Grid: Omni Visual Generation

Anonymous Authors[1]

Abstract

Visual generation has witnessed remarkable progress in single-image tasks, yet extending these capabilities to temporal sequences remains challenging. Current approaches either build specialized video models from scratch with enormous computational costs or add separate motion modules to image generators, both requiring learning temporal dynamics anew. We observe that modern image generation models possess underutilized potential in handling structured layouts with implicit temporal understanding. Building on this insight, we introduce GRID, which reformulates temporal sequences as grid layouts, enabling holistic processing of visual sequences while leveraging existing model capabilities. Through a parallel flow-matching training strategy with coarse-to-fine scheduling, our approach achieves up to **67**$\times$ faster inference speeds while using $< \frac{1}{1000}$ of the computational resources compared to specialized models. Extensive experiments demonstrate that GRID not only excels in temporal tasks from Text-to-Video to 3D Editing but also preserves strong performance in image generation, establishing itself as an efficient and versatile **omni-solution** for visual generation.

1. Introduction

Film strips demonstrate an elegant approach in visual arts: by arranging temporal sequences into structured grids, allowing time-based narratives to be displayed in layouts while maintaining their narrative coherence and visual connections. This organization does more than preserve chronological order - it enables efficient content manipulation, comparison, and editing. Drawing inspiration from this intuitive yet powerful organizational principle, we propose a fundamental question: **Can we directly reframe various temporal visual generation tasks as how to layout**, where key visual elements (such as multiple viewpoints or video frames) are treated as grid "layout"?

To answer this, a natural starting point emerges from the recent breakthroughs in text-to-image generation. For single image generation, models like (Esser et al., 2024; Baldridge et al., 2024; Betker et al., 2023) have demonstrated remarkable capabilities in understanding and generating complex spatial relationships. For temporal visual generation, current approaches typically follow two distinct paths: (a) building specialized video models from scratch (e.g., Sora), which requires learning both spatial and temporal relationships with prohibitive computational costs, (b) treating image generators as single-frame producers and mainly train additional motion modules - while this avoids learning spatial generation from scratch, it still requires learning temporal dynamics entirely anew.

Guided by our layout-centric perspective, we argue that the inherent capabilities of image generation models are significantly underestimated. Modern image models already possess implicit understanding of both spatial relationships and basic temporal coherence, suggesting we might not need to learn either aspect entirely from scratch. To validate this hypothesis, we first test the ability of current image generation models to handle grid-arranged layouts through simple prompting (Figure 8). Our experiments reveal that while these models show promising initial capabilities in understanding structured layouts, they still fall short in two fundamental aspects (detailed in Section A.1):

- **Layout Control:** They fail to maintain both consistent grid structures and visual appearances across layouts.
- **Motion Coherence:** When given specific motion instructions (*e.g.*, "rotate clockwise"), they cannot reliably create sequential movements across layouts.

To address these, we introduce GRID, which **reformulates temporal sequences as grid layouts**, allowing image generation models to process the entire sequence holistically and learn both spatial relationships and motion patterns.

Building on this grid-based framework, we develop a **parallel flow-matching** training strategy that leverages large-scale web datasets, where video frames are arranged in grid

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

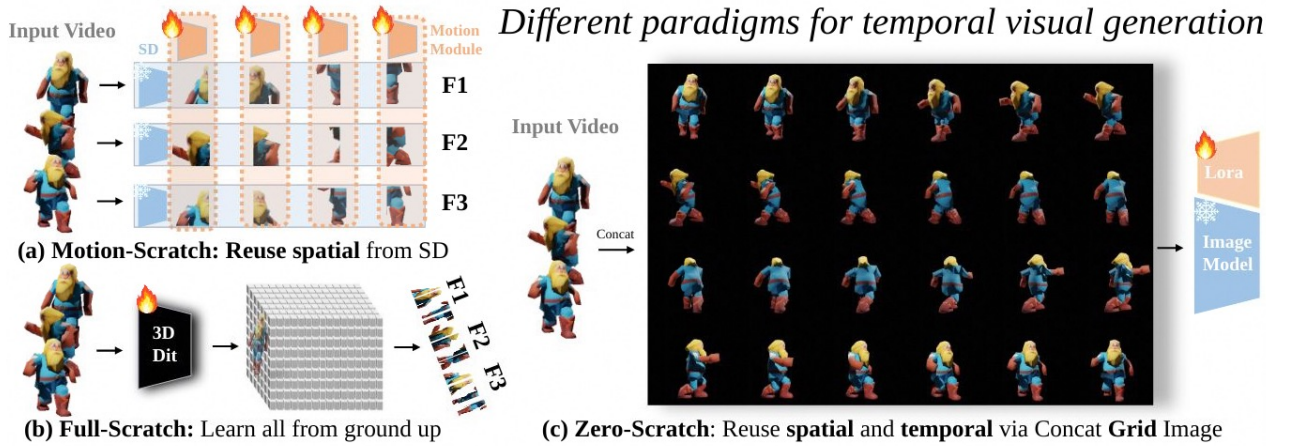

Figure 1: **Different paradigms for temporal visual generation**. (a) Motion-Scratch (e.g., SVD, AnimateDiff): learn temporal dynamics from scratch while reusing pretrained image models. (b) Full-Scratch (e.g., Sora): learn everything from scratch, requiring massive data and computational resources. (c) Zero-Scratch (**GRID**): reuse both spatial and temporal capabilities through grid-based reformulation, leveraging pretrained models' inherent understanding.

layouts. The model learns to simultaneously generate all frames in these structured layouts through a base parallel matching loss, achieving consistent visual appearances and proper grid arrangements. This approach naturally utilizes the models' self-attention mechanisms to capture and maintain spatial relationships across the entire layout.

For precise motion control, we further incorporate dedicated temporal loss and motion-annotated datasets during fine-tuning. The temporal loss ensures smooth transitions between adjacent frames, while the motion annotations help learn specific patterns like "rotate clockwise". These components are balanced through a coarse-to-fine training schedule to achieve both fluid motion and consistent spatial structure.

Through our carefully designed training paradigm, GRID achieves remarkable efficiency gains, demonstrating a substantial **6-35**$\times$ acceleration in inference speed compared to specialized expert models, while requiring merely $\frac{1}{1000}$ of the training computational resources. Our framework exhibits exceptional versatility, achieving competitive or superior performance across a diverse spectrum of generation tasks, including Text-to-Video, Image-to-Video, and Multi-view generations, with performance improvements of up to **23**%. Furthermore, we extend the capabilities of GRID to encompass Video Style Transfer, Video Restoration, and 3D Editing tasks, while preserving its original strong image generation capabilities for image tasks such as image editing and style transfer. This unique combination of expanded capabilities and preserved foundational strengths establishes GRID as a **omni-solution** for visual generation.

Our main contributions are summarized as follows:

- **Novel Grid-based Framework:** We introduce a new paradigm that reformulates temporal sequences as grid layouts, enabling holistic processing of visual sequences through image generation models.
- **Coarse-to-fine Training Strategy:** We develop a parallel flow-matching strategy combining layout matching and temporal coherence losses, with a coarse-to-fine training schedule that evolves from basic layouts to more precise motion control.
- **Omni Generation:** We demonstrate strong performance across multiple visual generation tasks while maintaining low computational costs. Our method achieves results comparable to task-specific approaches, despite using a single, efficient framework.

2. Layout Generation

Inspired by film strips that organize temporal sequences into structured grids, we present GRID, a grid layout-driven framework that reformulates multiple visual generation tasks through grid-based representation. Our GRID consists of three key components: 1) **Grid Representation**, which enables layout-based video organization for comprehensive visual generation; 2) **Parallel Flow Matching**, which ensures temporal coherence in successive grids; and 3) **Coarse-to-fine Training**, which enhances motion control capabilities. The framework architecture is illustrated in Figure 2 (left).

2.1. Grid Representation

Existing text-to-image models, with inherent attention mechanisms, enable image manipulation and editing by generat-

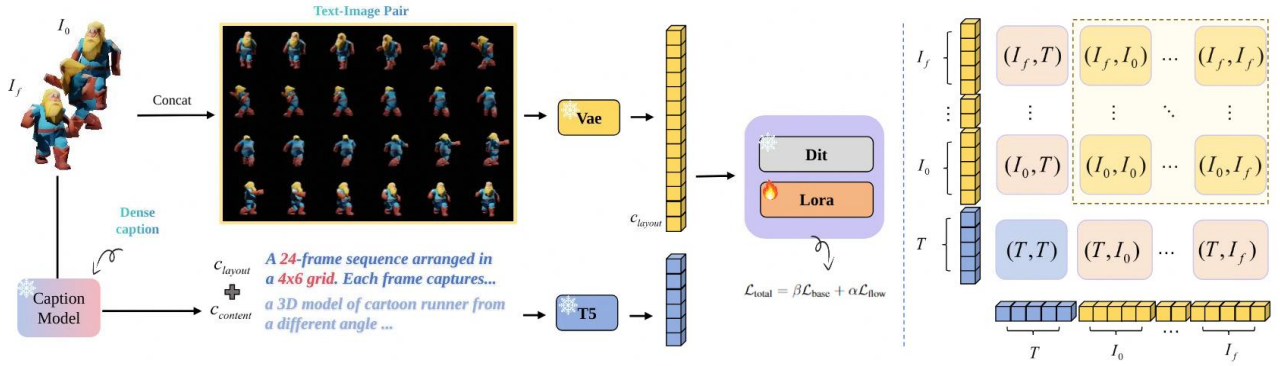

Figure 2: **Pipeline Overview**. Left: GRID arranges videos into grid layouts, with text annotations combining layout format prefix and LLM-generated captions. The model is trained using LoRA fine-tuning on DIT blocks, incorporating both base loss and temporal loss to capture inter-frame relationships. Right: Grid-based reformulation naturally extends model's built-in self-attention to include frame-wise self-attention, cross-frame attention, and text-to-frames cross-attention.

ing new content from partial image information and semantic instructions, which inspires us to extend this capability to temporal generation by introducing a novel input paradigm, termed Grid Representation, that generates temporal content from keyframe visuals and semantic instructions.

