SHORT-PAPER

ACM Multimedia 2025 Grand Challenge report for Image-to-Video Generation Model Acceleration

JIE YANG, Hunan University, Changsha, Hunan, China

SHIEN SONG

JIN CHEN

HAOYUAN XIE

HAN QI

YIFEI XUE, Hunan University, Changsha, Hunan, China

View all

Published: 27 October 2025

Citation in BibTeX format

MM '25: The 33rd ACM International Conference on Multimedia
October 27 - 31, 2025
Dublin, Ireland

Conference Sponsors:
SIGMM

MM '25: Proceedings of the 33rd ACM International Conference on Multimedia (October 2025)
https://doi.org/10.1145/3746027.3761991
ISBN: 9798400720352

ACM Multimedia 2025 Grand Challenge report for Image-to-Video Generation Model Acceleration

Jie Yang
Hunan University
Changsha, China
MGTV
Changsha, China
yangjie@mgtv.com

Shien Song
MGTV
Changsha, China
shien@mgtv.com

Jin Chen
MGTV
Beijing, China
chenjin@mgtv.com

Haoyuan Xie
MGTV
Changsha, China
xiehaoyuan@mgtv.com

Han Qi
MGTV
Changsha, China
qihan@mgtv.com

Yifei Xue
Hunan University
Changsha, China
iflyhsueh@gmail.com

Yizhen Lao*
Hunan University
Changsha, China
yizhenlao@hnu.edu.cn

Abstract

Recently, MGTV organized the Image-to-Video Model Acceleration Challenge, calling for participants to propose optimization solutions for the Wan 2.1-14B model[11]. The challenge emphasizes techniques such as quantization and GPU acceleration to improve the model's inference efficiency.

As AIGC technology advances rapidly, video generation large models exhibit great potential in content creation, yet they face critical challenges of high computing power consumption, long inference time, and excessive VRAM usage during inference, which severely hinder content production efficiency. This challenge aims to explore approaches for efficient video generation under limited computing resources, requiring participants to reduce the model's computing power and VRAM demands while improving inference speed, all without compromising generation quality.

To support participants' development and evaluation, the challenge provides a baseline framework and test dataset. For further details, please refer to the official challenge website (https://challenge.ai.mgtv.com/#/track/53).

CCS Concepts

• **General and reference → General conference proceedings**.

*Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3761991

Keywords

Artificial Intelligence Generated Content, Image-to-Video Generation, Model Acceleration

ACM Reference Format:
Jie Yang, Shien Song, Jin Chen, Haoyuan Xie, Han Qi, Yifei Xue, and Yizhen Lao. 2025. ACM Multimedia 2025 Grand Challenge report for Image-to-Video Generation Model Acceleration. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3746027.3761991

1 Introduction

Image-to-Video (I2V) generation[3] is a technology that takes one or multiple still images and automatically synthesizes a video sequence that is temporally coherent and semantically consistent. It preserves the identity of the original frame while adding realistic motion. With the explosive growth of Artificial Intelligence Generated Content (AIGC), I2V has transitioned from laboratory demonstrations to production pipelines. Netflix and Disney utilize it for rapid trailer iteration and scene previsualization. TikTok and Instagram Reels use it to power viral AR effects and personalized filters. Amazon, Taobao, and Shopify leverage it to create 360° product demos from a single photo. Game studios integrate it into virtual production stages to instantly animate concept art, significantly shortening creative iteration cycles.

Model acceleration refers to a category of optimization techniques covering algorithmic and system-level approaches. Its core goal is to reduce the computational overhead of deep neural networks (including floating-point operations (FLOPs), memory footprint, and energy consumption) while avoiding catastrophic degradation in video perceptual quality (e.g., clarity, color) or downstream performance (e.g., motion coherence, semantic consistency).

For diffusion-based I2V models, the necessity of acceleration is even more prominent. These models require hundreds of denoising iterations to generate videos and must handle temporal dependencies across multiple frames (such as object motion trajectories and continuity of light and shadow changes). As a result, the computational load of generating a 16-frame 1080p video is often more than 20 times that of generating a single image. Without optimization, real-time deployment or low-cost applications would be unfeasible. Therefore, I2V acceleration not only needs to reduce the computational cost per frame but also requires designing strategies to address "temporal dimension redundancy." It aims to compress generation latency while ensuring video semantic coherence (e.g., a cat does not turn into a dog due to acceleration) and smooth motion (e.g., no lag when a person waves their hand). Some common approaches include:

