# ACM Multimedia 2026 Grand Challenge Proposal: Single-Image Guided Multi-Angle Image Synthesis

Jin Chen
chenjin@mgtv.com
MGTV
Beijing, China

Haoyuan Xie
xiehaoyuan@mgtv.com
MGTV
Changsha, China

Hao Liu
liuhao@mgtv.com
MGTV
Changsha, China

Shien Song
shien@mgtv.com
MGTV
Changsha, China

Jie Yang
yangjie@mgtv.com
MGTV
Changsha, China

Han Qi
qihan@mgtv.com
MGTV
Changsha, China

Chengbo Wang
wangchb@hnu.edu.cn
Hunan University
Changsha, China

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

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

## ABSTRACT

We propose a Grand Challenge on Single-Image Guided Multi-Angle Image Synthesis, a cutting-edge task that transforms a single reference image into a sequence of multi-angle images with consistent spatial logic and content integrity. This technology addresses critical pain points in digital content creation, product visualization, and architectural spatial presentation by automating the labor-intensive multi-angle image generation process, while serving as a precise control signal for AIGC video generation to enhance the controllability and predictability of video outputs. To support this challenge, we introduce MGMultiAngle, a high-quality benchmark dataset featuring high-resolution reference images (1080p), precise 3D spatial calibration for diverse poses, multi-angle paired annotations, and detailed structured captions. MGMultiAngle overcomes limitations of existing datasets (e.g., insufficient pose coverage, incomplete spatial consistency annotations) and provides a rigorous evaluation foundation for model development. Through this challenge, we aim to foster innovations in single-image-driven novel view synthesis, spatial geometric modeling, and visual consistency preservation, accelerating the industrial application of multi-angle generation technologies over the next 3–5 years.

## KEYWORDS

single-image guided multi-angle synthesis, artificial intelligence generated content, convolutional neural networks, dataset, grand challenge


*MM '26, ,*

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

**ACM Reference Format:**
Jin Chen, Haoyuan Xie, Hao Liu, Shien Song, Jie Yang, Han Qi, Chengbo Wang, Yizhen Lao, and Yifei Xue. 2026. ACM Multimedia 2026 Grand Challenge Proposal: Single-Image Guided Multi-Angle Image Synthesis. In *Proceedings of ACM International Conference on Multimedia (MM '26).* ACM, Rio de Janeiro, Brazil, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

### 1.1 Who We Are?

The proposed challenge will be organized by the AI application team of MGTV and the visual intelligence perception research group of Hunan University (HNU-VIP).

*MGTV.* MGTV (https://w.mgtv.com/) is a new media service platform that focuses on audiovisual interaction, blending the characteristics of both internet and television. For six consecutive years (2015-2020), MGTV has been ranked among "The World's 500 Largest Media Companies". As one of the top four online video platforms in China, MGTV boasts a vast collection of copyrights for TV series, variety shows, animations, and more, totalling 3075 TV series, 75476 variety shows, and 616 animations. MGTV also has a substantial subscriber base, with over 50 million paying subscribers and 110 million daily active users.

We have extensive experience in organizing international competitions. Over the past three years, in collaboration with ACM MM, we have launched three grand challenges: the 2023 "Invisible Video Watermarking" Challenge[1], the 2024 "AI-Generated Image Detection" Challenge[2], and the 2025 "Image-to-Video Generation Model Acceleration" Challenge[3]. These events have attracted a total of more than 5,000 participants from well-known universities and leading enterprises across the globe.

*HNU-VIP.* Hunan University (http://www-en.hnu.edu.cn/) is a national key university in China, placed 195th in the 2022 U.S. News & World Report Best Global Universities Rankings. The HNU-VIP

[1]Grand Challenge 2023
[2]Grand Challenge 2024
[3]Grand Challenge 2025

**Table 1: Azimuths (Horizontal Rotation) for 96 Poses**

| Angle | Descriptor |
|---|---|
| 0° | Front view |
| 45° | Front-right quarter view |
| 90° | Right side view |
| 135° | Back-right quarter view |
| 180° | Back view |
| 225° | Back-left quarter view |
| 270° | Left side view |
| 315° | Front-left quarter view |