Consider a general visual generation task that transforms an input condition $c_{content}$ (such as a text description T) into a sequence of images $(I_0, ..., I_f)$. We propose a grid layout specification c_{layout} that arranges temporal frames into a structured grid within a single image, where each cell (i, j) contains a specific image I_{ij}. As shown in Figure 2 (right), when this grid structure is input into a conventional text-to-image model, the model's inherent attention mechanisms naturally extend their functionality to process this spatial arrangement as:

- **Self-attention Expansion:** The standard self-attention mechanism (I, I) (yellow block) expands into two distinct components:
  - Intra-frame attention (I_i, I_i): Maintains feature learning within individual grid cells
  - Cross-frame attention (I_i, I_j): Enables temporal relationships between different grid cells
- **Cross-attention Extension:** The text-image cross-attention (I, T) (pink block) extends naturally to provide uniform text conditioning across all frame positions

Our approach demonstrates that thoughtful problem restructuring can be more effective than architectural modifications. By reorganizing the input space into a grid representation, standard text-to-image models can naturally handle temporal generation without architectural changes (see Appendix A.2 for detailed attention mechanism analysis). This grid-based design offers two key advantages: First, it enables parallel generation of all frames and eliminates the error accumulation problems common in autoregressive approaches (Tian et al., 2024). Second, by leveraging the inherent consistency priors within pretrained image generation models, our approach effectively transfers their learned spatial consistency to temporal and multi-view coherence. This crucial advantage avoids the need for extensive pretraining on massive video datasets, as the grid representation naturally extends existing image-level understanding to sequence generation. Additionally, through flexible layout conditioning (c_{layout}), our model shows strong generalization capabilities beyond training constraints (Section A.5), suggesting a promising solution to the fixed-length limitations of existing methods. Additionally, our grid representation supports diverse input types, including multi-view images and multi-frame sequences, laying the foundation for a comprehensive omni-generation model that bridges image and video domains.

2.2. Parallel Flow Matching

To fully leverage the potential of our grid representation, we employ parallel flow matching (Esser et al., 2024) to ensure temporal coherence across consecutive grids. For each training sample $\mathbf{I} = (I_{ij})$, we generate a corresponding text representation by integrating layout specifications with content descriptions: $c' = [c_{\text{layout}}, c_{\text{content}}]$. Here, c_{layout} encodes the spatial structure (e.g., a sequence arranged in $m \times n$ grids), while c_{content} captures the visual content as well as the temporal relationships between frames.

Parallel Flow Evolution with Global Awareness. Our grid representation integrates seamlessly with flow matching by organizing temporal frames into a unified grid image $\mathbf{I}$. This enables parallel evolution of frames through the following

process:

$$\mathbf{I}_t = (1-t)\mathbf{I} + t\epsilon, \quad t \sim \mathcal{U}(0,1), \quad \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

Unlike autoregressive approaches that generate frames sequentially, our formulation allows all frames to evolve simultaneously from noise to target distribution through the model's native prediction process:

$$f : (\mathbf{I}_t, t, c') \rightarrow \epsilon - \mathbf{I} \tag{2}$$

Each frame $(I_{ij})_t$ interacts with others within the grid, enabling mutual influence. This interaction naturally enforces temporal consistency across all sequences.

2.3. Coarse-to-Fine Training

Training models for temporal understanding in grid representation demands extensive video data to achieve key capabilities like identity preservation and motion consistency - essential features for video and multi-view generation that text-to-image models typically lack. This training process faces two main challenges from mixed quality of available data: the abundance of low-quality internet videos, and high computational costs of processing high-resolution footage. We tackle these limitations through a coarse-to-fine training strategy that combines two key components: data curriculum and loss dynamic. This dual approach optimizes both training efficiency and model performance, enabling effective use of diverse data sources while minimizing computational overhead. Our strategy enhances the capabilities of our flow-based framework without sacrificing training efficiency.

Data Curriculum. Our training strategy follows a Coarse-to-Fine approach, starting with foundational learning and advancing to refinement:

- *Coarse Phase:* In the initial phase, we utilize large-scale Internet datasets, including WebVid, TikTok, and Objaverse, which are designed with uniform c_{layout} specifications. Although the content descriptions (c_{content}) are automatically generated by GLM-4V-9B (Du et al., 2022) and may lack precise control details, the vast scale and diversity of this data—albeit at lower resolutions—provide a strong basis for developing robust spatial understanding and basic layout structures.

- *Fine Phase:* Building on the foundational knowledge from the coarse phase, we transition to training with carefully curated, high-resolution samples. These samples are paired with detailed descriptions generated by GPT-4 (OpenAI, 2023), offering explicit spatial and temporal instructions. As shown in Figure 2, these high-quality captions facilitate fine-grained control over complex layout variations, enabling the model to handle intricate spatial and temporal dynamics effectively.

Loss Formulation. Our training objective combines appearance accuracy with temporal consistency through a weighted sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \alpha \mathcal{L}_{\text{flow}} \tag{3}$$

The base loss ensures accurate noise prediction at each position using mean squared error:

$$\mathcal{L}_{\text{base}} = \mathbb{E}_{t,\epsilon}[|\epsilon - \epsilon_\theta(\mathbf{I}, t, c')|^2] \tag{4}$$

The flow loss enforces smooth temporal transitions by penalizing inconsistent changes between adjacent positions. For any position (i,j) in the grid, directional changes are:

$$\Delta\epsilon^{ij} = \begin{cases} \epsilon^{ij} - \epsilon^{i,j-1} & \text{within row} \\ \epsilon^{i,0} - \epsilon^{i-1,n} & \text{across rows} \end{cases} \tag{5}$$

Similarly for predicted values:

$$\Delta\epsilon_\theta^{ij} = \begin{cases} \epsilon_\theta^{ij} - \epsilon_\theta^{i,j-1} & \text{within row} \\ \epsilon_\theta^{i,0} - \epsilon_\theta^{i-1,n} & \text{across rows} \end{cases} \tag{6}$$

The flow loss then minimizes inconsistencies in these directional changes:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,\epsilon}[|\Delta\epsilon - \Delta\epsilon_\theta(\mathbf{I}, t, c')|^2] \tag{7}$$

The weight α gradually increases from 0 to a preset upper bound, allowing the model to first establish precise content generation capabilities before focusing on temporal dynamics. This staged evolution of the loss function complements our data curriculum, enabling the model to effectively learn both the spatial and temporal aspects of generation in a coordinated manner.

2.4. Omni Inference

We propose an omni-inference framework designed to handle a wide range of generation tasks using a reference-guided grid layout initialization. The core idea of our approach is to unify different generation tasks by employing a well-structured initialization process combined with controlled grid noise injection. At the same time, we ensure consistency with the reference through the use of a binary mask.

Given a reference image I_{ref} or key frames $(I_0, ..., I_{m-1})$, we construct a grid structure $\mathbf{I} = (I_{ij})_{m \times n}$. For single-image expansion and frame interpolation tasks, we initialize the grid as:

$$I_{ij} = \begin{cases} I_{\text{ref}} & \text{expansion} \\ (1 - \frac{j}{n}) I_{i,0} + \frac{j}{n} I_{i+1,0} & \text{interpolation} \end{cases} \tag{8}$$

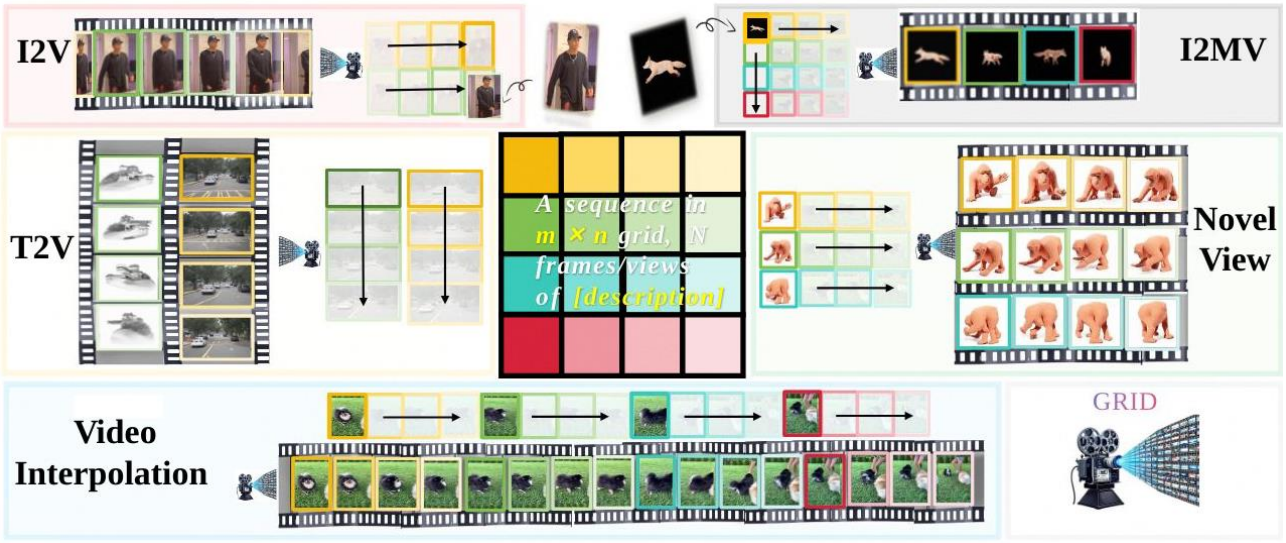

Figure 3: **Omni Inference Framework**: By transforming temporal and view sequences into structured layout spaces, we enable a pure image-based model FLUX to tackle diverse video and multi-view tasks (text/image-to-video generation, video interpolation, and multi-view synthesis) through a unified pipeline without additional video-specific architectures.

