Image as a World: Generating Interactive World from Single Image via Panoramic Video Generation

Anonymous Author(s)
Affiliation
Address
email

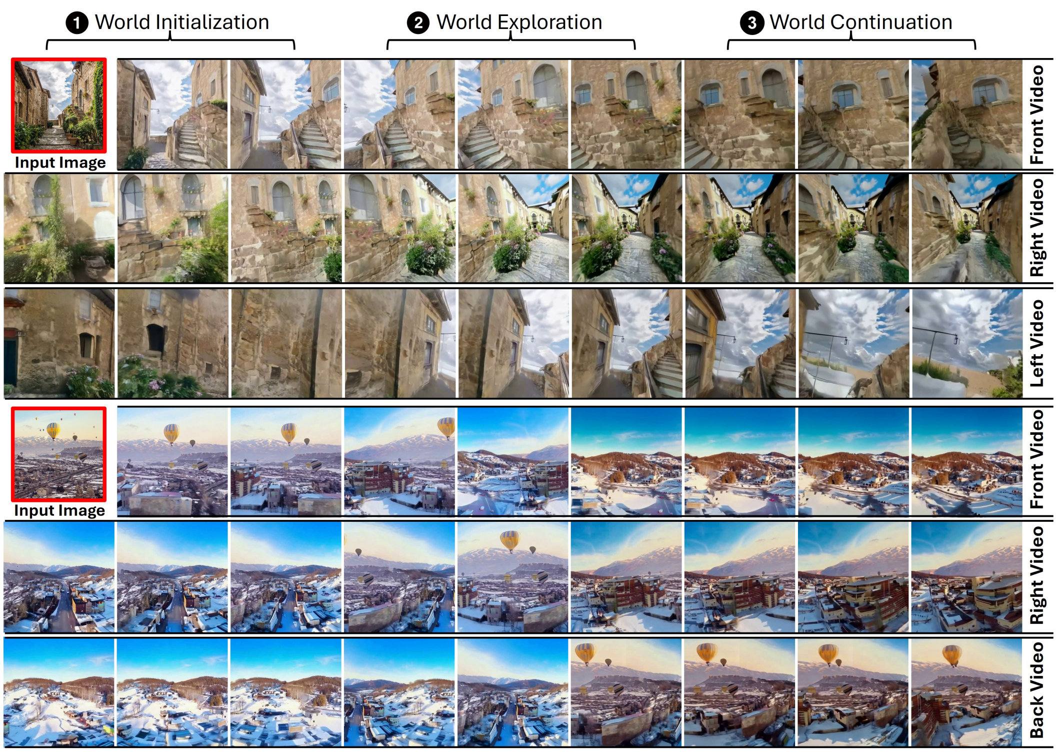

Figure 1: Generated world by our IaaW. Each line represents a fixed-view video that includes our proposed three stage of visual world generation: initialization, exploration, and continuation.

Abstract

Generating an interactive visual world from a single image is both challenging and practically valuable, as single-view inputs are easy to acquire and align well with prompt-driven applications such as gaming and virtual reality. This paper introduces a novel unified framework, Image as a World (**IaaW**), which synthesizes high-quality 360-degree videos from a single image that are both controllable and temporally continuable. Our framework consists of three stages: world initialization, which jointly synthesizes spatially complete and temporally dynamic scenes from a single view; world exploration, which supports user-specified viewpoint rotation; and world continuation, which extends the generated scene forward in time with temporal consistency. To support this pipeline, we design a visual world model based on generative diffusion models modulated with spherical 3D positional encoding and multi-view composition to represent geometry and view semantics. Additionally, a vision-language model (IaaW-VLM) is fine-tuned to produce both global and view-specific prompts, improving semantic alignment and controllability. Extensive experiments demonstrate that our method produces panoramic videos with superior visual quality, minimal distortion and seamless continuation in both qualitative and quantitative evaluations. To the best of our knowledge, this is the first work to generate a controllable, consistent, and temporally expandable 360-degree world from a single image.

1 Introduction

Recent advances in world models [11, 20] and video generation have enabled simulation and extension of environments in rich, multimodal ways. World models have evolved to handle raw visual inputs, producing videos conditioned on actions or inferred intent across domains such as robotics [39, 17], autonomous driving [44, 10], and interactive gaming [4]. Concurrently, large-scale diffusion models [16, 7, 26, 41] and vision transformers [25] have redefined the frontier of video generation, achieving high fidelity and temporal coherence across diverse conditions. These advancements collectively point toward a promising new direction: building dynamic, controllable, and immersive environments directly from visual cues.

In this work, we take a step further by proposing a novel problem: generating an explodable and temporally extendable panoramic world from a single image—one that not only predicts future frames, but also supports interactive viewpoint control, enabling arbitrary view rotations and continuous scene evolution. Compared to prior methods that rely on multi-view or panoramic input, our single-image setup significantly reduces the cost of data acquisition and aligns with the growing trend of prompt-based generative models. Unlike traditional world models that focus on action-conditioned prediction, our approach synthesizes immersive scenes that respond to user-specified actions, enabling both free-form viewpoint control and continuous scene expansion.

