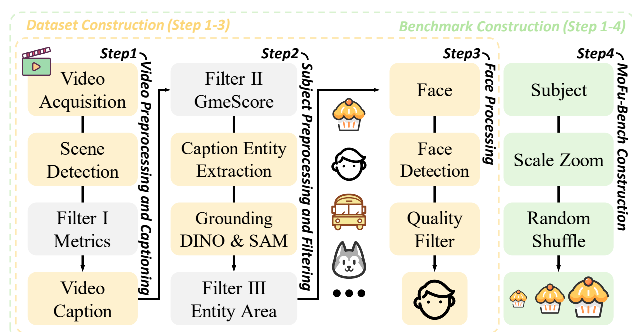

Figure 3: **Construction pipeline of the dataset and benchmark.** We first perform video preprocessing and captioning, and multi-subject extraction and filtering. Multi-scale data for subjects and faces are then generated and refined, resulting in a dataset tailored for evaluating scale consistency and permutation-invariance in multi-subject video generation.

- **Subject Processing and Filtering.** We first filter out videos with poor text–video alignment using GmeScore. Next, we extract entities (e.g., people, animals, objects) from captions using LLM-based parsing (GLM et al. 2024) and localize them in video frames with Grounded-SAM (Ren et al. 2024). Finally, we filter out videos where the extracted subjects are either too large or too small based on their relative area in the frames.

- **Face Processing.** Faces, which require higher fidelity, undergo additional detection and filtering (Deng et al. 2019) to preserve identity clarity, completing the final MoFu-1M dataset.
- **MoFu-Bench Construction.** To rigorously evaluate the model, we build MoFu-Bench by introducing controlled variations to the reference images. Specifically, we apply subject-centric zoom operations to create scale inconsistencies and randomly shuffle the order of references to simulate permutation changes. This benchmark directly tests a model's ability to maintain natural scale relationships and permutation-invariant generation under challenging conditions.

MoFu framework

As illustrated in Fig. 2, our MoFu framework extends a DiT-based backbone with two core modules: Scale-Aware Modulation (SMO) for scale consistency and Fourier Fusion for permutation-invariant conditioning, while the Scale-Permutation Stability Loss (SPSL) jointly enforces these objectives during training.

Scale-Aware Modulation The purpose of SMO is to explicitly capture scale relationships from the textual prompt and inject them into the generation process. We first encode the prompt p using a frozen LLM (Yang et al. 2025) to obtain a semantic embedding:

$$\mathbf{e}_p = \text{LLM}(p) \in \mathbb{R}^d. \tag{1}$$

This embedding is then passed through a lightweight Scale Control Adapter (SCA), implemented as an MLP, to predict modulation parameters:

$$\gamma, \beta, \eta = f_{\text{SCA}}(\mathbf{e}_p), \tag{2}$$

where $\gamma, \beta \in \mathbb{R}^d$ are scale and shift factors applied to normalized features, and $\eta \in \mathbb{R}^d$ is a gating factor to control residual connections.

Given a DiT block feature map $\mathbf{F} \in \mathbb{R}^{B\times L\times d}$, SMO performs adaptive feature modulation:

$$\hat{\mathbf{F}} = \gamma \odot \mathbf{F} + \beta, \quad \mathbf{F}_{out} = \mathbf{F} + \eta \cdot \text{Layer}(\hat{\mathbf{F}}), \tag{3}$$

where $\odot$ denotes element-wise multiplication and Layer represents either the MHA or FFN module. This modulation mechanism explicitly injects scale-aware cues into the features, thereby enforcing natural and consistent subject scales throughout the video generation process.