The generation process requires both flexibility and reference consistency. To achieve this, we introduce controlled grid noise injection instead of starting from pure noise:

$$\mathbf{I}_T = (1-T)\mathbf{I} + T\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{9}$$

where T denotes the time. This noise injection enables diverse generation while retaining the initialization structure.

To maintain reference consistency during generation, we employ a binary mask $M \in \{0,1\}^{m\times n}$:

$$M_{ij} = \begin{cases} 0 & \text{if } (i,j) \text{ contains reference frame} \\ 1 & \text{otherwise} \end{cases} \tag{10}$$

This mask modulates the update process:

$$\mathbf{I}_t = (1-M) \odot \mathbf{I}_{\text{ref}} + M \odot \mathbf{I}_t \tag{11}$$

ensuring reference frames remain unchanged while allowing other regions to evolve. The noise level T plays a key role in balancing generation quality. A large T leads to pure noise with poor reference consistency, while a small T yields near-duplicates. Our experiments show $T \in [0.8, 1.0]$ a good balance between diversity and fidelity.

3. Experiments

3.1. Experimental Setup

Datasets We train our model separately for video generation and multi-view generation tasks, both following a two-stage strategy: (1) For coarse-level training, we combine video clips from WebVid (Bain et al., 2021), and TikTok (Jafarian & Park, 2022) arranged in 8×8 and 4×4 grid layouts for video generation, and 30K sequences from Objaverse (Deitke et al., 2023) in 4×6 grids for multi-view generation. Each sequence is paired with automated captions and GLM-generated annotations emphasizing spatial and temporal relationships, using the sequence's inherent attributes (e.g., category labels) and visual content as queries. (2) For fine-grained control, we construct high-quality datasets of 1K sequences with structured annotations for each task. We first manually create exemplar annotations to establish a consistent format, then use these as few-shot examples for GPT-4o to generate precise control instructions while maintaining annotation consistency across the dataset.

Implementation Details We implement GRID based on the FLUX-dev, initializing from its pretrained weights. For video generation training, we adopt LoRA with ranks of 16-256, training for 10K steps with batch size 4 across 8 A800 GPUs using AdamW optimizer (learning rate 1e-4). The temporal loss weight α starts from 0 and gradually increases to a maximum of 0.5. For multi-view generation, we train on 30K sequences for 1.5K steps using LoRA rank 256 and Ours-EF using LoRA rank 16. During inference, we use a guidance scale of 3.5 and sampling step of 20.

Evaluation Protocol We evaluate our model on three distinct generation tasks: (1) Text-to-video generation on UCF-101 dataset (Soomro et al., 2012), evaluated using FVD (Unterthiner et al., 2019) (I3D backbone) and IS (Xu et al., 2018). We evaluate both 16-frame and 64-frame generation settings; (2) Image-to-video generation on a randomly sampled subset of 100 TikTok videos, measured by FVD and

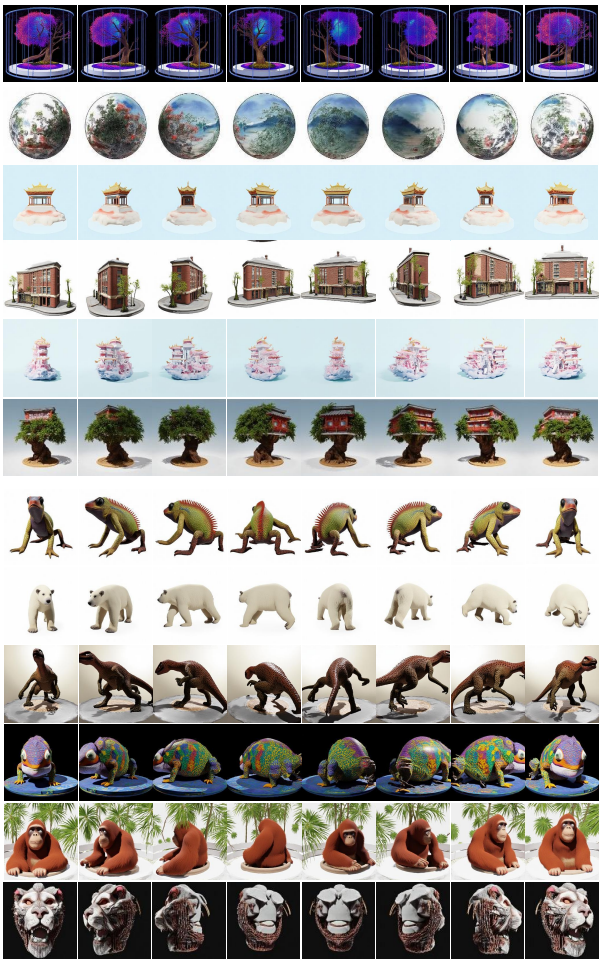

Figure 4: Multi-view generation results for static objects (top six rows) and dynamic subjects (bottom six rows), demonstrating consistent appearance and structure across different viewpoints.

CLIP$_{img}$ score; (3) Multi-view generation on Objaverse, where we evaluate on 30 randomly selected objects with 24 frames per sequence at different viewpoints to assess 4D generation capabilities. We compute FVD, CLIP metrics, following (Liang et al., 2024).

3.2. Main Results

We compare our approach with several state-of-the-art methods from well-established video/multyview generation model series, all of which represent the current frontiers in their respective domains.

Multi-view Generation We evaluate on the Objaverse test set with 30 3D objects. As shown in Table 1, our method achieves **state-of-the-art performance** on both text-to-multiview and image-to-multiview tasks. For T2MV, we improve CLIP-F to **0.9427** and reduce FVD to **324.3**, while achieving **67**× faster inference (6m vs. 405m) compared to 4DFY. For I2MV, we achieve **0.9486** CLIP-F score with **35**× speedup over STAG4D. Ours-EF (lora rank 16) also demonstrates strong performance-speed trade-off.

Figure 5: Text-to-Video Generation of driving scenes, showcasing complex multi-vehicle scenarios which represent the most challenging aspects of driving scene generation.

Figure 6: Image-to-Video Generation of dance sequences from TikTok dataset. The leftmost column shows the input reference image, followed by generated motion sequences.

Text-to-Video Generation As shown in Table 2, we achieve competitive FVD of 721.6 for 64-frame generation. For 16-frame generation, our method achieves **6.7**× **faster inference** (7.2s vs 48s) compared to CogVideo, with the efficiency gap widening to **5.5**× for 64-frame tasks. Our staged training shows clear progression: Stage1 achieves FVD 455.3, improving to **401.1** with fine-grained annotations, and further to **382.5** with $\mathcal{L}_{flow}$.

Image-to-Video Generation We evaluate on the TikTok dataset containing 100 diverse short videos. Our method achieves breakthrough performance with FVD of **93.7** (**23%** improvement) and CLIP$_{img}$ score of **0.9709**. Notably, our approach requires only **160M** parameters, compared to

Table 1: Quantitative comparison of Multi-view Generation Results on Text-to-Multiview and Image-to-Multiview Tasks. Inf Time indicates the **whole** time cost during inference.

Text-to-Multiview (T2MV)					**Image-to-Multiview (I2MV)**				
Method	CLIP-F↑	CLIP-O↑	FVD↓	Inf Time↓	Method	CLIP-F↑	CLIP-O↑	FVD↓	Inf Time↓
Animate124	0.7889	0.6005	411.6	180m	STAG4D	0.8803	0.6420	475.4	210m
4DFY	0.8092	0.6163	390.4	405m	4DGen	0.8724	0.6397	525.2	130m
Ours-EF	0.9060	0.6189	355.6	6m	Ours-EF	0.9392	**0.6580**	**333.7**	6m
Ours	**0.9427**	**0.6247**	**324.3**	**6m**	**Ours**	**0.9486**	0.6554	350.6	**6m**

Table 2: **Comprehensive Generation Results.** Our model achieves competitive quality with **superior efficiency** across tasks. While existing methods are limited to 16-frame generation, our approach efficiently scales to 64-frame sequences with linear time cost. Underlined and **bold** values indicate best results among our variants and all methods, respectively. Test Time shows average sampling time per sequence.

Text-to-Video (16-frame)				
Method	FVD↓	IS↑	Inf Time↓	Para↓
AnimateDiffv3	464.1	35.24	12s	419M
VideoCrafter2	424.2	32.00	15s	919M
OpenSora1.2	472.0	**39.07**	12s	1.5B
Cosmos	399.7	35.54	275s	7B
Ours(Stage1)	455.3	32.46	7.2s	**160M**
Ours(Stage1+2)	401.1	36.56	7.2s	
Ours(Full)	**382.5**	38.12	**7.2s**	
Text-to-Video (64-frame)				
Method	FVD↓	IS↑	Inf Time↓	Para↓
OpenSora1.2	1000.5	**37.11**	66s	1.5B
CogVideo5b	740.1	34.82	132s	5B
Ours(Stage1)	1003.2	32.48	24s	**160M**
Ours(Stage1+2)	994.6	36.47	24s	
Ours(Full)	**721.6**	36.63	**24s**	
Image-to-Video				
Method	FVD↓	CLIP$_{img}$ ↑	Inf Time↓	Para↓
AnimateDiffv3	250.9	0.9229	12s	419M
CogVideo5b	122.5	0.9185	48s	5B
Ours(Stage1)	115.5	0.9598	7.2s	**160M**
Ours(Stage1+2)	104.6	0.9695	7.2s	
Ours(Full)	**93.7**	**0.9709**	**7.2s**	

Table 3: **Video Frame Interpolation Results on UCF101.** We evaluate our full model following standard settings. All methods achieve comparable results, with our approach matching state-of-the-art EMA-VFI on PSNR.