- Weight Quantization[8]: Weight quantization reduces computational overhead by lowering the numerical precision of model weights, striking a balance between efficiency and video quality in I2V generation. Post-Training Quantization (PTQ) converts 32-bit floating-point (FP32) weights in spatial feature layers (e.g., ConvNet backbones) to 8-bit integers (INT8), cutting computational load by 75% while preserving motion coherence through static calibration of temporal transformer layers.
- TeaCache[7]: TeaCache (Timestep Embedding Aware Cache) is a training-free caching strategy that optimizes inference efficiency in video diffusion models by dynamically selecting which intermediate outputs to cache based on timestep characteristics. Unlike uniform caching, it leverages timestep embeddings to estimate output differences across denoising steps, ensuring efficient reuse of redundant computations. It achieves up to 4.41x acceleration over baseline methods (e.g., Open-Sora-Plan) with negligible visual quality degradation (-0.07% VBench score[13]), making it effective for real-time I2V deployment.
- Efficient Attention Implementations[15] [14]: For example, Sage Attention is an optimized attention mechanism designed specifically for video generation in frameworks like ComfyUI, and it particularly enhances the efficiency of I2V models such as WanVideo[11] and Cosmos by reducing the computational burden of attention layers through multiple approaches. This mechanism leverages the Triton library for optimized kernel execution, significantly accelerating attention computations in long video sequences. It also integrates seamlessly with torch compile and prompt enhancement pipelines to reduce the generation time of high-resolution I2V tasks.
- Operator Fusion and Compiler Optimization[12]: This method enhances the efficiency of I2V models by integrating sequential operations and leveraging advanced compilers to reduce overhead. Operator fusion combines contiguous spatiotemporal operations into a single composite kernel, reducing the number of kernel launches by 60% and memory access latency by 45%. Compiler-level optimizations use tensor compilers such as TensorRT, to optimize computation graphs. By adjusting convolution order and eliminating redundant operations, they boost the inference speed of 4K I2V generation by 2.2x compared to unoptimized implementations.

2 Challenge Description

We aim to address the challenge of accelerating Image-to-Video generation models in this competition. Participants are tasked with optimizing model performance to achieve faster inference while maintaining high-quality video output. The following resources and requirements are provided to facilitate this challenge:

- High-quality image prompt pairs to evaluate the optimized models;
- A collection of publicly available datasets, such as Panda-70M[1] and Miradata[4], for training and development purposes;
- A comprehensive set of baseline models and code for data processing and evaluation, enabling participants to focus on model acceleration techniques;
- A detailed set of metrics for assessing the efficiency, speed, and visual quality of the optimized models;
- A necessary requirement for inference time and video quality to ensure the practical applicability of the solutions.

2.1 Problem Definition

The central objective of this challenge is to accelerate the inference process of high-quality Image-to-Video generation models through quantization and other optimization techniques. Existing models face significant challenges in terms of computational resource consumption and long inference times, which severely impact the efficiency of content production. To address these challenges, this competition focuses on the research and development of quantization algorithms for video generation models, aiming to achieve efficient video generation with limited computational resources.

The key challenge lies in balancing the model's efficiency, quality, and robustness. The goal is to significantly reduce the computational demand and memory usage during inference while maintaining the high quality of the generated videos. This requires careful consideration of factors such as the model's capacity to generate detailed and coherent videos, its robustness against various optimization techniques, and the overall computational efficiency to enable fast and real-time processing.

For this competition, the Wan2.1-14B model[11] is chosen as the benchmark due to its status as one of the best open-source Image-to-Video models with a thriving community ecosystem. Participants are required to optimize this model through techniques such as quantization and GPU acceleration. The ultimate goal is to achieve a substantial reduction in computational requirements and memory usage while preserving the model's ability to generate high-quality, coherent videos.

2.2 Dataset

Existing open-source benchmarks such as VBench[13] already define standard splits and metrics for Image-to-Video evaluation. We therefore encourage participants to continue testing on these public corpora, including Panda-70M[1] and Miradata[4]. To provide a rigorous test bed for this competition, we release MGTV-I2V, a dedicated evaluation set of 150 high-resolution image-prompt pairs. Each pair originates from a quality-screened video whose first frame is extracted as the input image, while the accompanying prompt is first drafted by GPT-4o[9] from a rich video description and then polished by human reviewers for accuracy and clarity. The videos span a broad spectrum of motion magnitudes, from subtle facial shifts to sweeping camera moves, ensuring that the benchmark can reliably assess model fidelity, inference speed, and memory footprint under diverse real-world conditions.