Generating such a visual world from single image is both practically appealing and technically challenging. It requires the model to infer latent geometry, spatial layout, and temporally coherent dynamics from highly limited visual evidence—an under-constrained and ill-posed task. We formulate this as a new direction in panoramic video generation, where the goal is to synthesize panoramic, navigable, and temporally extensible video from minimal visual input.

Existing methods are not designed for this setting. Many prior approaches to panoramic video generation adopt one-shot generation strategies without temporal continuity or interaction capability. For instance, 360DVD [36] relies on text-to-video models with limited resolution, while 4K4DGen [21] generates each frame independently without temporal coherence. Others assume richer input such as multi-view videos [40, 23] or full panoramic images [19, 22].

To address these challenges, we structure our solution as a three-stage generative pipeline: (1) World Initialization, which synthesizes a spatially complete and temporally coherent panoramic video from a single image, which provides stable spatiotemporal foundation for the subsequent stages; (2) World Exploration, which enables interactive navigation by modeling viewpoint changes as actions, thereby embedding user control directly into the generation process; and (3) World Continuation, which extends the scene forward in time while maintaining temporal consistency beyond a fixed horizon. Each stage addresses a specific limitation in prior work, and they collectively enable consistent, controllable and infinitely extensible world synthesis. To support this pipeline, we design a *visual world model*, implemented by augmenting a diffusion-based video generator with 3D Spherical Rotary Positional Encoding (RoPE) and multi-view composition. These components equip the model with the ability to represent scene geometry and maintain diversity across dynamic panoramic sequences. To enhance controllability and prompt alignment, we also finetune a vision-language model (IaaW-VLM) that generates semantically grounded, view-specific prompts conditioned on the user's perspective. Comprehensive experiments demonstrate that IaaW is capable of generating high-fidelity, semantically plausible panoramic videos that are both spatially coherent and temporally smooth. To our knowledge, this is the first framework to achieve infinitely expandable, user-controllable panoramic world synthesis from a single image.

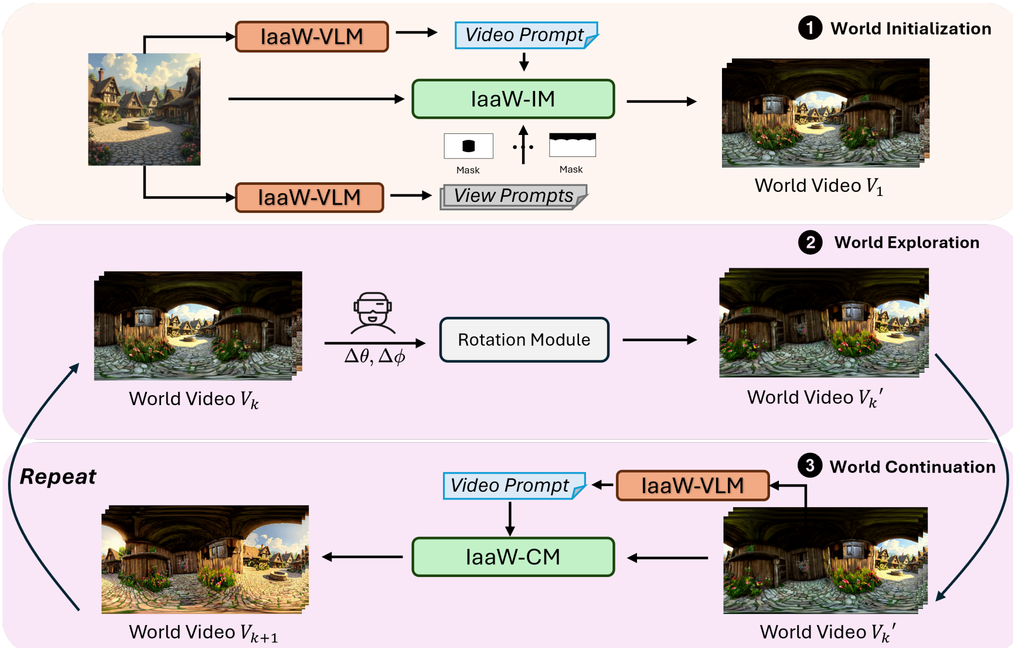

Figure 2: Pipeline of our proposed IaaW method, which consists of three core stages: world initialization, world exploration, and world continuation. In the world initialization stage, given a single reference image, we employ our finetuned IaaW-VLM to produce a holistic video-level prompt alongside multiple view-specific prompts. These, in conjunction with the input image, are processed by our IaaW-InitialModel (IaaW-IM) to generate the initial world video V_1. World exploration stage enables user's spatial control over the generated scene, the rotation module transforms the video V_k to reflect the desired viewpoint video V_k'. In the final world continuation stage, the rotated video and its associated prompt are fed into the IaaW-ContinualModel (IaaW-CM), which produces an extended segment of the world. This process is inherently recursive, allowing the newly generated video to undergo further view rotations and extensions.