Metrics	EMA-VFI	UPR-Net	VFIMamba	Ours
PSNR↑	**35.48**	35.47	35.45	**35.48**
SSIM↑	0.9701	0.9700	**0.9702**	0.9700

>400M for motion modeling or >1B for full generation in existing methods.

Video Frame Interpolation We evaluate on the UCF101 dataset for video frame interpolation (Zhang et al., 2023b; Jin et al., 2023; Zhang et al., 2024). As shown in Table 3, our approach achieves **state-of-the-art PSNR of 35.48**, matching EMA-VFI. For SSIM, all methods perform comparably around 0.970, with VFIMamba leading marginally.

3.3. Extension Capabilities

Beyond the primary generation tasks, we demonstrate GRID's strong zero-shot generalization capabilities across diverse video and multi-view applications without any task-specific training or architectural modifications. The layout-based design enables natural adaptation to various downstream tasks through prompt engineering alone.

Video Motion Clone Our framework enables natural video motion cloning through image redrawing without additional training. As demonstrated in Figure 7, we transform a cat video into videos featuring a fox, red panda, and tiger, while faithfully preserving the original motion patterns, temporal dynamics, and scene aesthetics.

Video Restoration Our architecture's multi-scale processing capability enables effective video restoration without explicit training. Figure 13 shows our model's performance in recovering high-quality videos from severely degraded inputs (with Gaussian blur and block masking).

3D Editing We demonstrate our model's potential for practical 3D appearance editing through an innovative virtual try-on application. As shown in Figure 12, given an uncolored 3D human walking sequence from multiple viewpoints, our model can dress and style the figure through simple text prompts. This enables diverse appearance variations - from adding hair to rendering outfits - while maintaining consistent 3D structure and motion.

More results and applications are shown in Appendix A.7.

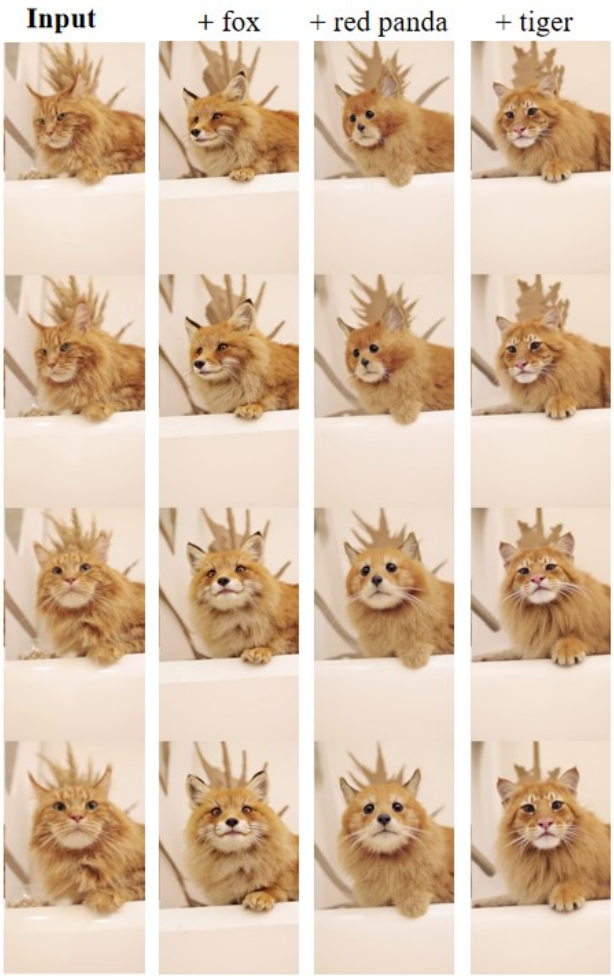

Figure 7: Zero-shot video motion clone results. Our model incorporates characteristics from different animals (fox, red panda, tiger) while maintaining motion pattern.

4. Related Work

Text-to-Image Generation Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) have fundamentally transformed image generation by employing iterative denoising processes to synthesize high-quality outputs. Subsequent advancements (Rombach et al., 2022; Podell et al., 2023; Ramesh et al., 2022; Saharia et al., 2022) have refined this paradigm leveraging latent spaces with significantly reduced computational costs. Diffusion Transformers (DiT) (Peebles & Xie, 2023) further advanced this area by replacing the U-Net architecture with transformer-based designs. This architectural shift improved training efficiency, paving the way for more scalable and versatile generative frameworks. Building on these, flow matching (Lipman et al., 2022; Esser et al., 2024) reformulates the generation process as a straight-path trajectory between data and noise distributions. More recently, FLUX (BlackForest, 2024), has combined the strengths of DiT and flow matching to achieve efficient and high-quality image generation. These models also integrate powerful language models (Raffel et al., 2020) and joint text-image attention mechanisms. This multimodal understanding has unlocked new possibilities for instruction-following and creative applications. Beyond generating high-quality images, text-to-image models demonstrate a strong spatial understanding that can be naturally extended to temporal dimensions through layout representations, enabling diverse downstream tasks.

Task-Specific Generation Diffusion-based approaches have shown remarkable progress in generalized video generation tasks (Ho et al., 2022; Blattmann et al., 2023b; Zhang et al., 2023a; Blattmann et al., 2023a; He et al., 2023; Zhou et al., 2022; Wang et al., 2023a; Ge et al., 2023; Wang et al., 2023c;b; Singer et al., 2022; Zhang et al., 2023a; Zeng et al., 2023; Agarwal et al., 2025). Notable works like VideoLDM (Blattmann et al., 2023b), Animatediff (Guo et al., 2023), and SVD (Chai et al., 2023) advance temporal modeling through specialized architectures. In the multi-view domain, various approaches (Watson et al., 2022; Liu et al., 2023a; Shi et al., 2023b; Long et al., 2024; Shi et al., 2023a; Lu et al., 2024; Li et al., 2023; Liu et al., 2023b; Li et al., 2024; Yang et al., 2024; Zhao et al., 2023; Yin et al., 2023) focus on cross-view consistency through different attention mechanisms and feature space alignments. Recent 4D generation methods (Ren et al., 2023; Liang et al., 2024; Xie et al., 2024b; Sun et al., 2024; Wu et al., 2024) further extend to joint spatial-temporal synthesis, though often facing efficiency challenges or requiring multi-step generation. While these methods achieve remarkable results, they are typically tailored to specific tasks, relying on specialized architectures for image, video, or multi-view generation. Additionally, methods like VideoPoet (Kondratyuk et al., 2023) employ complex cross-modal alignment mechanisms to bridge different generation modes. In contrast, our approach introduces layout generation, an omni framework that transforms temporal and spatial generation into layout representations. This enables seamless multi-modal generation, to address a wide range of tasks through straightforward modifications to input representations, without the need for complex cross-modal alignment mechanisms.

5. Conclusion

We present GRID, an omni visual generation framework through grid representation. Our two-stage training strategy enables both robust generation and precise control, while the temporal refinement mechanism enhances motion coherence. Experiments demonstrate significant computational efficiency gains while maintaining competitive performance across tasks. The framework's strong zero-shot generalization capabilities further enable adaptation to diverse applications without task-specific training, suggesting a promising direction for efficient visual sequence generation.

Impact Statement

This paper introduces research aimed at advancing visual sequence generation through an efficient layout-based framework. However, we must emphasize the potential risks associated with this technology, particularly in facial manipulation applications (Xie et al., 2024a; Luo et al., 2024), where our method could be misused to compromise identity security. Nevertheless, recent advances in adversarial perturbation protection mechanisms (Wan et al., 2024) provide solutions to help users protect their personal data against unauthorized model fine-tuning and malicious content generation. Therefore, we call for attention to these risks and encourage the adoption of defensive techniques to ensure the protection of personal content while advancing the development of generative AI technologies.

References

Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Bai, Y., Wu, D., Liu, Y., Jia, F., Mao, W., Zhang, Z., Zhao, Y., Shen, J., Wei, X., Wang, T., et al. Is a 3d-tokenized llm the key to reliable autonomous driving? *arXiv preprint arXiv:2405.18361*, 2024.

Bain, M., Nagrani, A., Varol, G., and Zisserman, A. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1728–1738, 2021.

Baldridge, J., Bauer, J., Bhutani, M., Brichtova, N., Bunner, A., Chan, K., Chen, Y., Dieleman, S., Du, Y., Eaton-Rosen, Z., et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

BlackForest. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, pp. 22563–22575, 2023b.

Chai, W., Guo, X., Wang, G., and Lu, Y. Stablevideo: Text-driven consistency-aware diffusion video editing. In *CVPR*, pp. 23040–23050, 2023.

Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., and Farhadi, A. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023.

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., and Tang, J. GLM: general language model pretraining with autoregressive blank infilling. pp. 320–335, 2022.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.-B., Liu, M.-Y., and Balaji, Y. Preserve your own correlation: A noise prior for video diffusion models. In *CVPR*, pp. 22930–22941, 2023.

Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., and Dai, B. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

He, Y., Yang, T., Zhang, Y., Shan, Y., and Chen, Q. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2(3):4, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Huang, L., Wang, W., Wu, Z.-F., Dou, H., Shi, Y., Feng, Y., Liang, C., Liu, Y., and Zhou, J. Group diffusion transformers are unsupervised multitask learners. *arXiv preprint arxiv:2410.15027*, 2024a.

Huang, L., Wang, W., Wu, Z.-F., Shi, Y., Dou, H., Liang, C., Feng, Y., Liu, Y., and Zhou, J. In-context lora for diffusion transformers. *arXiv preprint arxiv:2410.23775*, 2024b.

Jafarian, Y. and Park, H. S. Self-supervised 3d representation learning of dressed humans from social media videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (7):8969–8983, 2022.

Jin, X., Wu, L., Chen, J., Chen, Y., Koo, J., and Hahm, C.-h. A unified pyramid recurrent network for video frame interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2023.

Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Schindler, G., Hornung, R., Birodkar, V., Yan, J., Chiu, M.-C., et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

Li, J., Tan, H., Zhang, K., Xu, Z., Luan, F., Xu, Y., Hong, Y., Sunkavalli, K., Shakhnarovich, G., and Bi, S. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023.

Li, P., Liu, Y., Long, X., Zhang, F., Lin, C., Li, M., Qi, X., Zhang, S., Luo, W., Tan, P., et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *arXiv preprint arXiv:2405.11616*, 2024.

Liang, H., Yin, Y., Xu, D., Liang, H., Wang, Z., Plataniotis, K. N., Zhao, Y., and Wei, Y. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models. *arXiv preprint arXiv:2405.16645*, 2024.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., and Vondrick, C. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023a.

Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., and Wang, W. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023b.

Long, X., Guo, Y.-C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.-H., Habermann, M., Theobalt, C., et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9970–9980, 2024.

Lu, Y., Zhang, J., Li, S., Fang, T., McKinnon, D., Tsin, Y., Quan, L., Cao, X., and Yao, Y. Direct2. 5: Diverse text-to-3d generation via multi-view 2.5 d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8744–8753, 2024.

Luo, X., Zhang, X., Xie, Y., Tong, X., Yu, W., Chang, H., Ma, F., and Yu, F. R. Codeswap: Symmetrically face swapping based on prior codebook. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6910–6919, 2024.

OpenAI. GPT-4 technical report. *arXiv:2303.08774*, 2023.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Ren, J., Pan, L., Tang, J., Zhang, C., Cao, A., Zeng, G., and Liu, Z. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Karagol Ayan, B., Mahdavi, S. S., Gontijo Lopes, R., Salimans, T., Ho, J., Fleet, D., and Norouzi, M. Imagen: unprecedented photorealism × deep level of language understanding, 2022.

Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., and Su, H. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023a.

Shi, Y., Wang, P., Ye, J., Long, M., Li, K., and Yang, X. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023b.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Sun, W., Chen, S., Liu, F., Chen, Z., Duan, Y., Zhang, J., and Wang, Y. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024.

Tian, K., Jiang, Y., Yuan, Z., Peng, B., and Wang, L. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024.

Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. Fvd: A new metric for video generation. 2019.

Wan, C., He, Y., Song, X., and Gong, Y. Prompt-agnostic adversarial perturbation for customized diffusion models. *arXiv preprint arXiv:2408.10571*, 2024.

Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., and Zhang, S. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023a.

Wang, W., Yang, H., Tuo, Z., He, H., Zhu, J., Fu, J., and Liu, J. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023b.

Wang, X., Xie, L., Dong, C., and Shan, Y. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*.

Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Chen, X., Wang, Y., Luo, P., Liu, Z., et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023c.

Watson, D., Chan, W., Martin-Brualla, R., Ho, J., Tagliasacchi, A., and Norouzi, M. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022.

Wu, R., Gao, R., Poole, B., Trevithick, A., Zheng, C., Barron, J. T., and Holynski, A. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv preprint arXiv:2411.18613*, 2024.

Xie, Y., Xu, H., Song, G., Wang, C., Shi, Y., and Luo, L. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024a.

Xie, Y., Yao, C.-H., Voleti, V., Jiang, H., and Jampani, V. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*, 2024b.

Xu, Q., Huang, G., Yuan, Y., Guo, C., Sun, Y., Wu, F., and Weinberger, K. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*, 2018.

Yang, X., Shi, H., Zhang, B., Yang, F., Wang, J., Zhao, H., Liu, X., Wang, X., Lin, Q., Yu, J., et al. Hunyuan3d-1.0: A unified framework for text-to-3d and image-to-3d generation. *arXiv preprint arXiv:2411.02293*, 2024.

Yin, Y., Xu, D., Wang, Z., Zhao, Y., and Wei, Y. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*, 2023.

Zeng, Y., Wei, G., Zheng, J., Zou, J., Wei, Y., Zhang, Y., and Li, H. Make pixels dance: High-dynamic video generation. *arXiv preprint arXiv:2311.10982*, 2023.

Zhang, D. J., Wu, J. Z., Liu, J.-W., Zhao, R., Ran, L., Gu, Y., Gao, D., and Shou, M. Z. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023a.

Zhang, G., Zhu, Y., Wang, H., Chen, Y., Wu, G., and Wang, L. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5682–5692, 2023b.

Zhang, G., Liu, C., Cui, Y., Zhao, X., Ma, K., and Wang, L. Vfimamba: Video frame interpolation with state space models, 2024. URL https://arxiv.org/abs/2407.02315.

Zhao, Y., Yan, Z., Xie, E., Hong, L., Li, Z., and Lee, G. H. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023.

Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., and Feng, J. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

A. Appendix

A.1. Why Flux? Zero-shot Analysis of Foundation Models

To better understand the layout capabilities of existing models before fine-tuning, we conducted a comprehensive zero-shot evaluation comparing three state-of-the-art models: DALLE-3, Flux, and Imagen3. Figure 8 presents their generation results, with each row corresponding to DALLE-3 (top), Flux (middle), and Imagen3 (bottom) respectively.

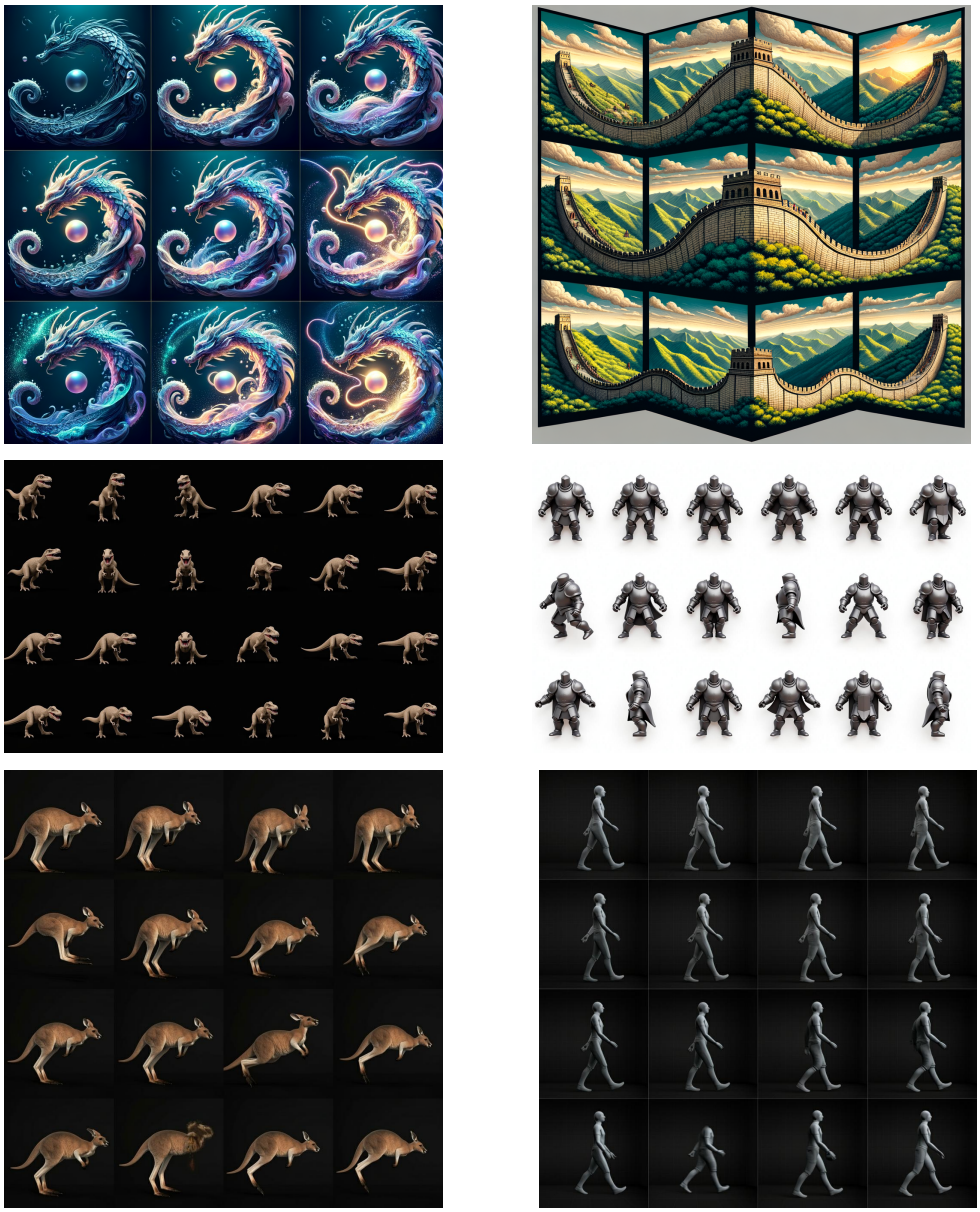

Figure 8: Zero-shot evaluation of foundation models on grid-based multi-view generation tasks before we begin to train. Using the prompt "a * from different angles in a mxn grid layout," First row: Dalle3, Second row: Flux, Third row: Imagen3.

Our analysis reveals varying degrees of grid layout understanding across models. While all models demonstrate basic grid comprehension, they exhibit different strengths and limitations. For motion control, we observe that precise directional instructions (e.g., clockwise rotation) often result in random orientations across all models, indicating limited spatial-temporal control capabilities.

In terms of grid structure accuracy, DALLE-3 shows inconsistent interpretation of specific layout requirements (e.g., 4×4 or

4×6 grids), while Flux and Imagen3 demonstrate better adherence to specified grid configurations. Notably, Flux exhibits superior understanding of spatial arrangements.

Content consistency across grid cells varies significantly. Both Imagen3 and DALLE-3 show noticeable variations in object appearance across frames, while Flux maintains better consistency in object characteristics throughout the sequence. This superior consistency, combined with its open-source nature, motivated our choice of Flux as the base model for our framework.