3 Challenge Evaluation

3.1 Data Setting

The challenge data are divided into two disjoint portions. Fifty samples form the public A-board set released at the start of the contest, allowing teams to iterate and self-check their solutions. One hundred distinct samples make up the B-board set, which is used exclusively for the final ranking. Official scores and prizes are determined solely by performance on the B-board.

3.2 Evaluation Environment

To guarantee fair comparison and reproducibility every submission is executed in an identical containerized environment. The host offers a 16-core CPU and enforces a 64 GB RAM ceiling while the accelerator is locked to a single NVIDIA GeForce RTX 4090 with 24 GB of VRAM. This deliberately modest yet powerful setup is chosen for two reasons. First, a uniform configuration forces teams to confront the same compute budget, ensuring that improvements stem from algorithmic ingenuity rather than hardware advantages. Second, the RTX 4090 tier represents a mid-range GPU that many independent researchers and small labs can access, widening participation beyond elite clusters equipped with A800 or H100 devices.

3.3 Metric

When evaluating the performance of submitted Image-to-Video (I2V) acceleration methods for this challenge, we take into consideration two core dimensions that reflect both technical effectiveness and practical deployment value: quality index (covering five key sub-indicators) and speed indicator. These dimensions collectively measure whether the acceleration method can balance video quality preservation and generation efficiency enhancement.

The quality index focuses on ensuring that acceleration does not compromise the usability and visual value of generated videos, with five sub-indicators targeting different aspects of video quality.

- IndexA (I2V Subject) refers to the use of DINO[5] features to evaluate the similarity between the subject in the input image and the subject in the generated video. Its core goal is to ensure that the core subject of the input image (e.g., a person, a product) is accurately represented in the generated video, avoiding subject distortion or replacement caused by acceleration.
- IndexB (I2V Background) evaluates the consistency and logical coherence between the input image background and the generated video background, for example, ensuring that a 'park background' in the input image does not turn into an 'indoor background' in the video, and that background elements (e.g., trees, benches) maintain reasonable spatial relationships between frames.
- IndexC (Motion Smoothness[13]) is used to assess whether the motion generated in the video is smooth and adheres to the laws of physics in the real world. It avoids disjointed motion (for example, a character's arm suddenly jumps) or unnatural movements (for example, floating objects) that may result from over acceleration.
- IndexD (Imaging Quality) mainly focuses on the distortion present in the generated video frames, such as overexposure, noise, and blur. This indicator ensures that acceleration does not degrade basic visual fidelity, keeping frames clear and true to the input's texture details. We evaluate it using the MUSIQ[6] image quality predictor.
- IndexE (Dynamic Degree) addresses the potential issue of overly static videos from conservative acceleration. It evaluates the dynamic level of the generated content, ensuring that the video maintains appropriate motion activity (rather than being completely still) while achieving acceleration. We use RAFT[10] to estimate the degree of dynamics in synthesized videos.

The speed indicator serves as a critical measure of the acceleration effect of submitted methods. It quantifies the extent to which a method reduces I2V generation latency when compared to non-accelerated baseline models. The detailed scoring criteria for this speed indicator are outlined in the accompanying Table 1.

Table 1: the score of generation speed

generation speed per sample	score
< 60 seconds	1.0
60 seconds - 120 seconds	0.9
120 seconds - 240 seconds	0.8
240 seconds - 360 seconds	0.75
360 seconds - 600 seconds	0.7
> 600 seconds	The submitted method is invalid

By jointly examining generation fidelity and computational efficiency we obtain a clear picture of each acceleration technique's real-world value. The final score is:

$$\begin{aligned}\text{Final Score} = 0.2 * I2VSubject + 0.2 * I2VBackground \\ + 0.2 * MotionSmoothness \\ + 0.2 * ImagingQuality \\ + 0.2 * speed\end{aligned} \tag{1}$$

The evaluation framework strikes a deliberate balance between the visual fidelity of the generated videos and the speed gains achieved through model acceleration, ensuring that each solution not only preserves the richness of motion and detail demanded by viewers but also delivers the rapid turnaround required for real world deployment.

4 Challenge Results

A total of 355 participants registered for the challenge, reflecting significant interest in Image-to-Video generation model acceleration. Of these, 44 teams submitted at least one valid and executable entry, highlighting both the complexity and the difficulty of the task. Owing to limited computing resources and long inference times, each team was restricted to one submission per day. In total,

Table 2: ACM MM 2025 I2V Generation Model Acceleration Challenge

Rank	Team name	Score
1	EYUAN	0.9121
2	mg2773588djv99	0.9082
3	LI2V	0.9064
4	mg4771798oAc15	0.9057
5	QQ_f8JKELSzraCG	0.8963
6	mantianxing	0.8949
7	sdjfalkjfa	0.8873
8	mg5834155177	0.8872
9	mg56492545eWl	0.8859
10	webbzhou	0.8734
-	baseline	0.8658

these 44 teams submitted 771 entries. As summarized in Table 2, Team EYUAN achieved the best overall performance.