2 Related Work

2.1 World Model

World models [11, 20] aim to predict the future evolution of an environment in response to specific actions. Traditionally, these models operated in abstract spaces and were predominantly used for planning [13, 29, 28] or policy learning [12] in reinforcement learning contexts. Recent advances in generative modeling have extended world models to the visual domain, enabling video generation conditioned on control inputs [45]. In autonomous driving [44, 10], models predict based on driver actions, while in robotics [39, 17], predictions are conditioned on control signals of robots. Genie [4] further generalizes this by learning action-conditioned dynamics from raw gameplay videos in an unsupervised manner. In contrast, we conceptualize a panoramic video generative model as a world model. Leveraging the interactive nature of the panoramic videos, we treat user-specified view rotations as world actions and generate the corresponding next-step visual evolution, enabling immersive, controllable, and continuable world generation.

2.2 Video Generation

Recent advancements in diffusion models [31, 16, 32, 7, 26, 24] have propelled video generation, with hierarchical U-Net [27] and diffusion vision transformer (DiT) [25] architectures leading the way in spatiotemporal modeling. Approaches like Imagen Video [15] and Make-A-Video [30] extend these models with temporal attention, while Sora [3] scales diffusion transformers for video synthesis. In the open-source landscape, CogVideoX [41] introduces a 3D causal VAE with adaptive LayerNorm for efficient spatiotemporal modeling, and Hunyuan Video [37] employs a dual-stream transformer for enhanced text-video alignment. Wan [35] addresses high-resolution generation with dynamic 3D-VAE compression. These models highlight the growing trend toward hybrid architectures and scaling strategies for balancing fidelity and efficiency.

2.3 Panoramic Video Generation

Recent advances in panoramic video generation explore diverse paradigms [36, 23, 19]. 360DVD [36] is the first to tackle text-to-panoramic video synthesis by integrating a 360-adapter into early-stage T2V pipelines. However, it is limited by low-resolution training data and underpowered base models, yielding suboptimal quality. 4K4DGen [21] proposes a training-free approach for animating 4K panoramic images by independently rendering perspective views and spatially fusing them, while OmniDrag [22] enables interactive control via drag-based motion manipulation. DynamicScaler [19] enhances spatial scalability using an offset-shifting denoiser to synthesize spherical panoramas, followed by a learned upscaling stage. Several works address panoramic video generation by outpainting from nFOV inputs. VideoPanda [40] introduces multi-view attention to maintain spatiotemporal consistency, whereas [23] reframes the task as video-to-video generation. In driving applications, Panacea [38] leverages BEV representations for conditional panoramic synthesis. In contrast, our method directly generates high-fidelity, temporally coherent and continuable panoramic videos from a single image, achieving both spatial diversity and temporal infinity without auxiliary inputs.

3 Method

3.1 Visual World Model

Our visual world model is built on a diffusion-based video backbone, enhanced with multi-view composition and 3D Spherical RoPE. To support different generation goals, we introduce two finetuned variants: IaaW-InitialModel (IaaW-IM) for world initialization and IaaW-ContinualModel (IaaW-CM) for world continuation. While sharing the same architecture, the two models are optimized for different stages, which are scene reconstruction from a single image vs. temporally coherent extension.

3.1.1 Multi-View Composition

To address the limitations of one-shot conditioning in existing video generation models, we propose a multi-view composition method that significantly enhances the quality and diversity in world initialization. As depicted in Fig. 3, our method begins with a single reference image, a corresponding video prompt, and a set of auxiliary view prompts, each of which is paired with spatially aligned masks. These view prompts are semantically and spatially diverse renderings of the scene, intended to provide additional geometric and contextual priors.

Building on recent powerful video generative models, such as CogVideoX [41], which uses MM-DiT blocks [8] and concatenate textual (z_{text}) and visual (z_{vision}) features, we introduce a multi-view conditioning mechanism for improved world initialization. Here, z_{text} comes from the main video prompt, and z_{vision} encodes a padded reference image with noisy frames. We add a parallel attention path using view-aware features: $z_{\text{text}}^{\text{view}}$ from IaaW-VLM is concatenated with z_{vision} and modulated by view masks for localized 3D full attention. This stream runs in parallel with the base attention and is adaptively

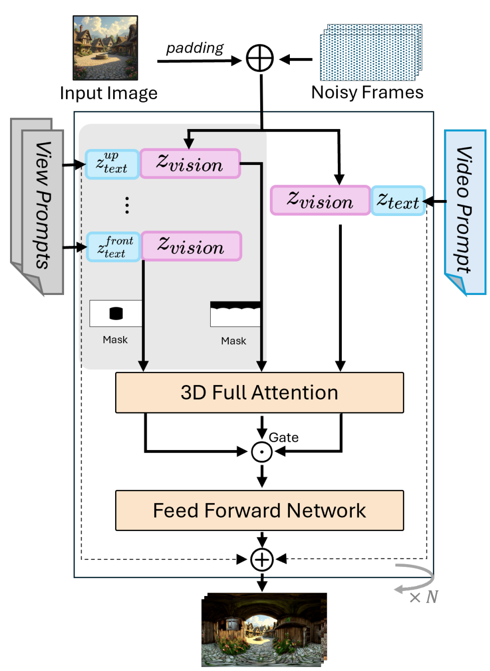

Figure 3: Multi-view composition used in IaaW-IM's MM-DiT blocks in world initialization.

gated to fuse multi-view cues with global context. For clarity, AdaLN and scale-and-shift components are omitted from Fig. 3.