A.2. Why is it Natural for GRID to Leverage Built-in Attention Mechanism

Video generation fundamentally requires three key capabilities: spatial understanding within frames, temporal consistency between frames, and semantic control across the entire sequence. Traditional approaches tackle these requirements by implementing separate attention modules, as shown in Figure 9(a). While this modular design directly addresses each requirement, it introduces architectural complexity and potential inconsistencies between modules.

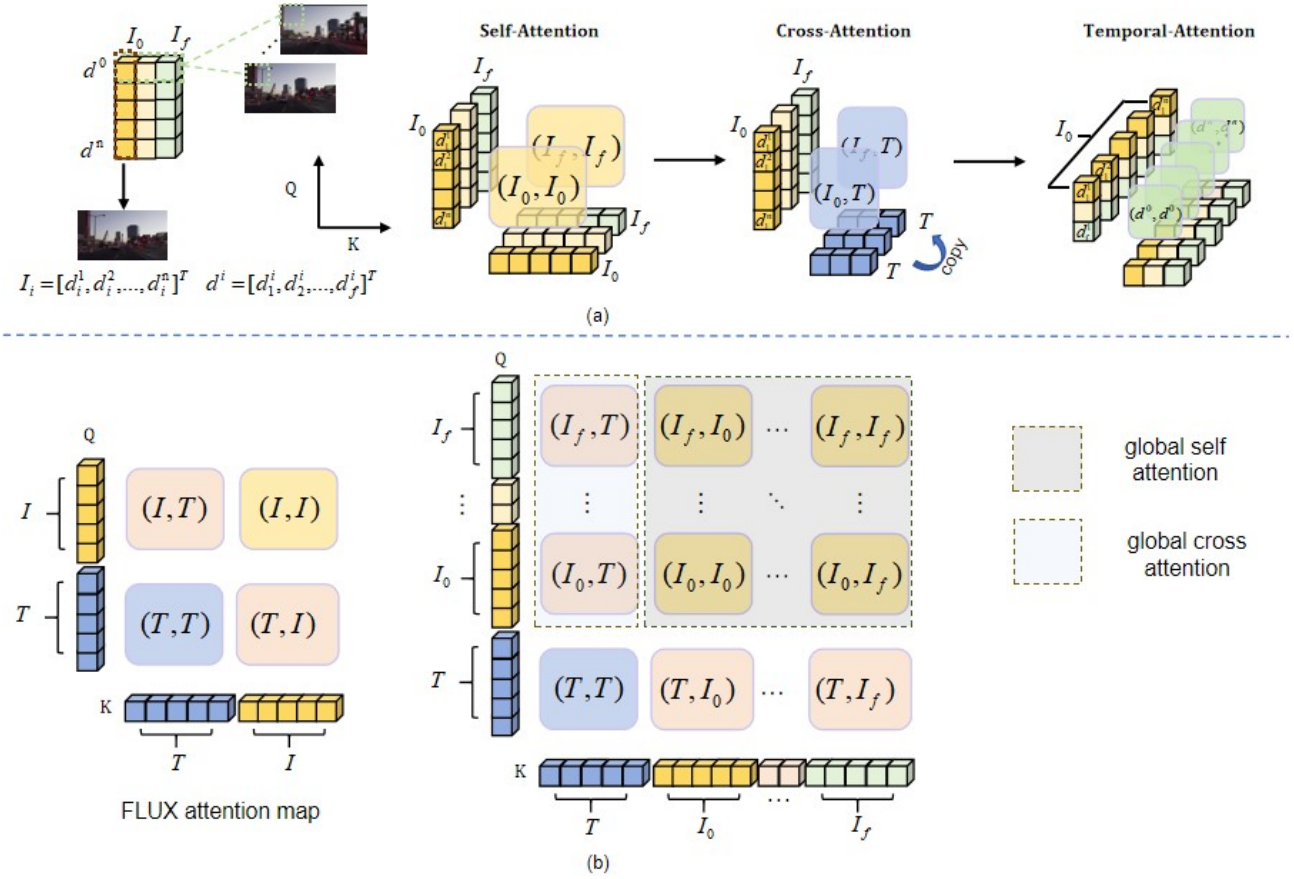

Figure 9: Comparison of attention mechanisms. (a) Traditional video diffusion models rely on three separate attention modules to handle spatial understanding, semantic guidance, and temporal consistency respectively. (b) Through our grid layout reformulation, FLUX's unified self-attention naturally encompasses both inner-frame (I_i, I_i) and cross-frame (I_i, I_j) relationships, while its global text-image attention (T, I) enables consistent control across all frames. This simplification eliminates the need for specialized temporal modules while maintaining effective spatio-temporal understanding.

Our key insight is that these seemingly distinct requirements can be unified through spatial reformulation. By organizing temporal sequences into grid layouts, we transform temporal relationships into spatial ones, allowing FLUX's native attention mechanism to naturally handle all requirements through a single, coherent process.

This unification works through two complementary mechanisms, as illustrated in Figure 9(b). First, the original image self-attention (I, I) automatically extends across the grid structure. When processing grid cells containing different temporal

frames, this self-attention naturally splits into inner-frame attention (I_i, I_i) and cross-frame attention (I_i, I_j). The inner-frame component maintains spatial understanding within each frame, while the cross-frame component captures temporal relationships - effectively handling both spatial and temporal coherence through a single mechanism.

Second, the text-image cross-attention $(T, [I_i]_{i=0}^{f})$ operates globally across all grid cells, enabling unified semantic control. This global operation ensures that textual instructions consistently influence all frames, maintaining semantic coherence throughout the sequence. The grid layout allows this semantic guidance to naturally incorporate both content and temporal specifications, as the attention mechanism can reference the spatial relationships between grid cells.

This reformulation fundamentally changes how temporal information is processed. Rather than treating temporal relationships as a separate problem requiring specialized mechanisms, we transform them into spatial relationships that existing attention mechanisms are already optimized to handle. This approach not only simplifies the architecture but also provides more robust temporal understanding, as it leverages the well-established capabilities of spatial attention mechanisms.

The elegance of this solution lies in its ability to achieve complex temporal processing without architectural modifications. By thoughtfully restructuring the problem space, we enable standard attention mechanisms to naturally extend their capabilities, demonstrating how strategic problem reformulation can be more powerful than architectural elaboration.

A.3. Comparison with Existing Approaches and Computational Efficiency Analysis

Current approaches to video generation can be categorized into two distinct paradigms, each with fundamental limitations in terms of architectural design and computational requirements. We provide a detailed analysis of these approaches and contrast them with our method:

Paradigm 1: Image Models as Single-Frame Generators Methods like SVD and AnimateDiff utilize pre-trained text-to-image models as frame generators while introducing separate modules for motion learning. This approach presents several fundamental limitations:

First, these methods require complex architectural additions for temporal modeling, introducing significant parameter overhead without leveraging the inherent capabilities of pre-trained image models. For instance, AnimateDiff introduces temporal attention layers that must be trained from scratch, while SVD requires separate motion estimation networks.

Second, the sequential nature of frame generation in these approaches leads to substantial computational overhead during inference. This sequential processing not only impacts generation speed but also limits the model's ability to maintain long-term temporal consistency, as each frame is generated with limited context from previous frames.

Paradigm 2: End-to-End Video Architectures Recent approaches like Sora, CogVideo, and Huanyuan Video attempt to solve video generation through end-to-end training of video-specific architectures. While theoretically promising, these methods face severe practical constraints:

The computational requirements are particularly striking:

- CogVideo requires approximately 35M video clips and an additional 2B filtered images from LAION-5B and COYO-700M datasets
- Open-Sora necessitates more than 35M videos for training
- These models typically demand multiple 80GB GPUs with sequence parallelism just for inference
- Training typically requires thousands of GPU-days, making reproduction and iteration challenging for most research teams

Our Grid-based Framework: A Resource-Efficient Alternative In contrast, GRID achieves competitive performance through a fundamentally different approach:

1. Architectural Efficiency: Our grid-based framework requires only 160M additional parameters while maintaining competitive performance. This efficiency stems from:

- Treating temporal sequences as spatial layouts, enabling parallel processing

- Leveraging existing image generation capabilities without architectural complexity
- Efficient parameter sharing across temporal and spatial dimensions

2. Data Efficiency: We achieve remarkable data efficiency improvements:

$$\text{Data Reduction} \approx \frac{> 35M \text{ videos (previous methods)}}{< 35K \text{ videos (our method)}} = 1000\times \tag{12}$$

This efficiency is achieved through:

- Strategic use of grid-based training that maximizes information extraction from each video
- Effective transfer learning from pre-trained image models
- Focused training on essential video-specific components

3. Computational Accessibility: Our approach enables high-quality video generation while maintaining accessibility for research environments with limited computational resources:

- Training can be completed on standard research GPUs
- Inference requires significantly less memory compared to end-to-end approaches
- The model maintains strong performance across both video and image tasks

This comprehensive analysis demonstrates that our approach not only addresses the limitations of existing methods but also achieves substantial improvements in computational efficiency while maintaining competitive performance. The significant reductions in data requirements and computational resources make our method particularly valuable for practical applications and research environments with limited resources.

A.4. Distinction from In-Context LoRA

Recent work IC-LoRA (Huang et al., 2024b;a) also utilizes grid-based layouts for image generation, which might superficially appear similar to our approach. However, a careful analysis reveals fundamental differences in both theoretical foundation and technical implementation.

Different Theoretical Foundations: The core principle of IC-LoRA is to use grid layouts as a prompt engineering technique, where multiple images are arranged together to provide in-context examples for task adaptation. This is essentially an extension of in-context learning from language models to visual domain. Their grid layout serves merely as a presentation format for example-based learning.

