"Identity-Preserving Video Generation Challenge"

A Grand Challenge Proposal to ACM Multimedia 2026

Challenge Overview

Recent advancements in Multimodal Artificial Intelligence-Generated Content (AIGC) have catalyzed a paradigm shift in digital content creation. Cutting-edge models such as DALL·E, Stable Diffusion, Imagen, Flux, and Midjourney have demonstrated unprecedented capabilities in synthesizing high-fidelity images from textual prompts, while systems like Sora, Runway Gen-3, and HiDream.ai have extended these breakthroughs to the video domain. The emergence of such technologies has not only astonished the research community but also underscored the transformative potential of AI-driven video generation across industries, including entertainment, education, virtual reality, and personalized media.

Video generation, as a fundamental task in multimedia field, demands the simultaneous modeling of spatial details, temporal dynamics, and semantic consistency. State-of-the-art text-to-video systems excel at producing visually compelling content directly from textual prompts. However, a critical limitation persists: the inherent stochasticity of diffusion processes often leads to inconsistent appearances of key subjects (e.g., characters, objects) across frames/scenes. For instance, facial features of a character may drift unnaturally over time, or object attributes might degrade during prolonged sequences. This instability stems from inherent challenges in disentangling identity-related features from contextual variables (e.g., pose, lighting, occlusion) within high-dimensional latent spaces—a problem exacerbated by the stochastic nature of generative processes. In this grand challenge, we aim to alleviate this limitation by introducing Identity-Preserving Video Generation (IPVG) task, which maintains the consistency of given reference identity along text-to-video generation process. Practically, IPVG is pivotal for applications such as personalized avatar animation, historical figure reenactment, and brand-specific content creation, where even minor identity deviations can compromise usability or authenticity.

This year's IPVG grand challenge includes two tracks: **Facial Identity-Preserving Video Generation** and **Sequential Action Identity-Preserving Video Generation**. In general, the goal of this grand challenge is two-fold: (a) coalescing community effort around new challenging identity-preserving video generation datasets, and (b) offering a fertile ground for designing controllable generative models to facilitate precise identity binding in video generation, aiming to propel the field toward more accountable and user-steerable video synthesis systems.

To further motivate and challenge the academic and industrial research communities, we are releasing updated benchmarks, including the **Identity-Preserving Video Benchmark (VIP-200K)** and the newly constructed **ReactID-Data**. Note that these datasets can only be used for research purpose.

By participating in this challenge, you can:

- Leverage VIP-200K/ReactID-Data benchmark to boost research on the emerging task of identity-preserving video generation;
- Try out your identity-preserving video generation systems using real world data;
- See how them compare to the rest of the community's entries;
- Get to be a contender for ACM Multimedia 2026 Grand Challenge.

Task Description

This year we will focus on two tasks, i.e., facial identity-preserving text-to-video generation and sequential action identity-preserving text-to-video generation.

Track 1: Facial Identity-Preserving Text-to-Video Generation

In the first track, given the videos and the corresponding prompts plus reference identity facial images, the goal is to learn a video generation model capable of synthesizing temporally consistent videos that are both semantically aligned with prompts and faithful to the identity-preserving constraints derived from the reference facial images.

Track 2: Sequential Action Identity-Preserving Video Generation

This track introduces structured timeline-based constraints and generalized subject consistency, expanding the scope beyond facial identity to include full-body humans, animals, and specific objects. Addressing the fundamental trade-off between subject consistency and action realism, this track requires models to generate videos from reference images and structured timeline prompts that specify multiple sub-actions with precise timestamps. The goal is to synthesize videos where the subject faithfully performs the sequence of actions, maintaining strict visual consistency of the reference subject—whether living or non-living—throughout the video, while ensuring accurate synchronization of actions to avoid artifacts or identity loss during complex movements.

Contestants are asked to develop identity-preserving video generation systems based on the provided datasets. For evaluation purposes, a contesting system is asked to produce at least one video for each <reference identity, prompt> pair given in the testing set .

Relevance to Previous Challenges

Most of the organizers in this proposal have successfully co-organized MSR Video to Language Challenge in ACM MM 2016 and ACM MM 2017, Pre-training for Video Captioning Challenge in ACM MM 2020, Pre-training for Video Understanding Challenge in ACM MM 2021 and ACM MM 2022, and Identity-Preserving Video Generation Challenge in ACM MM 2025, as listed below. Similar to the vision-to-language objectives in previous challenges, our challenge in this Generative AI era is also specific to the problems of bridging vision and language, i.e., language-to-vision, thereby offering a valuable venue to foster multimedia research into identity-preserving text-to-video generation.