**Table 2: Elevations (Vertical Angle) for 96 Poses**

| Angle | Descriptor | Description |
|---|---|---|
| -30° | Low-angle shot | Camera below, looking up |
| 0° | Eye-level shot | Camera at object level |
| 30° | Elevated shot | Camera slightly above |
| 60° | High-angle shot | Camera high, looking down |

**Table 3: Distances for 96 Poses**

| Factor | Descriptor | Usage |
|---|---|---|
| ×0.6 | Close-up | Highlight details/textures |
| ×1.0 | Medium shot | Balanced, standard view |
| ×1.8 | Wide shot | Show context/environment |

research group currently consists of three faculty members[4][5][6] and multiple PhD and MSc students. They have published over 20 papers in prestigious conferences and journals including T-PAMI, IJCV, T-IP, NeurIPS, CVPR, ECCV, ACM MM, IJCAI, AAAI etc.

## 1.2 What is Single-Image Guided Multi-Angle Image Synthesis?

Single-Image Guided Multi-Angle Image Synthesis refers to the task of automatically generating a sequence of **spatially consistent, content-aligned multi-angle images** from a single reference image or a frame from a dynamic scene, with a core requirement of covering **96 predefined poses** (4 elevations × 8 azimuths × 3 distances, detailed in Tables 1, 2, 3). The task can be decomposed into four long-term research sub-tasks to drive sustained progress:

(1) **Single-Image Spatial Parsing**: Accurately extracting 3D structure, object positions, lighting, and scene elements from a single 2D reference image;
(2) **96-Pose Compliance**: Generating images that strictly follow 96 predefined poses (physical perspective rules) without spatial distortion;
(3) **Cross-Angle Consistency Preservation**: Maintaining core element consistency (shape, texture, lighting) across all 96 poses;
(4) **Efficient High-Resolution Generation**: Balancing generation speed and visual quality for 1080p images (critical for industrial deployment).

*1.2.1 96 Predefined Poses Specification.* The 96 poses are designed to cover typical industrial scenarios (product display, architectural visualization) and follow physical spatial rules, with three components: azimuths representing horizontal rotation angles, elevations denoting vertical shooting angles, and shooting distance scaling factors for different view scales.

*1.2.2 State-of-the-Art (SOTA) Limitations.* Current mainstream approaches to this task can be categorized into three technical paradigms, each with distinct strengths and limitations:

(1) **LoRA-Finetuned Generative Models**: Represented by Qwen-Image LoRA, this paradigm leverages large-scale image generation models (e.g., Qwen-Image-Edit[6]) and fine-tunes them with LoRA adapters to learn angle-specific generation patterns. It supports multiple preset viewing angles (combining azimuth, elevation, and distance) and excels at content consistency with the reference image. However, it often struggles with strict spatial geometric constraints, leading to subtle perspective distortions in extreme views (e.g., ±90° elevation).
(2) **Point Cloud-Driven Generation**: Exemplified by Apple's Sharp Point Cloud[4] scheme and related open-source variants, this approach first reconstructs a dense 3D point cloud from the reference image (via techniques like Structure-from-Motion or depth estimation) and then uses the point cloud as geometric guidance for view synthesis. It ensures strong spatial consistency by adhering to point cloud projections but faces challenges with texture detail preservation and computational efficiency—especially for high-resolution image generation.
(3) **3D Gaussian Splatting Integration[3]**: Building on the real-time radiance field rendering capability of 3D Gaussian Splatting, this paradigm reconstructs a 3D Gaussian representation from the reference image and renders novel views by projecting Gaussians to 2D. It balances spatial accuracy and visual quality but requires careful optimization of anisotropic covariance and density control to avoid "splotchy" artifacts in image outputs[5].

Closed-source models like Sora demonstrate strong visual coherence for multi-angle generation, while open-source solutions (e.g., Qwen-Image-Edit-Multiple-Angles-LoRA, Sharp-derived point cloud tools) focus on specific technical tradeoffs. However, existing methods still struggle with three critical gaps: (1) spatial distortion in extreme views, (2) content inconsistency for fine-grained details (e.g., texture patterns, small object features), and (3) inefficient high-resolution generation—gaps this challenge aims to address.

## 1.3 Why Single-Image Guided Multi-Angle Image Synthesis is Important?

Single-Image Guided Multi-Angle Image Synthesis is a transformative technology for multimedia content creation, with profound value in both industrial and academic domains.