In contrast, our approach fundamentally re-conceptualizes temporal sequences into spatial layouts. Rather than using grids for example presentation, we treat them as an inherent representation of temporal information, where spatial relationships in the grid directly correspond to temporal relationships in the sequence. This enables our model to learn and generate temporal dynamics in a holistic manner.

Distinct Technical Objectives: IC-LoRA's technical implementation focuses on task adaptation through example pairs. Their method relies on LoRA-based fine-tuning and natural language prompts to define relationships between grid elements. However, this approach has inherent limitations in handling temporal dynamics, as it treats each grid element independently without explicit modeling of their temporal relationships.

Our method, on the other hand, is specifically designed for temporal sequence generation. We introduce parallel flow-matching and dedicated temporal loss functions that explicitly model motion patterns and temporal coherence. This allows our approach to capture and generate complex temporal dynamics that are beyond the capability of example-based methods like IC-LoRA.

Different Application Scopes: While IC-LoRA excels at static, example-based generation tasks through prompt engineering, it struggles with temporal sequence generation due to its fundamental design limitations. Our method, specifically designed

for temporal modeling, naturally handles both static and dynamic visual generation tasks while maintaining precise control over temporal dynamics.

This analysis demonstrates that despite the superficial similarity in using grid layouts, our approach represents a fundamentally different direction in visual generation. We independently developed our method to address the specific challenges of temporal sequence generation, resulting in distinct technical contributions that go beyond the capabilities of example-based frameworks like IC-LoRA.

These crucial differences are evidenced by our method's superior performance in temporal tasks and its ability to maintain consistent motion patterns across sequences - capabilities that are fundamentally beyond the scope of IC-LoRA's example-based approach.

A.5. Inference Details

For extension tasks (style transfer, restoration, and editing), we modify the omni-inference framework to process full sequences while maintaining temporal coherence. Unlike the reference-guided generation that requires partial initialization and masking, these tasks operate on complete sequences with controlled noise injection for appearance modification.

Given an input sequence represented as a grid structure $\mathbf{I} = (I_{ij})_{m\times n}$, we initialize the generation process with noise-injected states:

$$\mathbf{I}_T = (1 - T)\mathbf{I} + T\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{13}$$

where $T \in [0.8, 0.9]$ represents a lower noise level compared to the reference-guided generation. This lower T value helps preserve the original temporal structure while allowing sufficient flexibility for appearance modifications.

A.6. Post-Processing Pipeline

For multi-view generation results, we employ a two-stage enhancement process. First, the generated sequences are processed as video frames to ensure temporal consistency. Subsequently, we apply super-resolution using Real-ESRGAN (Wang et al.) with anime-video-v3 weights, upscaling from 256×256 resolution to 1024×1024. This enhancement pipeline significantly improves visual quality while maintaining temporal coherence.

Table 4 shows parts of our inference prompts for multyview generation. We basically follow this prompt format.

A.7. Potential Applications

Our framework demonstrates significant potential beyond its primary applications.

A.7.1. Creative Multi-view Generation

As shown in Figure 10, our method exhibits remarkable flexibility in combining different conceptual elements to create novel multi-view compositions. The grid-based layout allows for intuitive arrangement and manipulation of various visual elements, enabling creative expressions that would be challenging for traditional approaches. This capability suggests promising applications in creative design, artistic visualization, and content creation.

A.7.2. Flexible Frame Extension

Notably, our model demonstrates strong generalization capability in sequence length. Despite being trained on 4×4 (16-frame) driving scenarios, the model can effectively generate 4×8 (32-frame) sequences by simply adjusting the c_layout prompt at inference time. As shown in Figure 11, the extended sequences maintain temporal consistency and visual quality comparable to the original training length. This flexibility suggests that our layout-based approach naturally accommodates variable-length generation without requiring explicit retraining, opening possibilities for dynamic content generation across different temporal scales.

A.7.3. Future Extension to Video Understanding

Our layout-based framework shows potential in transforming traditional video understanding tasks into image-domain problems. Unlike conventional autoregressive approaches (Bai et al., 2024) that process frames sequentially, our method arranges frames in a grid layout, enabling parallel processing and global temporal modeling. This approach could benefit

Common Format	A 24-frame sequence arranged in a 4x6 grid. Each frame captures a 3D model of [subject] from a different angle, rotating 360 degrees. The sequence begins with a front view and progresses through a complete clockwise rotation
Category	**Subject Description**
Creative Fusion	a skyscraper with knitted wool surface and cable-knit details a mechanical hummingbird with clockwork wings and steampunk gears hovering near a neon flower a bonsai tree with spiral galaxies and nebulae blooming from its twisted branches a phoenix crafted entirely from woven bamboo strips with intricate basketwork details glowing from within a jellyfish with a transparent porcelain bell decorated in blue-and-white patterns and ink-brush tentacles a coral reef made entirely of rainbow-hued blown glass with intricate marine life formations an urban street where buildings are shaped as giant functional musical instruments including a violin apartment and piano mall a butterfly with stained glass wings depicting medieval scenes catching sunlight a floating city where traditional Chinese pavilions rest on clouds made of flowing silk fabric in pastel colors a lion composed of moving gears and pistons that transforms between mechanical and organic forms a garden where geometric crystal formations grow and branch like plants with rainbow refractions a tree whose trunk is a twisting pagoda with branches of miniature traditional buildings and roof tile leaves a phoenix-dragon hybrid creature covered in mirrored scales that create fractal reflections a celestial teapot with constellation etchings pouring a stream of stars and nebulae an origami landscape where paper mountains continuously fold and unfold to reveal geometric cities and rivers a sphere where traditional Chinese ink and wash paintings flow continuously between day and night scenes
Natural Creatures	a Velociraptor in hunting pose with detailed scales and feathers a Mammoth with detailed fur and tusks a chameleon changing colors with detailed scales a white tiger in mid-stride with flowing muscles a Pterodactyl with spread wings in flight pose an orangutan showing intelligent behavior a polar bear with detailed fur texture

Table 4: **Prompt format for 360° object rotation generation.** All prompts follow the same structural template, varying only in the subject description. The subjects are categorized into creative fusion designs that combine different artistic elements and concepts, and natural creatures that focus on realistic animal representations.

various video understanding tasks: for video-text retrieval, the layout representation allows direct comparison between video content and text embeddings across all frames simultaneously; for video question answering, it enables the model to attend to relevant frames across the entire sequence without sequential constraints; for video tracking and other analysis tasks, it avoids error accumulation common in traditional sequential processing. While we have not conducted specific experiments in these directions, our framework's ability to convert temporal relationships into spatial ones through layouts offers a promising alternative to conventional video understanding paradigms, potentially enabling more efficient and effective multi-modal video analysis.

A.7.4. Maintained Image Generation Ability

Our framework preserves the original Flux model's image generation capabilities while extending its functionality to handle video sequences. As demonstrated in Figure 14, the model maintains high-quality performance on various image generation tasks such as text-to-image synthesis, image editing, and style transfer. This preservation of original capabilities alongside newly acquired video generation abilities creates a versatile model that can seamlessly handle both single-image and multi-frame tasks. The ability to maintain original image generation quality while adding new functionality demonstrates the effectiveness of our training approach and the robustness of the layout-based framework.

A.8. Limitations

Our approach faces two primary limitations. First, the grid-based layout design inherently constrains frame resolution due to limitations of the based Text-to-Image models when processing multiple frames simultaneously. Second, our training strategy, based on lora finetuning, shows limitations in text-to-video generation tasks that significantly deviate from the base model's capabilities. Combined with our relatively small training dataset, this makes it challenging to achieve competitive performance in open-world video generation scenarios requiring complex motion understanding.

A.9. Multyview Camera Parameters

Building upon the dataset opensourced by Diffusion4D (Liang et al., 2024), Table 5 presents camera trajectory parameters, which serve as the foundation for consistent 4D content generation and subsequent reconstruction tasks.

Our camera configuration follows precise mathematical relationships, with cameras positioned at 15-degree intervals along a circle of radius 2 units in the horizontal plane. The systematic progression of coordinate bases ensures optimal coverage while maintaining consistent inter-frame relationships. Each camera's orientation is defined by orthogonal basis vectors, with the Y vector consistently aligned with the negative Z-axis to establish stable up-direction reference.