5 Challenge Methods

In this section, we outline the top-performing solutions from the ACM MM 2025 Image-to-Video Generation Model Acceleration Challenge.

5.1 Winning Method

The EYUAN team won first place in the competition with an overall score of 0.9121, and sub-scores of 0.9711 for I2V subject, 0.9761 for I2V background, 0.9884 for motion smoothness, 0.7246 for imaging quality, and 0.9 for time score. To meet challenge requirements of balancing generation quality, lowering computation/memory use, and speeding up inference, the team proposed a core solution integrating three components: split inference optimization, diffusion sampling acceleration, and post-processing enhancement.

For split inference optimization, the team designed targeted strategies for key modules (text encoding, image encoding, VAE, DiT). Text encoding used a bf16-precision model for high-accuracy text control and was paired with fp8 inference to cut memory use. Image encoding retained the original model to ensure full reference image information. VAE adopted an fp16-precision model, with testing confirming better inference results. DiT used a GGUF-quantized model[2], which sharply reduced inference memory while preserving performance and speeding up inference.

For diffusion sampling acceleration, the team used optimized measures to balance speed and quality. It applied an optimized UNet model with Q6_K quantization, cutting memory use sharply with minimal performance loss. The default sampling steps were set to 8, and combining this with Classifier-Free Guidance (cfg) achieved a reasonable trade-off between inference speed and generated content quality. Cfg strength was set to 1.0, letting the model use the conditional branch directly (skipping unconditional computation) for 1.5x acceleration. The uni_pc sampler was chosen to balance generated content quality and inference speed without overcompromise.

To address excessive smoothness and intensified inter-frame flicker in accelerated video generation, the team built a post-processing module combining texture enhancement and color balance. Input images first underwent preprocessing. Next, skin regions were detected via color space conversion (BGR to YCrCb) and mask processing. Gaussian noise was generated and blurred for naturalness. Image details were enhanced by separating high/low frequencies and scaling details. Final images were synthesized by combining smooth base, enhanced details, and noise to simulate real skin texture. Post-synthesis, pixel values were adjusted, and saturation/contrast tuning plus Gaussian filtering reduced inter-frame noise, improving video realism.

5.2 Other Deep Learning Plans

To optimize the Wan2.1 image-to-video model[11] on a single RTX 4090 and achieve an overall score of 0.9082, the mg2773588djv99 team developed a four-tier core strategy. First, mixed-precision quantization via Hugging Face optimum-quanto addressed the 70 GB original model's memory overload: parameters were stored in INT8 (15.5 GB, 4:1 compression) and dynamically converted to BF16 for inference with differentiated tactics for components, keeping total memory under 20 GB and boosting loading speed by 60%. Second, SageAttention2 [14] replaces the original attention mechanism: it quantizes Q/K to INT4 (leveraging the RTX 4090's Tensor Cores) and PV to FP8 using precision-preserving techniques, achieving a 2x speed-up, 481 TOPS on the RTX 4090, and 1.8x end-to-end acceleration without any loss in quality. Third, torch.compile optimized via selective Transformer block compilation (avoiding overhead for lightweight components like VAE) and CudaGraph disabling, delivering 30% inference speedup, 15% higher memory efficiency, and compatibility with SageAttention2[14]. Fourth, TeaCache[7], a timestep-aware cache using noise input differences as indicators (optimal threshold 0.3), reduced redundant computation for 2.5x speedup.

Another team employs TensorRT conversion for the key components of the model. Due to issues such as numerical anomalies or lack of noticeable acceleration in some components, only the T5 and Wan-DiT were ultimately converted to TensorRT engines while the remaining components retained their PyTorch implementations. Subsequently, the forward logic of Wan DiT was reconstructed to eliminate operators incompatible with TensorRT conversion and optimize node localization, which laid the groundwork for the subsequent injection of additional logics. Following this, fp8 quantization calibration was conducted with a differentiated strategy in which key layers such as embedding and head layers were excluded from quantization, while some layers were quantized to fp8, and others retained bf16 precision to balance generation quality and inference performance. Thereafter, the quantized model was exported as an ONNX file, which was further optimized using onnx-graphsurgeon to streamline the model structure through operations like constant folding and redundant node cleanup. Next, Sparse Attention[16] and Cache logics were injected at the ONNX level. For Sparse Attention, custom nodes were introduced to implement conditional branches that switch between standard attention and block sparse attention, and for the Cache logic, the MagCache scheme was adopted to skip redundant block computations. Before the final conversion, the optimized ONNX model was transformed into a TensorRT Engine: specifically, a plugin for Sparse Attention was implemented, and an engine supporting weight streaming was built to prevent out-of-memory issues during inference.