3.1.2 3D Spherical RoPE

We propose a unified 3D Spherical Rotary Positional Encoding (RoPE) that extends traditional rotary embeddings [33, 41] to spherical video domains. By embedding positional information in both spherical space [6, 42, 43] and time, our method aligns with the geometric structure of equirectangular panoramic video while preserving the rotation-equivariant properties of RoPE.

Let a video $V \in \mathbb{R}^{H\times W\times D\times T}$ represent a sequence of frames with height H, width W, feature dimension D, and temporal length T. Each spatial coordinate (x, y) is mapped to spherical angles via:

$$\theta = \frac{\pi}{2}\left(\frac{2y}{H} - 1\right), \quad \phi = \pi\left(\frac{2x}{W} - 1\right), \tag{1}$$

where θ and ϕ denote latitude and longitude, respectively. We then construct a unified 3D positional encoding by modulating angular and temporal components in a factorized trigonometric basis:

$$\mathrm{RoPE}_{x,y,t,d} = [\cos(2^d\theta)\cdot\cos(2^d\phi)\cdot\cos(2^d\cdot 2\pi t), \sin(2^d\theta)\cdot\cos(2^d\phi)\cdot\cos(2^d\cdot 2\pi t), \ldots] \tag{2}$$

which compactly encodes the 3D positional across spatial angles (θ, ϕ), frequency d and normalized time t. 3D Spherical RoPE captures rotational symmetries on the spherical surface while enabling temporal phase alignment, resulting in a compact and geometry-aware encoding mechanism for panoramic video generation.

3.2 IaaW Pipeline

3.2.1 World Initialization

World Initialization serves as the entry point for visual world synthesis, which establishes the spatiotemporal foundation for subsequent user-controlled exploration and continuation. Given only a single-view image, the model must generate an initial panoramic video clip that is both spatially complete and semantically coherent, despite the severe ambiguity posed by missing multi-view context.

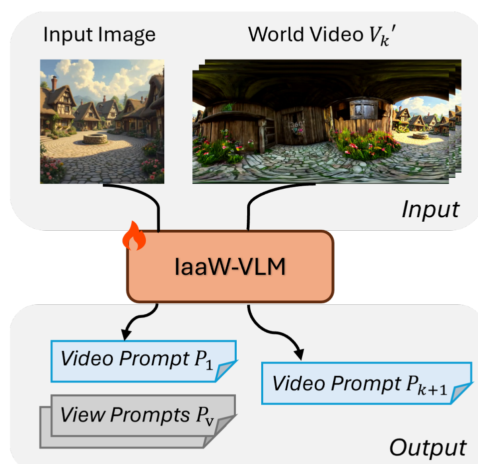

Figure 4: Functionality of IaaW-VLM.

To enhance semantic suitability and consistency in video generation, we introduce a world context model IaaW-VLM that generates both global and view-specific prompts. For each equirectangular video V, we first employ a caption model for a global prompt P that summarizes the entire scene. The video is then spatially segmented into multiple views $\{V_v\}$ and individually captioned to yield prompts $\{P_v\}$, capturing the localized context. From each V_v, we extract a representative frame I_v, which forms the dataset $\{V, P, \{V_v, P_v\}, \{I_v\}\}$. This corpus supports training for IaaW-VLM, whose functionality is shown in Fig. 4. IaaW-VLM can generate $\{P_v\}, P$ from single view image I_v and P from video V, which supports IaaW-IM and IaaW-CM separately. By grounding generation in this multi-granular context, IaaW-VLM acquires an enriched understanding of both spatial structure and temporal coherence.

With visual world model IaaW-IM and above IaaW-VLM, as shown in Fig. 2, we send entire video prompt and view prompts with corresponding masks into IaaW-IM model. This process generates the world video as

$$V_1 = \mathcal{IM}(I, P_1, \{M_v, P_v, v \in \text{views}\}) \tag{3}$$

where P_1 and P_v represent initial prompt and view prompts respectively, M_v represents masks.

3.2.2 World Exploration

Panoramic video enables immersive navigation by allowing users to rotate their virtual viewpoint within a spherical environment. We model this interaction as a transformation in spherical coordinates applied to the kth equirectangular video $V_k \in \mathbb{R}^{H\times W\times D\times T}$, where $W = 2H$, and D, T denote the channel and temporal dimensions. Each pixel $(x, y) \in [0, W) \times [0, H)$ corresponds to spherical coordinates (θ, ϕ) following Eq. (1). These angles represent latitude $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ and longitude $\phi \in [-\pi, \pi)$. User-specified pitch and yaw rotations $(\Delta\theta, \Delta\phi) \in \mathbb{R}^2$ simulate view changes by adjusting the angles:

$$\theta' = \mathrm{clip}(\theta + \Delta\theta, -\tfrac{\pi}{2}, \tfrac{\pi}{2}), \quad \phi' = \phi + \Delta\phi \tag{4}$$

Here, clip ensures that the elevation stays within the bounds of the spherical domain. To map back to image coordinates, we have

$$x' = \left(\frac{\phi'}{2\pi} + \frac{1}{2}\right) W \bmod W, \quad y' = \left(\frac{\theta'}{\pi} + \frac{1}{2}\right) H \tag{5}$$

The complete process yields the rotated video $V'_k \in \mathbb{R}^{H\times W\times D\times T}$, which is obtained by sampling the original video at $V_k(x', y', :, t)$ for each (x, y, t).