Frame	X Vector	Y Vector	Z Vector	Origin
1	[1.0, 0.0, 0.0]	[-0.0, 0.0, -1.0]	[-0.0, 1.0, 0.0]	[0.0, -2.0, 0.0]
2	[0.96, 0.27, -0.0]	[0.0, -0.0, -1.0]	[-0.27, 0.96, -0.0]	[0.54, -1.93, 0.0]
3	[-0.92, 0.4, -0.0]	[0.0, 0.0, -1.0]	[-0.4, -0.92, -0.0]	[0.8, 1.83, 0.0]
4	[-0.99, 0.14, -0.0]	[0.0, 0.0, -1.0]	[-0.14, -0.99, -0.0]	[0.27, 1.98, 0.0]
5	[-0.99, -0.14, 0.0]	[-0.0, 0.0, -1.0]	[0.14, -0.99, -0.0]	[-0.27, 1.98, 0.0]
6	[-0.92, -0.4, 0.0]	[-0.0, 0.0, -1.0]	[0.4, -0.92, -0.0]	[-0.8, 1.83, 0.0]
7	[-0.78, -0.63, 0.0]	[-0.0, -0.0, -1.0]	[0.63, -0.78, 0.0]	[-1.26, 1.55, 0.0]
8	[-0.58, -0.82, -0.0]	[0.0, 0.0, -1.0]	[0.82, -0.58, 0.0]	[-1.63, 1.15, 0.0]
9	[-0.33, -0.94, -0.0]	[0.0, -0.0, -1.0]	[0.94, -0.33, 0.0]	[-1.88, 0.67, 0.0]
10	[-0.07, -1.0, -0.0]	[0.0, 0.0, -1.0]	[1.0, -0.07, 0.0]	[-2.0, 0.14, 0.0]
11	[0.2, -0.98, 0.0]	[0.0, -0.0, -1.0]	[0.98, 0.2, 0.0]	[-1.96, -0.41, 0.0]
12	[0.46, -0.89, 0.0]	[0.0, -0.0, -1.0]	[0.89, 0.46, 0.0]	[-1.78, -0.92, 0.0]
13	[0.85, 0.52, 0.0]	[-0.0, 0.0, -1.0]	[-0.52, 0.85, 0.0]	[1.04, -1.71, 0.0]
14	[0.68, -0.73, -0.0]	[-0.0, 0.0, -1.0]	[0.73, 0.68, 0.0]	[-1.46, -1.37, 0.0]
15	[0.85, -0.52, -0.0]	[0.0, 0.0, -1.0]	[0.52, 0.85, 0.0]	[-1.04, -1.71, 0.0]
16	[0.96, -0.27, 0.0]	[-0.0, -0.0, -1.0]	[0.27, 0.96, -0.0]	[-0.54, -1.93, 0.0]
17	[1.0, -0.0, 0.0]	[0.0, 0.0, -1.0]	[0.0, 1.0, 0.0]	[-0.0, -2.0, 0.0]
18	[0.68, 0.73, 0.0]	[0.0, 0.0, -1.0]	[-0.73, 0.68, 0.0]	[1.46, -1.37, 0.0]
19	[0.46, 0.89, -0.0]	[-0.0, -0.0, -1.0]	[-0.89, 0.46, 0.0]	[1.78, -0.92, 0.0]
20	[0.2, 0.98, -0.0]	[-0.0, -0.0, -1.0]	[-0.98, 0.2, 0.0]	[1.96, -0.41, 0.0]
21	[-0.07, 1.0, 0.0]	[-0.0, 0.0, -1.0]	[-1.0, -0.07, 0.0]	[2.0, 0.14, 0.0]
22	[-0.33, 0.94, 0.0]	[-0.0, -0.0, -1.0]	[-0.94, -0.33, 0.0]	[1.88, 0.67, 0.0]
23	[-0.58, 0.82, 0.0]	[-0.0, 0.0, -1.0]	[-0.82, -0.58, 0.0]	[1.63, 1.15, 0.0]
24	[-0.78, 0.63, -0.0]	[0.0, -0.0, -1.0]	[-0.63, -0.78, 0.0]	[1.26, 1.55, 0.0]

Table 5: Camera Parameters for 24 Frames

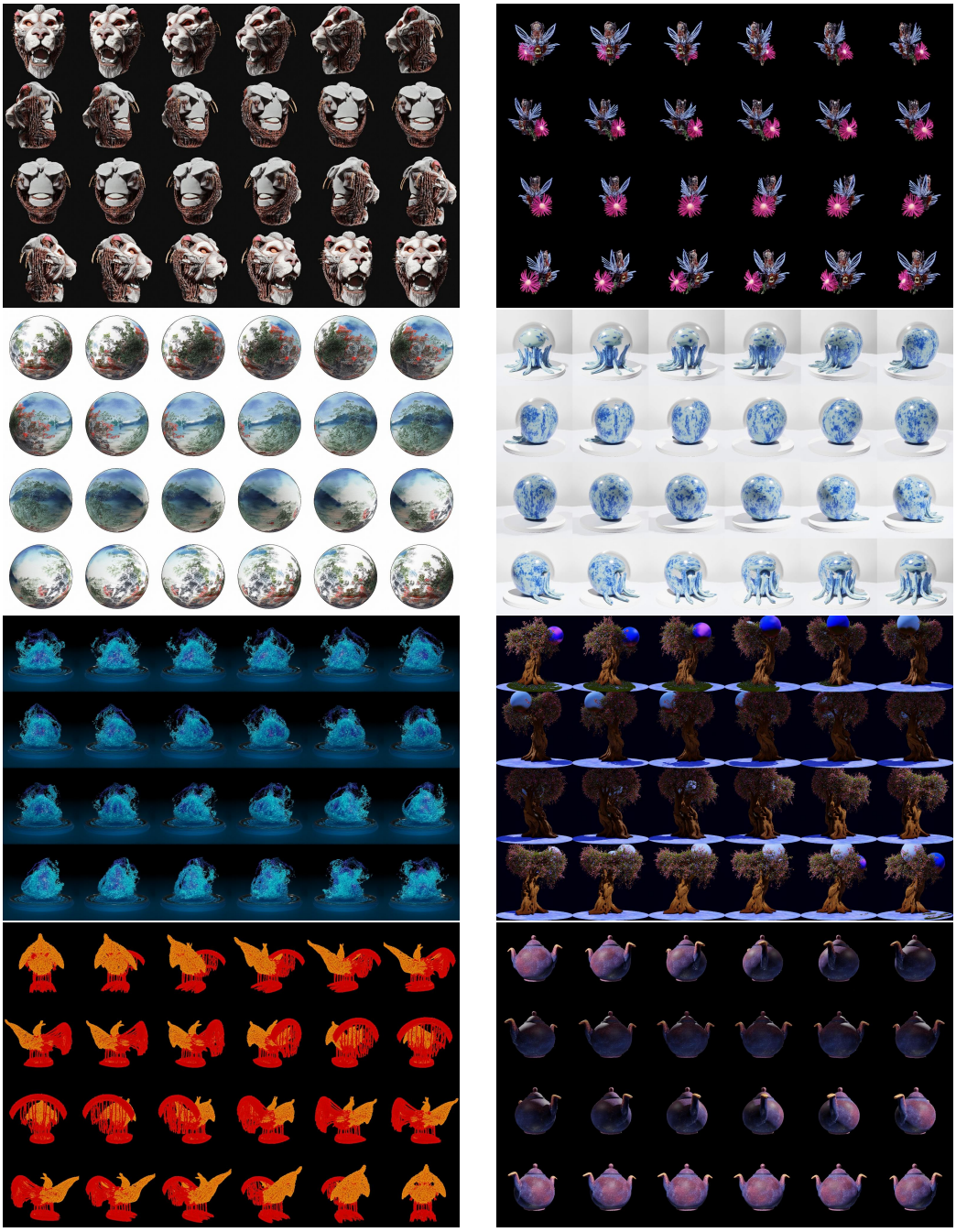

Figure 10: Creative multy-view concept generation.

Figure 11: We only train our model using 4×4 datasets, but when at inference, we directly change prompt to ask to layout 4×8 grid. The model has not trained on these kind of dataset, but show a zero-shot generalization ability.

Figure 12: Zero-shot 3D editing with attribute control. Our model generates diverse variations by modifying appearance attributes through text prompts while preserving motion patterns.

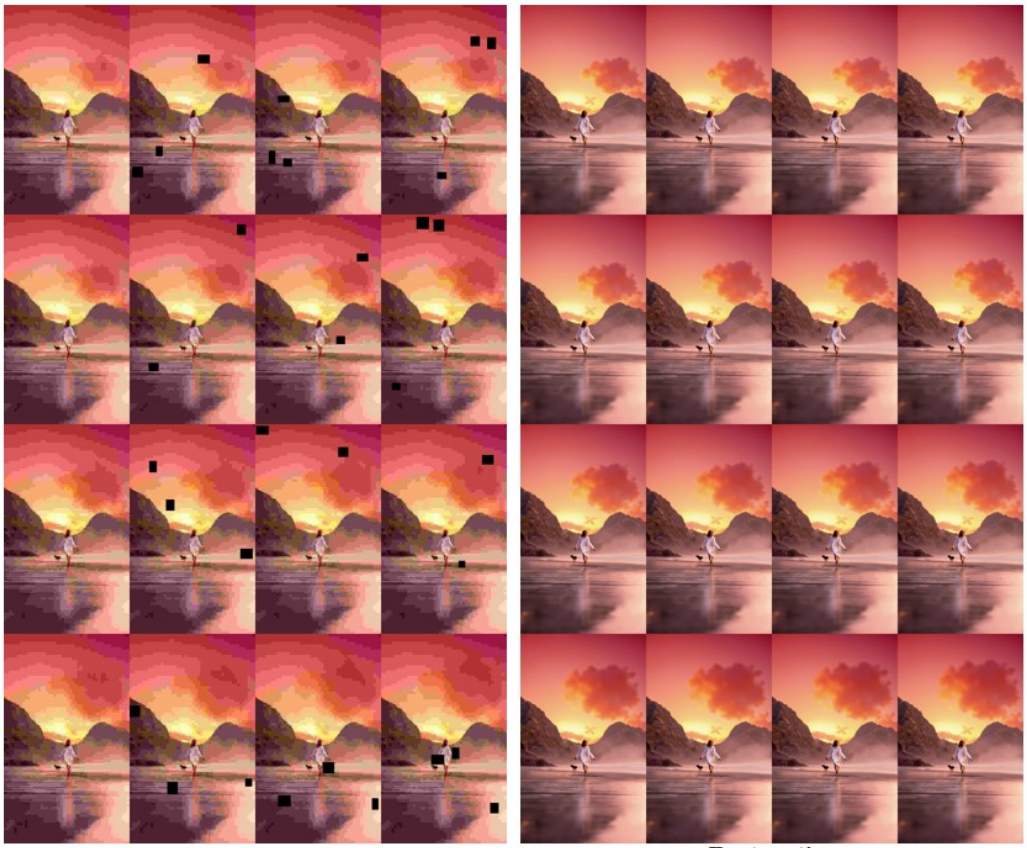

Figure 13: Video restoration from degraded inputs. Left: Input sequences with Gaussian blur and block masking. Right: Restored high-quality outputs maintaining temporal consistency.

Figure 14: Demonstration of maintained image generation capabilities. Our model preserves high-quality single-image generation performance across diverse scenarios including: basic objects, nature scenes, character interactions, indoor/outdoor environments, artistic styles, and lighting effects. Each image is generated from text prompts testing different aspects of the model's generation abilities.