6 Conclusion

The Image-to-Video Model Acceleration Challenge addresses the core dilemma in AIGC video generation: balancing high-quality output including subject consistency, motion smoothness, and imaging clarity with low resource consumption and fast inference. Through the competition, participants proposed innovative solutions such as differentiated quantization that preserves key layer precision while compressing size, hardware-aware cache mechanisms, and compiler-level kernel optimization, all of which break traditional limitations. These solutions achieved significant efficiency gains including reduced generation time and lower VRAM use under strict constraints that prohibit full-parameter training and third-party quantized models, while mitigating quality loss, thus verifying the feasibility of "high-quality and efficiency coexistence" for large video models.

Beyond technical validation, the challenge serves as a pivotal platform for exchanging cutting-edge acceleration technologies. It aggregates diverse technical paths and establishes a comprehensive evaluation paradigm (integrating quality and speed metrics), clarifying future optimization directions, particularly the value of layer-specific optimization and hardware-software synergy for further efficiency gains.

Looking ahead, consumer-grade demands (real-time content creation, small-batch production) will drive innovation. Future research should focus on three priorities: deepening integration of model compression and inference optimization, exploring content-adaptive dynamic adjustment mechanisms, and standardizing accelerated video generation evaluation systems.

Ultimately, the challenge accelerates the transition of image-to-video technology from lab research to practical deployment, laying a foundation for improving AIGC content production efficiency and unlocking broader applications in digital marketing, education, and entertainment.

References

[1] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. 2024. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[2] Kazuki Egashira, Robin Staab, Mark Vero, Jingxuan He, and Martin Vechev. 2025. Mind the Gap: A Practical Attack on GGUF Quantization. arXiv:2505.23786 [cs.CR] https://arxiv.org/abs/2505.23786

[3] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. 2023. Structure and Content-Guided Video Synthesis with Diffusion Models. arXiv:2302.03011 [cs.CV] https://arxiv.org/abs/2302.03011

[4] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. 2024. MiraData: A Large-Scale Video Dataset with Long Durations and Structured Captions. arXiv:2407.06358 [cs.CV] https://arxiv.org/abs/2407.06358

[5] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht. 2021. Emerging properties in self-supervised vision transformers. In *ICCV.*

[6] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. MUSIQ: Multi-scale Image Quality Transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV).* 5128–5137. doi:10.1109/ICCV48922.2021.00510

[7] Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. 2024. Timestep Embedding Tells: It's Time to Cache for Video Diffusion Model. *arXiv preprint arXiv:2411.19108* (2024).

[8] Yang Liu, Hong Wang, Lei Zhang, and Xin Li. 2023. Quantization-Aware Training for Low-Latency Image-to-Video Generation. *arXiv preprint arXiv:2311.16548.*

[9] OpenAI 2025. GPT-4o. https://openai.com/index/hello-gpt-4o/.

[10] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV.*

[11] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, and Di Chen. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314* (2025).

[12] Zhi Wang, Hui Li, Yu Chen, Sen Zhang, and Wei Liu. 2024. I2V-Accel: Efficient Image-to-Video Generation via Temporal-Aware Cache and Kernel Fusion. *arXiv preprint arXiv:2403.12876.*

[13] Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, and Ziwei Liu. 2024. Evaluation Agent: Efficient and Promptable Evaluation Framework for Visual Generative Models. *arXiv preprint arXiv:2412.09645* (2024).

[14] Jintao Zhang, Haofeng Huang, Pengle Zhang, Jia Wei, Jun Zhu, and Jianfei Chen. 2025. Sageattention2: Efficient attention with thorough outlier smoothing and per-thread int4 quantization. In *International Conference on Machine Learning (ICML).*

[15] Jintao Zhang, Jia Wei, Pengle Zhang, Jun Zhu, and Jianfei Chen. 2025. SageAttention: Accurate 8-Bit Attention for Plug-and-play Inference Acceleration. In *International Conference on Learning Representations (ICLR).*

[16] Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. 2025. Spargeattn: Accurate sparse attention accelerating any model inference. In *International Conference on Machine Learning (ICML).*