Fourier Fusion To address permutation sensitivity, we propose the Fourier Fusion strategy, which aggregates multiple reference images without introducing positional or order bias. Given N reference images $\{x_1, x_2, \ldots, x_N\}$, we first apply Grounded-SAM to segment each image and obtain subject masks, from which the subject regions are cropped and resized. Each processed reference is then passed through a 3×3 CNN encoder $\mathcal{E}$ to extract the corresponding feature maps:

$$\mathbf{F}_i = \mathcal{E}(x_i) \in \mathbb{R}^{d\times H\times W}. \tag{4}$$

Each feature map is transformed into the frequency domain using FFT:

$$\mathcal{F}_i = \text{FFT}(\mathbf{F}_i). \tag{5}$$

We then decompose $\mathcal{F}_i$ into high-frequency (HF) and low-frequency (LF) components using a frequency mask M_{freq}:

$$\mathcal{F}_i^{\text{HF}} = M_{\text{freq}} \odot \mathcal{F}_i, \quad \mathcal{F}_i^{\text{LF}} = (1 - M_{\text{freq}}) \odot \mathcal{F}_i, \tag{6}$$

where M_{freq} is a binary mask derived from the radial frequency map that separates HF and LF regions.

Leveraging the near-orthogonality of high-dimensional vectors, we aggregate these components across all references by simple summation:

$$\mathcal{F}^{\text{HF}} = \sum_{i=1}^{N} \mathcal{F}_i^{\text{HF}}, \quad \mathcal{F}^{\text{LF}} = \sum_{i=1}^{N} \mathcal{F}_i^{\text{LF}}. \tag{7}$$

The fused frequency representation is reconstructed via IFFT:

$$\mathbf{F}_{\text{fused}} = \text{IFFT}(\mathcal{F}^{\text{HF}} + \mathcal{F}^{\text{LF}}). \tag{8}$$

This permutation-invariant representation is then concatenated with the video features to condition the generation process.

Scale-Permutation Stability Loss To jointly enforce scale consistency and permutation-invariant generation, we design the Scale-Permutation Stability Loss (SPSL). SPSL consists of two complementary terms: a scale loss, which encourages the model to respect the relative scale of each subject, and a permutation loss, which learns permutation-invariant generation and penalizes inconsistencies caused by changing the order of reference images.

Scale loss. The scale loss leverages reference masks and their relative area ratios to adaptively re-weight the standard flow-matching loss. Given predicted and target noise tensors $\epsilon_{\text{pred}}, \epsilon_{\text{true}} \in \mathbb{R}^{B\times C\times T\times H\times W}$, and their base mean-squared error (MSE) loss $\mathcal{L}_{\text{mse}}$, we compute normalized spatial weights from the reference masks:

$$\mathbf{M} = \sum_{r=1}^{R} w_r \cdot \text{Resize}(m_r), \quad w_r = \frac{\exp(a_r)}{\sum_{r'} \exp(a_{r'})}, \tag{9}$$

where m_r is the mask of the r-th reference image for the sample, a_r is its area ratio and w_r is the normalized weight. The resized masks are combined using the area ratio weights. The final scale loss is:

$$\mathcal{L}_{\text{scale}} = \frac{\sum(\mathcal{L}_{\text{mse}} \odot \mathbf{M})}{\sum \mathbf{M} + \epsilon}, \tag{10}$$

where $\odot$ denotes element-wise multiplication and ϵ is a small constant to avoid division by zero. This weighting scheme places stronger emphasis on accurately reconstructing subjects with larger relative areas while ensuring all frames remain normalized.