- MSR Video to Language Challenge 2016, ACM MM 2016 Grand Challenge (**30 teams**)
- MSR Video to Language Challenge 2017, ACM MM 2017 Grand Challenge (**15 teams**)
- Pre-training for Video Captioning Challenge 2020, ACM MM 2020 Grand Challenge (**50 teams**)
- Pre-training for Video Understanding Challenge 2021, ACM MM 2021 Grand Challenge (**40 teams**)
- Pre-training for Video Understanding Challenge 2022, ACM MM 2022 Grand Challenge (**40 teams**)
- Identity-Preserving Video Generation Challenge 2025, ACM MM 2025 Grand Challenge (**20 teams**)

Datasets

Dataset	Context	Source	#Video	#IDs	#Hours	#Prompt
VIP-200K	Video-Prompt-identity triplets	Automatic crawling from web	500,000	200,000	1,700	500,000

To formalize the task of identity-preserving text-to-video generation, we are offering the following datasets to participants:

For Track 1:

To formalize the task of identity-preserving text-to-video generation, we are offering the following datasets to participants:

- Training Dataset: Comprising 500,000 videos in the VIP-200K dataset. Each video is coupled with a textual prompt and one or more identity images. Each identity (ID) is defined by one or more video frames, with bounding boxes around the face.
- Testing Dataset: Containing 200 person IDs (unseen in the training data). Each ID is defined by one or more portrait images, and we provide five textual prompts for video generation. Accordingly, we have 1000 <identity images, prompt> pairs are utilized for testing.

For Track 2:

- We provide the ReactID-Data, a large-scale dataset constructed with a high-precision pipeline to ensure reliable subject-video correspondence. This dataset includes subject-to-video pairs with timeline annotations, where each video is associated with structured descriptions detailing sub-actions and their specific timestamps. The data is filtered to ensure high aesthetic quality and precise subject alignment.
- **Testing Dataset:** Containing unseen identities paired with complex, multi-action timeline prompts (e.g., "0-2s: [Action A]; 2-5s: [Action B]") to evaluate the model's ability to handle sequential dynamics and identity preservation simultaneously.

Submission Format

For each track, each team is allowed to submit the results of at most three runs and selects one run as the primary run of the submission (we do not guarantee to evaluate second and third runs), which will be measured for performance comparison across teams. Each run must be formatted in a self-contained ZIP file as following:

```
submission/
├──── id001/
│       ├──── prompt1.mp4
│       ├──── prompt2.mp4
│       ├──── prompt3.mp4
│       ├──── prompt4.mp4
│       └──── prompt5.mp4
... ...
├──── id200/
        ├──── prompt1.mp4
        ├──── prompt2.mp4
        ├──── prompt3.mp4
        ├──── prompt4.mp4
        └──── prompt5.mp4
```

Note: All submitted videos must be encoded using the H.264 video encoder and saved in MP4 format. Any submission that does not meet these requirements may be disqualified.

Evaluation Metric

For both tracks, the final score for the generated videos will be assessed using a combination of objective metrics and human studies, focusing on key dimensions: identity preservation, video quality and action-timeline alignment (especially for Track 2).

1. Identity Preservation:
   The Identity Preservation will be evaluated by feature similarity between the generated video and the given reference identity image. In addition, multiple annotators will provide manual scores to assess the accuracy of the Identity Preservation.
2. Video Quality:
   The video quality will be assessed by considering three factors: visual quality, motion dynamics, and text alignment. Objective scoring models will be used to evaluate each of these three factors, while multiple annotators will also assess the subjective quality of the videos.
3. Action-Timeline Alignment (Track 2):
   For the Track 2, we will additionally evaluate the text-to-video alignment and temporal consistency to measure how accurately the generated actions match the specified timestamps in the timeline prompts.

Finally, the overall score for each submission will be derived from a weighted combination of these objective evaluation results and a subjective user study conducted on the testing set.

Participation

The Challenge is a team-based contest. Participants are welcome to enter in one or both tracks. Each team can have one or more members, and an individual cannot be a member of multiple teams.