3.2.3 World Continuation

View-aware world continuation stage enables the synthesis of temporally extended and visually coherent video sequences conditioned on a user-defined reference view. Our approach is built upon the visual world model IaaW-CM, which operates in an autoregressive manner, progressively generating video segments while maintaining view and content consistency over time in Eq. (6).

$$V_{k+1} = \mathcal{CM}(V'_k, P_{k+1}) \qquad k = 2, \ldots, n \tag{6}$$

Specifically, following the paradigm of IaaW-IM, we substitute the single view image with video V'_k from previously rotated video chunk. At step k, the IaaW-VLM produces the next prompt P_{k+1} based on the evolving visual context, guiding the generation of segment V_{k+1} towards arbitrary length. This stage establishes a foundation for open-ended and infinite scene generation, where a coherent and semantically meaningful world can emerge over extended temporal horizons, grounded in a user-defined viewpoint trajectory.

4 Experiments

4.1 Experimental Setup

Models In the field of video generation, there are few open-source video diffusion models available for experimentation. We use CogVideoX1.5-5B-I2V [41], a text-image conditional video generator that supports arbitrary resolution and is well suited to our 2:1 aspect-ratio video setup. We use equirectangular videos to finetune IaaW-IM, where the input image is padded before being fed into the model. IaaW-CM is finetuned on top of IaaW-IM, using the previous video chunk as input and the next video chunk as output. Finetuning is conducted over two weeks on 4×A100 GPUs, followed by one week of progressive finetuning. Due to the absence of released code from prior panoramic methods [19, 23, 21], we implemented two baselines for comparison. One is 360I2V, a panoramic animation baseline fine-tuned from CogVideoX, which takes panoramic image as input to generate panoramic videos. Another is FETA (First Expand, Then Animate), a two-stage baseline for world initialization, where we combine Diffusion360 [9] for NFoV-to-panorama expansion with 360I2V for subsequent animation. We compare our world initialization results with FETA, compare the world continuation results with 360I2V, and compare our whole IaaW pipeline with FETA+360I2V.

Data We consider several panoramic video datasets, including WEB360 [36] and 360-1M from ODIN [34]. Due to WEB360's limited scale and low resolution (2K videos at 1024×512), it is excluded from our study. From 360-1M, we curate a high-resolution, equirectangular subset by filtering out static scenes and selecting diverse, dynamic content. Captions are generated using Qwen-VL-2.5 [2], and low-quality samples are removed based on caption quality. Augmented with an internal collection, our final dataset comprises 120K videos at 2048×1024 resolution. An 8K high-quality subset is further collected for progressive finetuning.

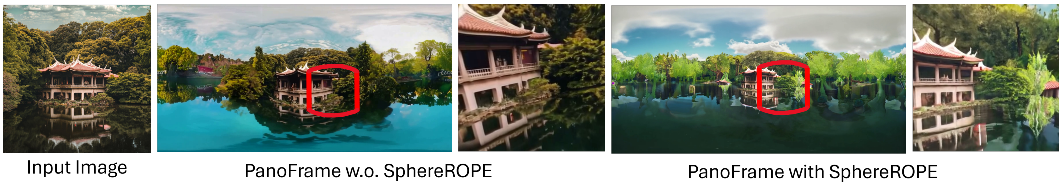

Figure 5: The results of reducing distortions of 3D Spherical RoPE in world initialization.

Figure 6: Generation results of multi-view composition for world initialization.

Metrics To evaluate video generation quality, we consider both overall and per-view fidelity and consistency using metrics from VBench [18] and VideoBench [14]. *Subject Consistency* measures temporal coherence via the average cosine similarity of DINO [5] features between each frame and the first. *Motion Smoothness* is quantified by the mean absolute error between interpolated and dropped frames, while *Aesthetic Quality* is predicted using the LAION aesthetic model [1]. *Video-Text Consistency* assesses semantic alignment with the prompt, computed as the average score (1–5) assigned by a vision-language model. To evaluate continuous generation results, we concatenate videos from preceding steps, rotational transitions, and subsequent generations to evaluate coherence over extended sequences.

4.2 Qualitative Analysis

World Initialization Results We first demonstrate the effectiveness of our 3D spherical RoPE in Fig. 5. The figure compares panoramic frame produced by our IaaW-IM. When rendering the panoramic image from a specified viewpoint, the model with spherical RoPE exhibits fewer distortions. In particular, it preserves the correct perspective geometry of structures such as the pavilion, whereas the model without that yields deformed objects with incorrect perspective relationships.