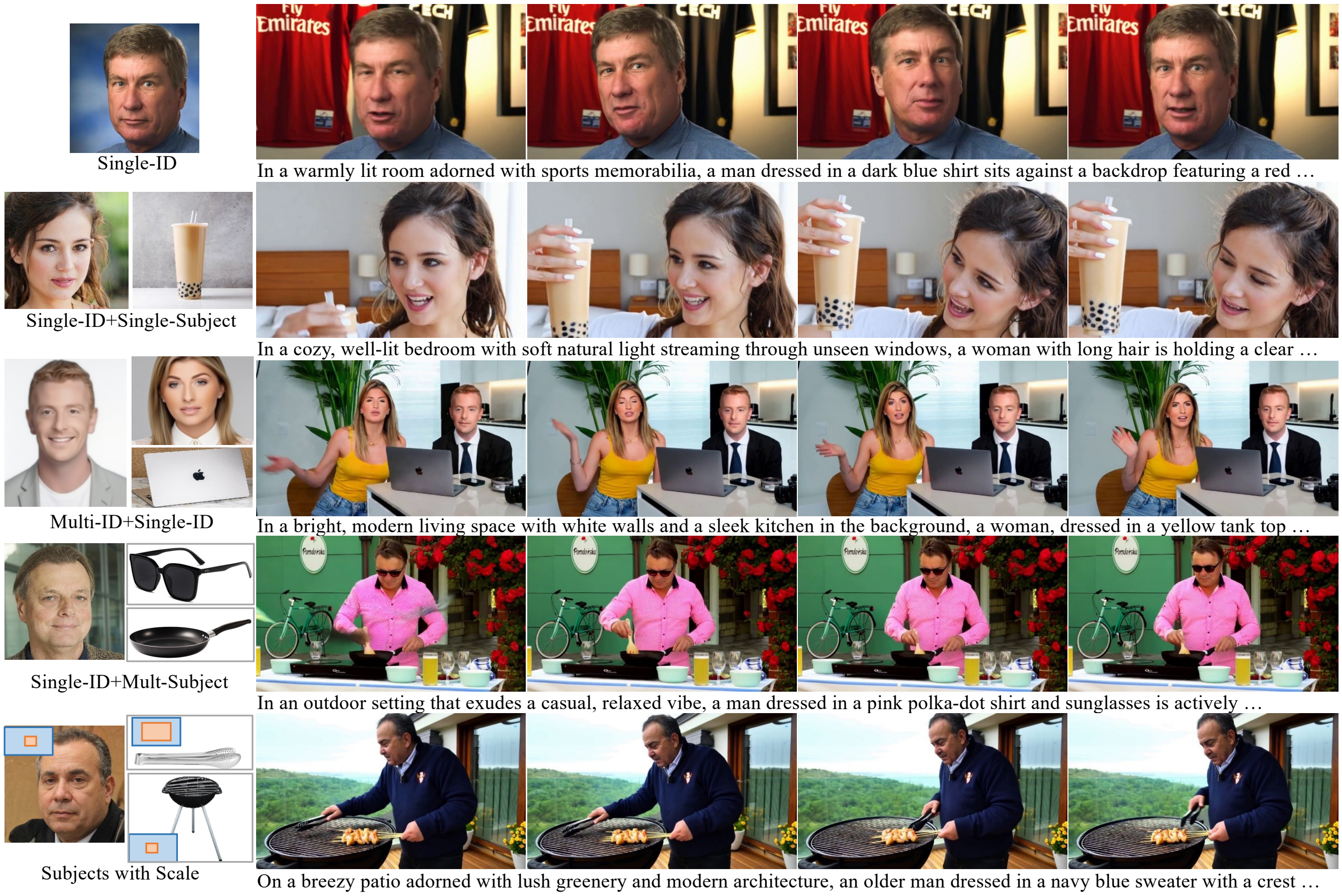

Figure 4: Qualitative evaluation results of our method on different cases. Our model consistently generates video that maintains natural scale and permutation-invariance while accurately following the input text prompt. In row 5, the small rectangles overlaid on the reference images represent the relative area occupied by each subject in the frame. Specifically, the blue regions denote the padded white background, while the red regions indicate the actual subject area.

Permutation loss. To ensure that the change in the permutation of references does not affect the generation, we compute an MSE loss in the frequency domain across P permutations of reference inputs:

$$\mathcal{L}_{\text{prem}} = \frac{1}{P}\sum_{p=1}^{P} \|F(\mathcal{R}_p) - F(\mathcal{R}_{\text{ref}})\|_2^2, \tag{11}$$

where $F(\cdot)$ denotes the Fourier Fusion output, $\mathcal{R}_p$ is a permutation of references.

Final objective. The complete SPSL is a weighted combination of the above terms:

$$\mathcal{L}_{\text{SPSL}} = \mathcal{L}_{\text{scale}} + \mathcal{L}_{\text{perm}}. \tag{12}$$

SPSL provides explicit guidance for MoFu to maintain natural subject scales and robust conditioning regardless of the input reference permutation.