At the end of the Challenge, all teams will be ranked based on evaluation described above. For each track, the top three performing teams will receive award certificates and/or cash prizes (prize amounts TBD). At the same time, all accepted submissions are qualified for the conference's grand challenge award competition.

Timeline

- March 10, 2026: Web Site and Call for Participation Ready
- March 15, 2026: Dataset available for download (training and validation sets)
- May 10, 2026: Testing set of each track available for download
- May 16, 2026: Results submission
- May 17 - May 21, 2026: Objective evaluation
- May 22, 2026: Evaluation results announce
- May 28, 2026: Paper submission deadline

Paper Submission

Please follow the guideline of ACM Multimedia 2026 Grand Challenge for the paper submission.

Challenge Organizers

Yingwei Pan, HiDream.ai, Beijing, China
Email: panyw.ustc@gmail.com

Dr. Yingwei Pan is currently a technical director with HiDream.ai, Beijing, China. His research interests include vision and language, and visual content understanding. He has authored or co-authored about 60 papers in top-notch Conferences and Journals. Dr. Pan was one of core designers of top-performing multimedia analytic systems in worldwide competitions such as COCO Image Captioning, ActivityNet Dense-Captioning Events in Videos Challenge 2017 and Visual Domain Adaptation Challenge 2018 and 2019. He received Ph.D. degree in Electrical Engineering from University of Science and Technology of China in 2018. For his contributions to vision and language, and multimedia search, he was awarded the 2015 Microsoft Research Asia PhD Fellowship.

Yiheng Zhang, HiDream.ai, Beijing, China

Email: yihengzhang.chn@gmail.com

Dr. Yiheng Zhang is currently a senior researcher at HiDream.ai in Beijing, China. His research interests include high-quality image/video generation, identity-preservation generation, and efficient inference for generative models. He has participated in several prestigious competitions, including the Visual Domain Adaptation Challenge and the ActivityNet Challenge, with a focus on video activity understanding and visual generation.

Zhaofan Qiu, HiDream.ai, Beijing, China

Email: zhaofanqiu@gmail.com

Dr. Zhaofan Qiu is currently a senior researcher at HiDream.ai, Beijing, China. His current research interests include large-scale video classification, high-quality video generation, and multimedia understanding. He has participated several large-scale video analysis competitions, such as ActivityNet Large Scale Activity Recognition Challenge, and THUMOS Action Recognition Challenge. He was awarded the MSRA Fellowship in 2017.

Ting Yao, HiDream.ai, Beijing, China

Email: tingyao.ustc@gmail.com

Dr. Ting Yao currently the CTO of HiDream.ai, a high-tech startup company focusing on generative intelligence for creativity. Previously, he was a Principal Researcher with JD AI Research in Beijing, China and a researcher with Microsoft Research Asia in Beijing, China. He has co-authored more than 100 peer-reviewed papers in top-notch conferences/journals. He has developed one standard 3D Convolutional Neural Network, i.e., Pseudo-3D Residual Net, for video understanding, and his video-to-text dataset of MSR-VTT has been used by 400+ institutes worldwide. He serves as an associate editor of IEEE Transactions on Multimedia, Pattern Recognition Letters, and Multimedia Systems. His works have led to many awards, including 2015 ACM-SIGMM Outstanding PhD Thesis Award, 2019 ACM-SIGMM Rising Star Award, 2019 IEEE-TCMC Rising Star Award, 2022 IEEE ICME Multimedia Star Innovator Award, and the winning of more than 10 championship in worldwide competitions.

Tao Mei, HiDream.ai, Beijing, China

Email: tmei@live.com

Dr. Tao Mei is the Founder and CEO of HiDream.ai. Previously, He was a vice president with JD.COM and a senior research manager with Microsoft Research. He has authored or co-authored more than 200 publications (with 12 best paper awards) in journals and conferences, 10 book chapters, and edited five books. He holds more than 25 US and international patents. He is or has been an Editorial Board Member of IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Multimedia, ACM Transactions on Multimedia Computing, Communications, and Applications, Pattern Recognition, etc. He is the general co-chair of IEEE ICME 2019, the Program co-chair of ACM Multimedia 2018, IEEE ICME 2015 and IEEE MMSP 2015. He is a fellow of IAPR (2016), a distinguished scientist of ACM (2016), and a Distinguished Industry Speaker of IEEE Signal Processing Society (2017).

URL: https://taomei.me/