[4]Yizhen Lao: Professor
[5]Yifei Xue: Researcher
[6]Chengbo Wang: Researcher

From an industrial perspective, it addresses three key pain points in modern content production. First, it enhances the controllability of AI video production. Currently, AIGC video generation (text-to-video and image-to-video) has become an essential multimedia tool in industries like advertising, film, and e-commerce, but it suffers from insufficient control over the final output. Existing methods often produce results deviating from user expectations (e.g., unexpected camera angles, mismatched scene perspectives) due to limited fine-grained control over camera movement and spatial logic. In contrast, single-image guided multi-angle synthesis provides a solution: generating predefined multi-angle images first allows users to directly use these consistent perspectives to control the camera trajectory of AI-generated videos (e.g., selecting a sequence of 0°→45°→90° angles to simulate a smooth orbiting shot). This "multi-angle as control signal" approach drastically improves the predictability and controllability of video outputs, ensuring alignment with creative intentions.

Second, it streamlines multi-view storyboard production. Manual creation of multi-view storyboards—an essential step in film pre-production—requires repeatedly refining spatial logic based on a single reference image or scene sketch, which is time-consuming (often taking days for a single project) and prone to perspective errors. A high-performance single-image guided multi-angle algorithm can quickly generate spatially compliant multi-view storyboard sequences (e.g., 5–8 key angles from the 96 predefined poses) without manual intervention, reducing production time by over 80% and eliminating human-induced perspective biases.

Third, it solves long-standing inefficiencies in traditional multi-angle content production. Manual creation of multi-angle content (e.g., product 360° displays) requires multi-camera shooting or complex 3D modeling, which is costly and time-intensive. Automated generation lowers entry barriers for small businesses and independent designers, enabling scalable applications in product design, architectural visualization, and e-commerce.

From a technical perspective, the task pushes the boundaries of AIGC and computer vision by integrating single-image 3D understanding, novel view synthesis, and visual consistency preservation—core challenges in modern AI research. It requires models to "understand" spatial geometry from 2D inputs and "generate" plausible, consistent views, bridging the gap between 2D content and 3D spatial reasoning, and laying the foundation for advanced tasks like 4D video generation (3D space + time) and interactive AR/VR content creation.

Furthermore, the technology enhances content diversity and accessibility. It allows creators to generate dynamic multi-angle content from static assets, eliminating the need for expensive equipment or specialized skills. This democratizes high-quality multi-angle content production, empowering small businesses, educators, and hobbyists to create professional-grade visual materials.

# 2 DATASET

## 2.1 Dataset Introduction

Existing datasets for view synthesis (e.g., DTU[2], BlendedMVS[7]) have critical limitations for single-image guided multi-angle image synthesis:

- lack of spatial calibration for multi-angle-specific views;
- incomplete annotation of content consistency (e.g., texture detail preservation);
- low resolution (often <= 720p) and simplistic scene composition (insufficient for complex industrial scenarios);
- lack of structured captions describing view intentions.

To address these gaps, we introduce **MGMultiAngle**, a large-scale, high-quality benchmark dataset tailored for single-image guided multi-angle image synthesis. The dataset includes 1,000 high-resolution reference images (1080p) covering diverse scenarios: product design, architectural spaces, daily objects, indoor scenes, and other relevant scenarios. Each reference image is paired with more than 96 annotated multi-angle images (total images) that follow physical perspective rules (e.g., front, side, 45° oblique, overhead views). To ensure the highest quality and reliability, all data in MGMultiAngle undergoes strict quality control:

- **Spatial Accuracy**: Each reference image and its paired multi-angle images are calibrated using Structure-from-Motion (SfM) and 3D Gaussian Splatting. The azimuth and elevation errors are guaranteed to be < 0.5°, ensuring precise alignment with the 96 predefined poses.
- **Content Consistency**: Core object features (shape, texture, color) are verified by a team of three professional visual designers to ensure no discrepancies between the reference image and the multi-angle pairs.
- **Visual Quality**: All images are free of blur, compression artifacts, and color distortion. The LAR-IQA[1] ( A Lightweight, Accurate, and Robust No-Reference Image Quality Assessment Model) score for every image is < 0.5, indicating high perceptual quality.

A demo dataset[7] is currently available (with diverse sample types for demonstration). The full dataset is being prepared, to be completed before the challenge starts and released in phases according to the challenge timeline. All datasets will be publicly accessible for academic research after the challenge ends.