Experiment

Implementation Details

Dataset and Benchmark. We train our model on a high-quality, self-curated video dataset (see Section 3.2). Starting from 15M clips obtained after scene detection, we perform a two-stage filtering process to ensure semantic alignment and appropriate subject visibility. This yields 2.5M high-quality clips, each paired with multiple reference images stored as RGBA. Each clip consists of 81 frames. From these, we select 1M unique clips to form the final training set MoFu-1M. To evaluate scale consistency and permutation invariance, we construct MoFu-Bench, a dedicated benchmark with 1,000 subject-text pairs. Cases are partially adapted from ConsisID (Yuan et al. 2025b), A2-Bench (Fei et al. 2025), and OpenS2V-Bench (Yuan et al. 2025a), with the remainder curated for broad subject diversity. Each case includes up to three reference images with systematic scale variations and randomized permutations, paired with a high-quality prompt for semantic alignment. Compared to existing benchmarks, MoFu-Bench is the first to explicitly assess both scale consistency and permutation invariance, providing diverse subjects and fine-grained annotations as a strong standard for future evaluation.

Evaluation Metrics. To comprehensively evaluate generated videos, upon MoFu-Bench, we adopt six metrics

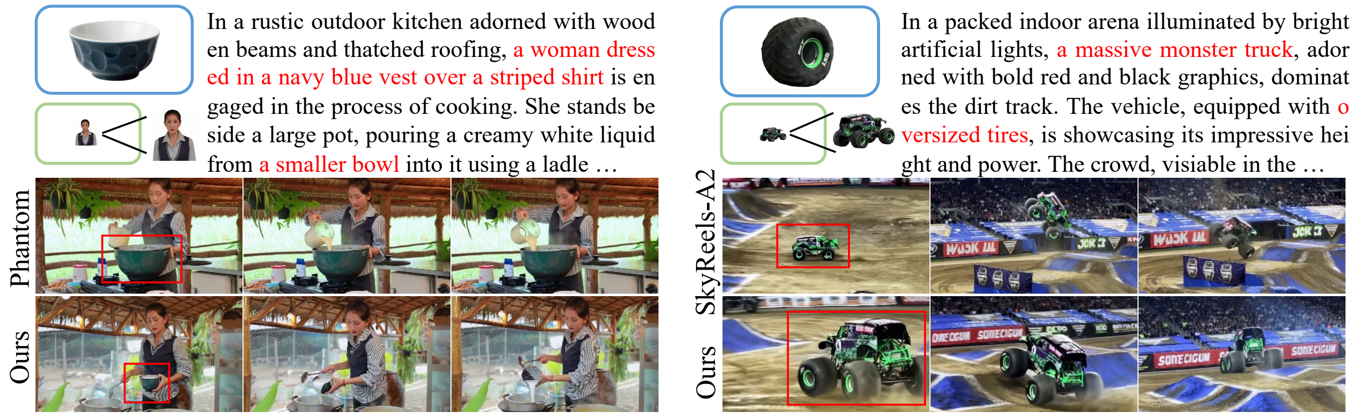

Figure 5: Qualitative comparison on scale-inconsistent scenarios. In the left case, Phantom fails to maintain the correct relative scale of the bowl, while our approach produces a proportionally accurate result. Similarly, in the right case, the large monster truck is generated at a natural size relative to its environment.