To assess the impact of multi-view composition in our initialization model IaaW-IM, we visualize generation results under varying view prompts in Fig. 6. Using a fixed video prompt and identical spatial masks, we observe that distinct view prompts (e.g., wooden buildings vs. flower yards) yield semantically diverse scene expansions. This demonstrates the fine-grained controllability afforded by multi-view composition mechanisms and underscores their ability to guide content-specific scene synthesis during world initialization.

In Fig. 7, we compare our initialization strategy with a baseline that first performs panoramic extrapolation then animates the results in two separate stages. This decoupled spatial-temporal generation often leads to pronounced spatial artifacts and temporal discontinuities. In contrast, our method jointly models spatial structure and temporal dynamics and delivers coherent expansions that maintain global scene structure while enabling temporally smooth motion, establishing a superior world base.

World Continuation Results We evaluate the continuation model IaaW-CM in Fig. 8, where the initialized world is an aerial view towards a lighthouse. Our model maintains directional consistency across extended sequences after rotational transformations. Specifically, our generated continuation

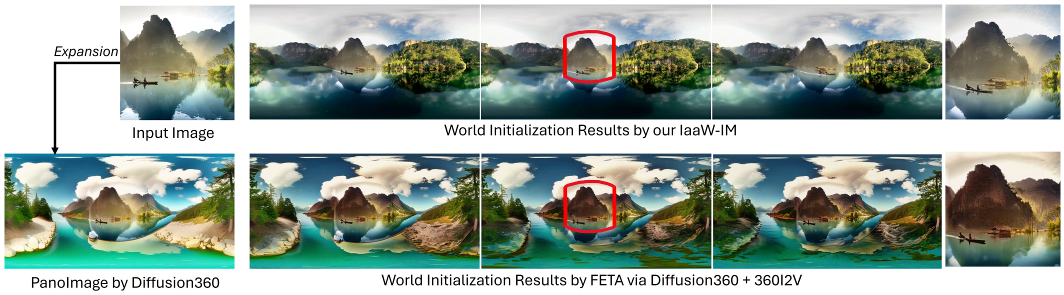

Figure 7: World initialization results compared with First Expanding Then Animating(FETA).

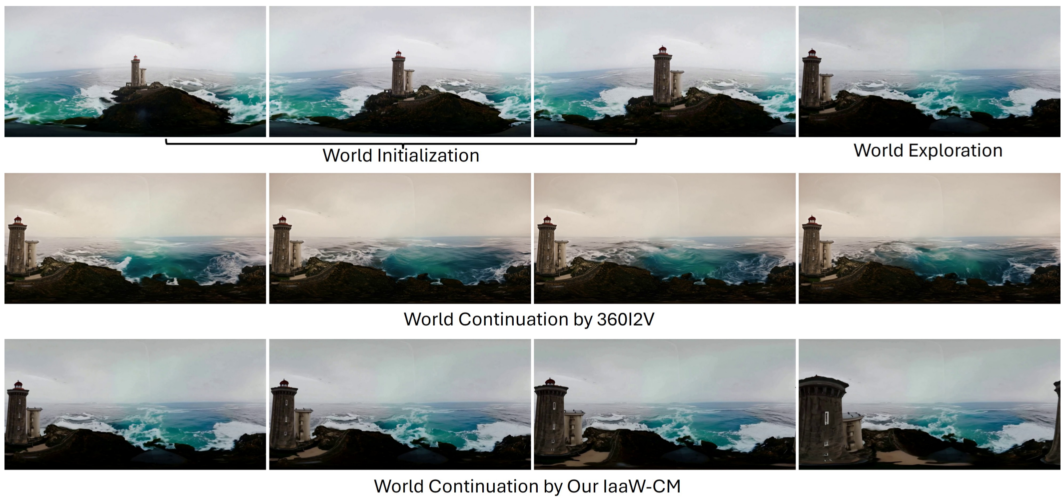

Figure 8: World continuation results of our IaaW-CM compared with baseline 360I2V.

video persistently advances toward the lighthouse while remaining both temporally stable and spatially coherent. In comparison, the baseline, which conditioned solely on the last frame, suffers from abrupt motion discontinuities and visual degradation, and fails to preserve global motion dynamics. These findings highlight the efficacy of our IaaW-CM in capturing long-consistent motion trajectories.

Overall, our IaaW framework demonstrates qualitatively superior results in both initialization and continuation stages, affirming its effectiveness in generating visually coherent, controllable, and temporally consistent panoramic video worlds.

4.3 Quantitative Analysis

We present the quantitative results in Table 1, where we evaluate the videos in three setups: world initialization videos, world continuation videos, and entire world videos. The latter refers to the concatenated video of initialization video, world exploration video, and world continuation video.

Our IaaW-IM outperforms the baseline FETA across most metrics, demonstrating superior spatial-temporal quality. Temporal metrics averaged across views are higher than overall due to motion discontinuities introduced by splits in the equirectangular format. Aesthetic quality is lower when averaged per view, as certain angles (e.g., top, bottom) naturally lack visual appeal (e.g., sky, floor). VTC-View scores are lower than VTC-All because some view-specific videos inadequately capture the full prompt, reducing alignment.