## 2.2 What Are The Differences Over The Previous Datasets?

MGMultiAngle outperforms existing datasets in four key aspects:

- **96-Pose Spatial Calibration**: Unlike existing datasets that provide only a few sparse views, MGMultiAngle provides full spherical coordinate annotations (azimuth, elevation, distance) for all 96 predefined poses for each reference image. This enables quantitative evaluation of pose accuracy—a critical metric for this task.
- **Industrial Scenario Diversity**: Most existing datasets (e.g., DTU) focus on simple, isolated objects in controlled lab environments. MGMultiAngle prioritizes real-world industrial scenarios (products, architecture, indoor scenes) to ensure that models developed on this dataset are directly applicable to industry needs.
- **High-Resolution Annotations**: All images in MGMultiAngle are 1080p, significantly higher than the 720p resolution of many existing datasets. Additionally, each image includes

[7] Google Drive

detailed texture annotations (e.g., "matte black metal," "wood grain") to facilitate fine-grained consistency evaluation.
- **Legal copyright ownership**: All images in MGMultiAngle are either self-shot by MGTV or licensed from professional content providers, ensuring no copyright issues for academic research and challenge participation.

# 3 EVALUATION

## 3.1 Data Setting

The MGMultiAngle dataset contains 1,000 high-resolution reference images (1080p). All data will be publicly released in full after the challenge concludes to ensure fair competition and promote further research. The dataset is split into three parts for evaluation:

- **Preliminary Test Set**: Released at the challenge launch. This set is intended for participants to test their model implementations, debug inference pipelines, and familiarize themselves with the dataset format and evaluation metrics.
- **Test Set A**: Used in Stage 1 (Qualification) of the challenge. This set is kept confidential and released in the end of challenge.
- **Test Set B**: Used in Stage 2 (Final) of the challenge. This set is kept confidential and only released to the evaluation system during the final stage to rigorously assess the generalizability of participants' models. It includes hidden scenarios and object categories not present in the Preliminary Test Set or Test Set A.

For each reference image in all test sets, participants are provided with the target parameters (azimuth, elevation, distance) for the 96 poses they need to generate. The evaluation system strictly checks whether the generated images match these target parameters.

## 3.2 Submission

The competition is structured into two distinct stages. During the initial stage, participants must submit a Docker file that encompasses their testing codes. Our scoring system will then evaluate the uploaded images to determine their performance. Advancing to the second stage, the top ten finalists, as determined by their cumulative scores, are obligated to provide a Docker file that not only contains the testing codes but also the training codes. Additionally, these finalists must furnish an explanatory document that outlines their comprehensive approach, which will be subject to verification.

## 3.3 Metric

We design a concise, objective evaluation suite focusing on **content consistency and visual quality**, the core requirements for single-image guided multi-angle image synthesis. All metrics are automatically computable, with no subjective evaluation. The scoring method for each metric is defined in Table 4.

### 3.3.1 *Scoring Rules.*

(1) **Single Sample Total Score**: Calculated as the sum of scores for all metrics:

$$\text{Single Sample Score} = \frac{\text{PSNR}}{40} + \text{SSIM} + (1 - \text{LPIPS}) + \text{LAR-IQA} \quad (1)$$

**Table 4: Metric Scoring Method**

| Metric | Scoring Formula |
|---|---|
| PSNR | PSNR/40 |
| SSIM | SSIM |
| LPIPS | 1 − LPIPS |
| LAR-IQA | LAR-IQA |

The scoring formulation in Table4 and the summation calculation in this Formula are designed to balance the different evaluation metrics, eliminating the impact of dimensional differences and varying value ranges across metrics to ensure each indicator contributes fairly to the comprehensive evaluation.

(2) **Final Score**: The final score of the participating algorithm is the arithmetic mean of the single sample total scores of all evaluated valid samples:

$$\text{Final Score} = \frac{1}{N} \sum_{i=1}^{N} \text{Single Sample Score}_i \quad (2)$$

where $N$ is the number of valid samples (samples that meet basic spatial perspective rules).