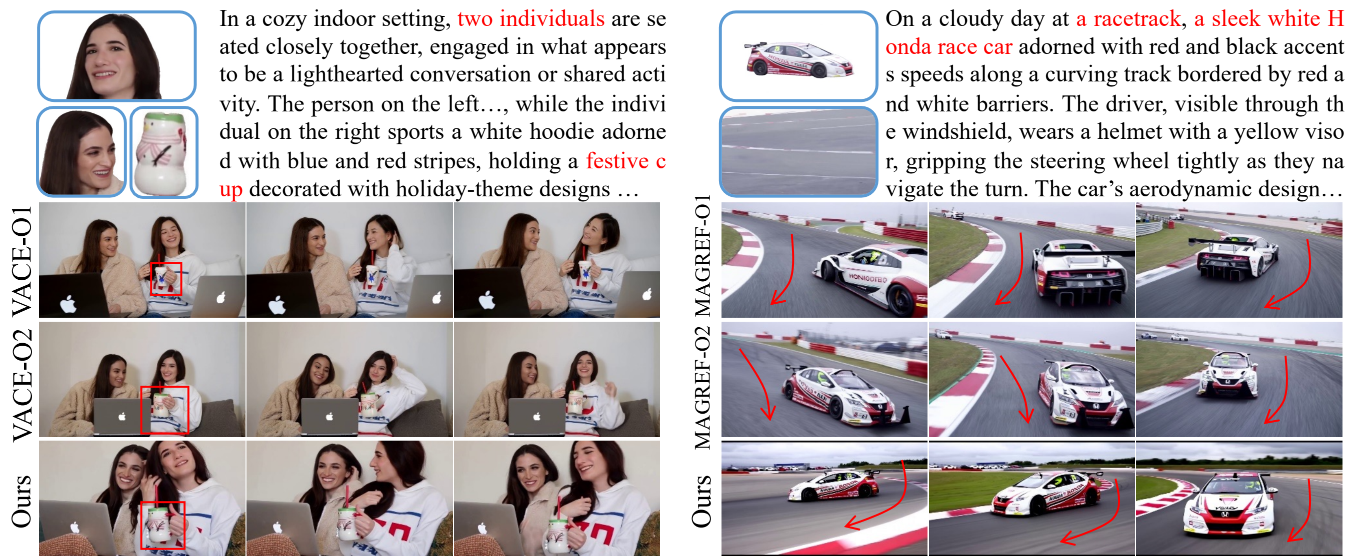

Figure 6: Qualitative comparison on permutation-variant scenarios. In the left case, other methods fail to maintain the correct representation of the cup when the input order changes, while our approach preserves it faithfully. In the right case, changing the input order causes other methods to generate a race car moving backwards, violating physical plausibility, whereas our method consistently produces videos with correct orientation and motion.

covering visual quality, subject fidelity, temporal stability, and scale consistency: Aesthetics (Schuhmann and contributors 2025), FaceSim (Deng et al. 2019), GmeScore, Motion Score, ScaleScore, and SubjectSim. Aesthetics evaluates perceptual quality, FaceSim and SubjectSim measure identity preservation, GmeScore quantifies text-video alignment, Motion Score assesses temporal coherence, and ScaleScore evaluates relative subject scales against prompt descriptions. Detailed definitions and implementations are provided in the Appendix.

Training Strategies. We train our model using the AdamW optimizer, configured with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.01. The learning rate is initialized at 1×10^{-5} and follows a cosine annealing schedule with periodic restarts. Model training is conducted on 16 NVIDIA H800 GPUs for 7 days. Input videos are processed at a resolution of 480P, with a sequence length of 81 frames.

Qualitative Analysis

Fig. 4 presents qualitative results of MoFu, showing its ability to generate high-quality videos that align with prompts while maintaining natural scales and stability across reference permutations. In single- and multi-subject cases (1–4), MoFu preserves subject identity and consistent scales even under challenging settings, while Case 5 demonstrates its robustness to scale discrepancies in reference images. As shown in Fig. 5 and Fig. 6, MoFu surpasses prior

Method	Aesthetics↑	FaceSim↑	GmeScore↑	Motion↔	ScaleScore↑	SubjectSim↑
Phantom (Liu et al. 2025)	0.355	0.375	0.706	0.229	0.536	0.748
SkyReels-A2 (Fei et al. 2025)	0.286	0.341	0.691	0.233	0.527	0.737
VACE (Jiang et al. 2025)	0.392	0.247	0.732	0.214	0.547	0.692
MAGREF (Deng et al. 2025b)	0.369	0.362	0.717	0.207	0.511	0.731
MoFu	**0.401**	**0.396**	**0.745**	0.221	**0.585**	**0.755**

Table 1: Quantitative comparison on MoFu-Bench. The best scores are **bolded**, while the second-best is underlined.

Figure 7: Ablation studies on MoFu. **Left:** Without SMO, the relative scale between the elephant and sparrow becomes unrealistic. With SMO, the natural size relationship is preserved. **Right:** Without Fourier Fusion, the model is sensitive to the order and may drop subjects. With Fourier Fusion, MoFu consistently generates all subjects regardless of reference permutations.

methods in both scale-inconsistent and permutation-variant scenarios, accurately preserving relative object sizes and spatial relations (e.g., bowl–truck, cup–car) with realistic motion patterns. Overall, MoFu achieves prompt-faithful, scale-consistent, and permutation-invariant video generation across diverse multi-subject scenes.