In continuation model comparisons, our IaaW-IM outperforms the baseline 360I2V across most metrics, indicating stronger temporal modeling. Temporal and spatial trends mirror those in initialization models, but SC-View is lower than SC-All due to reduced uncertainty when the full panoimage is available. Temporal results surpass those of initialization models as full panoramic input offers richer context than single-view inputs. Slightly lower spatial scores stem from decreased diversity and aesthetic richness when multi-view information is provided.

Model	SC-View	SC-All	MS-View	MS-All	AQ-View	AQ-All	VTC-View	VTC-All
FETA	89.4	86.8	98.1	98.3	49.9	56.7	3.19	3.93
IaaW-IM	91.8	88.2	99.0	98.9	55.9	59.8	3.72	4.00
360I2V	92.5	94.8	98.9	98.7	49.5	55.0	3.25	3.89
IaaW-CM	95.8	97.2	99.3	99.2	49.7	55.7	3.26	3.90
FETA+360I2V	81.0	88.7	98.8	98.7	50.1	55.9	3.39	3.93
IaaW-IM+CM	91.0	90.3	99.1	99.1	50.5	57.5	3.50	3.94

Table 1: Analysis of video generation results of our method and several baselines. SC, MS, AQ and VTC represent subject consistency, motion smoothness, aesthetic quality, and video-text consistency respectively, and for all of these metrics, higher scores are better. Postfix "View" means the numbers are calculated across different views and "All" means the numbers are calculated as a whole.

Model	SC-View	SC-All	MS-View	MS-All	AQ-View	AQ-All	VTC-View	VTC-All
IaaW-IM	91.8	88.2	99.0	98.9	55.9	59.8	3.72	4.00
IaaW-IM *w.o.* 3D SphereRoPE	86.3	83.7	98.1	97.9	48.8	56.8	3.17	3.97
IaaW-IM *w.o.* MultiViewComp	91.2	86.6	99.0	98.9	49.5	59.9	3.24	3.95

Table 2: Ablation study on our world initialization model IaaW-IM.

For whole-process comparison, our IaaW-IM+CM surpasses the baseline FETA+360I2V across all metrics, demonstrating enhanced temporal consistency and spatial quality. By effectively integrating initialization and continuation models, our pipeline generates visually consistent results, whereas the baseline exhibits fragmentation between two stages, leading to inferior performance.

4.4 Ablation Study

We conduct an ablation study on world initialization components in Table 2. Removing the 3D Spherical RoPE consistently degrades performance both in spatial and temporal metrics, especially for view-based metrics. This degradation is primarily due to spatial distortions, resulting in unsmooth motion and scene deformation. Excluding the Multi-View Composition module reduces VTC-View and AQ-View, as it limits the model's ability to capture view-specific textual cues and leads to a loss of visual quality in separate views. Temporal metrics remain relatively stable, since this module mainly enhances diversity rather than motion smoothness. The slight drop in VTC-All suggests the model still generates prompt-aligned content overall, as neither component directly influences overall textual understanding in video generation.

5 Limitations and Social Impact

IaaW excels at generating panoramic world from a single image but struggles with maintaining temporal consistency over very long durations. Specifically, IaaW-CM conditions on the most recent video chunk rather than the full video history, which can lead to a loss of coherence during super long-term video continuations. This long-term consistency presents a key challenge not only for IaaW but also for the broader field of video generation, which we leave as future work.

From a societal perspective, IaaW empowers content creation across VR/AR and gaming, potentially opening new avenues for immersive interactive experiences. However, it also introduces risks related to misinformation and visual deception, which may undermine trust in visual media. Implementing robust safeguards is essential to mitigate potential misuse and ensure responsible use.

6 Conclusion

We introduce Image as a World (IaaW), a novel framework for generating expandable, user-controllable panoramic world from a single image, which comprises three critical components: world initialization, world exploration, and world continuation. We design visual world models equipped with 3D spherical RoPE and multi-view composition, and two variants of which, IaaW-IM and IaaW-CM, tackle world initialization and continuation, respectively. Extensive experiments validate the effectiveness of our approach, demonstrating high fidelity, controllability, and scalability across diverse scenarios. Our work opens new potential for one-shot visual world generation in applications such as gaming and virtual reality, setting the stage for future research in generating interactive visual worlds.

References

[1] LAION AI. Laion-aesthetics.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[4] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[6] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022.

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[9] Mengyang Feng, Jinlin Liu, Miaomiao Cui, and Xuansong Xie. Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models. *arXiv preprint arXiv:2311.13141*, 2023.

[10] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024.

[11] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

[12] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[13] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

[14] Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li-jia Li, and Yongxin Ni. Video-bench: Human preference aligned video generation benchmark. *arXiv preprint arXiv:xxx*, 2024.

[15] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[17] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

[18] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[19] Liu Jinxiu, Lin Shaoheng, Li Yinxiao, and Yang Ming-Hsuan. Dynamicscaler: Seamless and scalable video generation for panoramic scenes, 2024.