### 3.3.2 *Metric Definitions.*

- **PSNR (Peak Signal-to-Noise Ratio)**: Measures pixel-level consistency between the generated multi-angle image and the reference image (higher = better pixel consistency).
- **SSIM (Structural Similarity Index)**: Measures structural consistency (e.g., object shape, edge contour) between the generated image and the reference image (range [0,1], higher = better structural consistency).
- **LPIPS (Learned Perceptual Image Patch Similarity)**: Measures human-perceptual consistency between the generated image and the reference image (range [0,1], lower = better perceptual consistency).
- **LAR-IQA (A Lightweight, Accurate, and Robust No-Reference Image Quality Assessment Model)**: A no-reference metric that measures the visual quality of the generated image (higher = better quality).

All metrics are computed after **unified preprocessing** of the reference image and generated image:

(1) Crop the core object region (exclude background) using a object detection model;
(2) Ensure both the reference image and generated image maintain a 1080P resolution during evaluation;
(3) Normalize pixel values to [0, 1].

# 4 AWARD

To motivate more participation, we will provide the following **bonus** to the winners of the proposed challenge:

- **1st**: 16,500 USD
- **2nd**: 9,000 USD
- **3rd**: 4,500 USD

## 5 COMMITMENT

Building on our experience organizing four consecutive Chinese algorithm competitions (average prize pool > 1M RMB), we solemnly commit to:

(1) **Dataset Release**: The MGMultiAngle dataset will be prepared prior to the launch of the challenge, and the full version of the dataset will be released upon the challenge's conclusion, accompanied by comprehensive documentation (including calibration tools and annotation guidelines) and dedicated technical support (via GitHub issue responses and email correspondence);
(2) **Website Maintenance**: Establish and maintain an official challenge website (https://challenge.ai.mgtv.com//#/match) for at least 3 years after the challenge. The website will include: - Competition rules, submission guidelines, and leaderboards; - Dataset downloads (after challenge conclusion); - Winning technical reports and open-source codes;
(3) **Fair Evaluation**: Use a closed-source evaluation system to prevent cheating; provide a pre-evaluation tool (based on the preliminary set) for participants to verify submissions; publish a detailed evaluation report (including top methods' performance breakdown) after the challenge;
(4) **ACMMM Collaboration**: Work closely with ACM MM 2026 organizers to publicize the challenge, including promotions on the ACM MM website, conference emails, and social media;
(5) **Long-Term Impact**: Host annual follow-up workshops to track progress in multi-pose synthesis, updating the dataset with new scenarios (e.g., more dynamic objects) to maintain relevance.

## 6 CONTACT

There are three organizers who will be responsible for organizing, publicizing, reviewing, and judging the Grand Challenge submissions as described in the proposal:

- Jin Chen. Researcher, MGTV. Email: chenjin@mgtv.com
- Haoyuan Xie. Researcher, MGTV.
  Email: xiehaoyuan@mgtv.com
- Yizhen Lao. Professor, Hunan University.
  Email: yizhenlao@hnu.edu.cn

## REFERENCES

[1] Nasim Jamshidi Avanaki, Abhijay Ghildyal, Nabajeet Barman, and Saman Zadtootaghaj. 2024. LAR-IQA: A Lightweight, Accurate, and Robust No-Reference Image Quality Assessment Model. *arXiv preprint arXiv:2408.17057* (2024).
[2] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. 2014. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 406–413.
[3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (July 2023). https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/
[4] Lars Mescheder, Wei Dong, Shiwei Li, Xuyang Bai, Marcel Santos, Peiyun Hu, Bruno Lecouat, Mingmin Zhen, Amaël Delaunoy, Tian Fang, Yanghai Tsin, Stephan R. Richter, and Vladlen Koltun. 2025. Sharp Monocular View Synthesis in Less Than a Second. *arXiv preprint arXiv:2512.10685*. https://arxiv.org/abs/2512.10685
[5] Lukas Radl, Michael Steiner, Mathias Parger, Alexander Weinrauch, Bernhard Kerbl, and Markus Steinberger. 2024. StopThePop: Sorted Gaussian Splatting for View-Consistent Real-time Rendering. arXiv:2402.00525 [cs.GR] https://arxiv.org/abs/2402.00525
[6] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. 2025. Qwen-Image Technical Report. arXiv:2508.02324 [cs.CV] https://arxiv.org/abs/2508.02324
[7] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. 2020. BlendedMVS: A Large-scale Dataset for Generalized Multi-view Stereo Networks. *Computer Vision and Pattern Recognition (CVPR)* (2020).