Quantitative Results

As shown in Tab. 1, MoFu achieves state-of-the-art performance across nearly all metrics on MoFu-Bench. It attains the highest Aesthetics and FaceSim scores, producing visually appealing videos with strong identity preservation. Although I2V-based methods such as MAGREF perform well on FaceSim and SubjectSim due to strong reference conditioning, their scale inconsistency limits overall quality. MoFu also leads on GmeScore, reflecting superior text–video alignment, and achieves balanced Motion results by maintaining temporal coherence without sacrificing fidelity. Moreover, it yields substantial gains on ScaleScore and SubjectSim, demonstrating robust scale consistency and subject stability across permutations.

Ablation studies

Effect of SMO and Fourier Fusion. We ablate SMO and Fourier Fusion to assess their contributions as shown in Fig. 7. Without SMO, the model fails to maintain natural relative scales. For example, the sparrow appears disproportionately large compared to the elephant. Incorporating SMO explicitly encodes scale cues from the prompt, preserving realistic proportions. Removing Fourier Fusion makes the model sensitive to reference order, causing missing subjects, like the wooden box in the case, under permutation. Fourier Fusion eliminates this sensitivity, producing consistent and complete outputs regardless of permutations.

Scale-Permutation Stability Loss. SPSL consists of a scale loss and a permutation loss. We do not provide ablations of these two components in Fig. 7 because SMO and FF strategy are supervised respectively by the scale and permutation information provided by this loss. Removing the loss while training SMO or FF independently would render the comparison meaningless, as these modules would lack the necessary supervision signals. Therefore, we omit ablations on the Scale-Permutation Stability Loss.

Conclusion

We proposed MoFu, a unified framework that addresses scale inconsistency and permutation sensitivity in multi-subject video generation. By introducing Scale-Aware Modulation and the Fourier Fusion strategy, MoFu explicitly enforces natural subject scales and permutation-invariant conditioning. Besides, we build MoFu-1M for training, and establish a dedicated benchmark MoFu-Bench for assessing scale consistency and permutation-invariance.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China under Grant No. 62576083.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bi, X.; Lu, J.; Liu, B.; Cun, X.; Zhang, Y.; Li, W.; and Xiao, B. 2025. Customttt: Motion and appearance customized video generation via test-time training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1871–1879.

Castellano, B.; and contributors. 2025. PySceneDetect: Automated Scene Detection for Videos.

Chen, T.-S.; Siarohin, A.; Menapace, W.; Fang, Y.; Lee, K. S.; Skorokhodov, I.; Aberman, K.; Zhu, J.-Y.; Yang, M.-H.; and Tulyakov, S. 2025. Multi-subject open-set personalization in video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6099–6110.

Deng, H.; Pan, T.; Diao, H.; Luo, Z.; Cui, Y.; Lu, H.; Shan, S.; Qi, Y.; and Wang, X. 2024. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Deng, Y.; Guo, X.; Wang, Y.; Fang, J. Z.; Wang, A.; Yuan, S.; Yang, Y.; Liu, B.; Huang, H.; and Ma, C. 2025a. Cinema: Coherent multi-subject video generation via mllm-based guidance. *arXiv preprint arXiv:2503.10391*.

Deng, Y.; Guo, X.; Yin, Y.; Fang, J. Z.; Yang, Y.; Wang, Y.; Yuan, S.; Wang, A.; Liu, B.; Huang, H.; et al. 2025b. MAGREF: Masked Guidance for Any-Reference Video Generation. *arXiv preprint arXiv:2505.23742*.

Fei, Z.; Li, D.; Qiu, D.; Wang, J.; Dou, Y.; Wang, R.; Xu, J.; Fan, M.; Chen, G.; Li, Y.; et al. 2025. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*.