[20] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

[21] Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, and Zhiwen Fan. 4k4dgen: Panoramic 4d generation at 4k resolution, 2024.

[22] Weiqi Li, Shijie Zhao, Chong Mou, Xuhan Sheng, Zhenyu Zhang, Qian Wang, Junlin Li, Li Zhang, and Jian Zhang. Omnidrag: Enabling motion control for omnidirectional image-to-video generation. *arXiv preprint arXiv:2412.09623*, 2024.

[23] Rundong Luo, Matthew Wallingford, Ali Farhadi, Noah Snavely, and Wei-Chiu Ma. Beyond the frame: Generating 360 {\deg} panoramic videos from perspective videos. *arXiv preprint arXiv:2504.07940*, 2025.

[24] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[25] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[28] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[29] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[30] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[31] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[33] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[34] Matthew Wallingford, Anand Bhattad, Aditya Kusupati, Vivek Ramanujan, Matt Deitke, Aniruddha Kembhavi, Roozbeh Mottaghi, Wei-Chiu Ma, and Ali Farhadi. From an image to a scene: Learning to imagine the world from a million 360° videos. *Advances in Neural Information Processing Systems*, 37:17743–17760, 2024.

[35] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[36] Qian Wang, Weiqi Li, Chong Mou, Xinhua Cheng, and Jian Zhang. 360dvd: Controllable panorama video generation with 360-degree video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6913–6923, 2024.

[37] Zijian Zhang Rox Min Zuozhuo Dai Jin Zhou Jiangfeng Xiong Xin Li Bo Wu Jianwei Zhang Kathrina Wu Qin Lin Aladdin Wang Andong Wang Changlin Li Duojun Huang Fang Yang Hao Tan Hongmei Wang Jacob Song Jiawang Bai Jianbing Wu Jinbao Xue Joey Wang Junkun Yuan Kai Wang Mengyang Liu Pengyu Li Shuai Li Weiyan Wang Wenqing Yu Xinchi Deng Yang Li Yanxin Long Yi Chen Yutao Cui Yuanbo Peng Zhentao Yu Zhiyu He Zhiyong Xu Zixiang Zhou Zunnan Xu Yangyu Tao Qinglin Lu Songtao Liu Dax Zhou Hongfa Wang Yong Yang Di Wang Yuhong Liu Weijie Kong, Qi Tian and along with Caesar Zhong Jie Jiang. Hunyuanvideo: A systematic framework for large video generative models, 2024.

[38] Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6902–6912, 2024.

[39] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pages 2226–2240. PMLR, 2023.

[40] Kevin Xie, Amirmojtaba Sabour, Jiahui Huang, Despoina Paschalidou, Greg Klar, Umar Iqbal, Sanja Fidler, and Xiaohui Zeng. Videopanda: Video panoramic diffusion with multi-view attention. *arXiv preprint arXiv:2504.11389*, 2025.

[41] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[42] Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360 panorama image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6347–6357, 2024.

[43] Dian Zheng, Cheng Zhang, Xiao-Ming Wu, Cao Li, Chengfei Lv, Jian-Fang Hu, and Wei-Shi Zheng. Panorama generation from nfov image done right. *arXiv preprint arXiv:2503.18420*, 2025.

[44] Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. In *European Conference on Computer Vision*, pages 87–104. Springer, 2024.

[45] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims stated in the abstract and introduction accurately reflect the paper's core contributions and scope. They are aligned with both the theoretical foundations and the experimental results presented in the main body.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper clearly outlines its limitations, including factors influencing the method's performance, demonstrating a responsible and transparent evaluation of its contributions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper's theoretical result provide the full set of assumptions but we do not propose any new theorems thus no proof needed.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper includes all necessary implementation details, dataset information, and evaluation protocols to ensure reproducibility of the key experimental findings supporting the main claims.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We are currently considering the possibility of open access to the data and code, but due to policy constraints, it is not yet feasible to make them publicly available. We will continue to evaluate this option as the project progresses.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All essential training and testing configurations—including data splits, hyperparameters, and optimization strategies—are thoroughly documented to support reproducibility and understanding.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports error bars and statistical significance metrics with clarity, detailing how they were calculated and what variability factors they represent.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Detailed information on hardware used, compute time, and resource requirements for experiments is provided, allowing for practical replication of the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research adheres to the NeurIPS Code of Ethics in all respects, ensuring ethical standards are met throughout the study.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Both the potential benefits and risks of the proposed work are discussed, including societal impacts and possible unintended consequences, along with considerations for mitigation.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The paper outlines measures to prevent misuse in section about social impacts.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All third-party assets are properly cited with explicit mention of licenses and terms of use, demonstrating respect for intellectual property.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets released as part of the paper are accompanied by comprehensive documentation detailing usage, licensing, and limitations.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve any form of crowdsourcing or research involving human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: The study does not involve human subjects and thus does not require IRB or equivalent review.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper clearly describes how LLMs are used as a key, non-standard component in the proposed methodology.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.