Gao, Y.; Guo, H.; Hoang, T.; Huang, W.; Jiang, L.; Kong, F.; Li, H.; Li, J.; Li, L.; Li, X.; et al. 2025. Seedance 1.0: Exploring the Boundaries of Video Generation Models. *arXiv preprint arXiv:2506.09113*.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

HaCohen, Y.; Chiprut, N.; Brazowski, B.; Shalem, D.; Moshe, D.; Richardson, E.; Levin, E.; Shiran, G.; Zabari, N.; Gordon, O.; et al. 2024. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*.

Huang, C.-P.; Wu, Y.-S.; Chung, H.-K.; Chang, K.-P.; Yang, F.-E.; and Wang, Y.-C. F. 2025a. Videomage: Multi-subject and motion customization of text-to-video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17603–17612.

Huang, Y.; Yuan, Z.; Liu, Q.; Wang, Q.; Wang, X.; Zhang, R.; Wan, P.; Zhang, D.; and Gai, K. 2025b. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*.

Jiang, Z.; Han, Z.; Mao, C.; Zhang, J.; Pan, Y.; and Liu, Y. 2025. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*.

Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes.

Kling. 2025. Image to video elements feature.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.

Liang, F.; Ma, H.; He, Z.; Hou, T.; Hou, J.; Li, K.; Dai, X.; Juefei-Xu, F.; Azadi, S.; Sinha, A.; et al. 2025. Movie Weaver: Tuning-Free Multi-Concept Video Personalization with Anchored Prompts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13146–13156.

Liu, L.; Ma, T.; Li, B.; Chen, Z.; Liu, J.; Li, G.; Zhou, S.; He, Q.; and Wu, X. 2025. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*.

OpenAI. 2025. Sora.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Peng, X.; Zheng, Z.; Shen, C.; Young, T.; Guo, X.; Wang, B.; Xu, H.; Liu, H.; Jiang, M.; Li, W.; et al. 2025. Open-sora 2.0: Training a commercial-level video generation model in 200 k. *arXiv preprint arXiv:2503.09642*.

Pika. 2025. Pika Art 2.0's Scene Ingredients: Redefining Personalized Video Creation.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Runway. 2025. Runway.

Schuhmann, C.; and contributors. 2025. Improved Aesthetic Predictor.

Teng, H.; Jia, H.; Sun, L.; Li, L.; Li, M.; Tang, M.; Han, S.; Zhang, T.; Zhang, W.; Luo, W.; et al. 2025. MAGI-1: Autoregressive Video Generation at Scale. *arXiv preprint arXiv:2505.13211*.

Vershynin, R. 2018. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press.

Vidu. 2025. Reference to video.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Wang, Z.; Li, A.; Zhu, L.; Guo, Y.; Dou, Q.; and Li, Z. 2024. Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962*.

Xiao, G.; Yin, T.; Freeman, W. T.; Durand, F.; and Han, S. 2025. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133(3): 1175–1194.

Xie, D.; Xu, Z.; Hong, Y.; Tan, H.; Liu, D.; Liu, F.; Kaufman, A.; and Zhou, Y. 2025. Progressive autoregressive video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6322–6332.

Yang, A.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Huang, H.; Jiang, J.; Tu, J.; Zhang, J.; Zhou, J.; et al. 2025. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Yuan, S.; He, X.; Deng, Y.; Ye, Y.; Huang, J.; Lin, B.; Luo, J.; and Yuan, L. 2025a. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *arXiv preprint arXiv:2505.20292*.

Yuan, S.; Huang, J.; He, X.; Ge, Y.; Shi, Y.; Chen, L.; Luo, J.; and Yuan, L. 2025b. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12978–12988.

Zhang, H.; Deng, Y.; Yuan, S.; Jin, P.; Cheng, Z.; Zhao, Y.; Liu, C.; and Chen, J. 2025a. MagicComp: Training-free Dual-Phase Refinement for Compositional Video Generation. *arXiv preprint arXiv:2503.14428*.

Zhang, Z.; Liao, J.; Meng, X.; Qin, L.; and Wang, W. 2025b. Tora2: Motion and Appearance Customized Diffusion Transformer for Multi-Entity Video Generation. *arXiv preprint arXiv:2507.05963*.