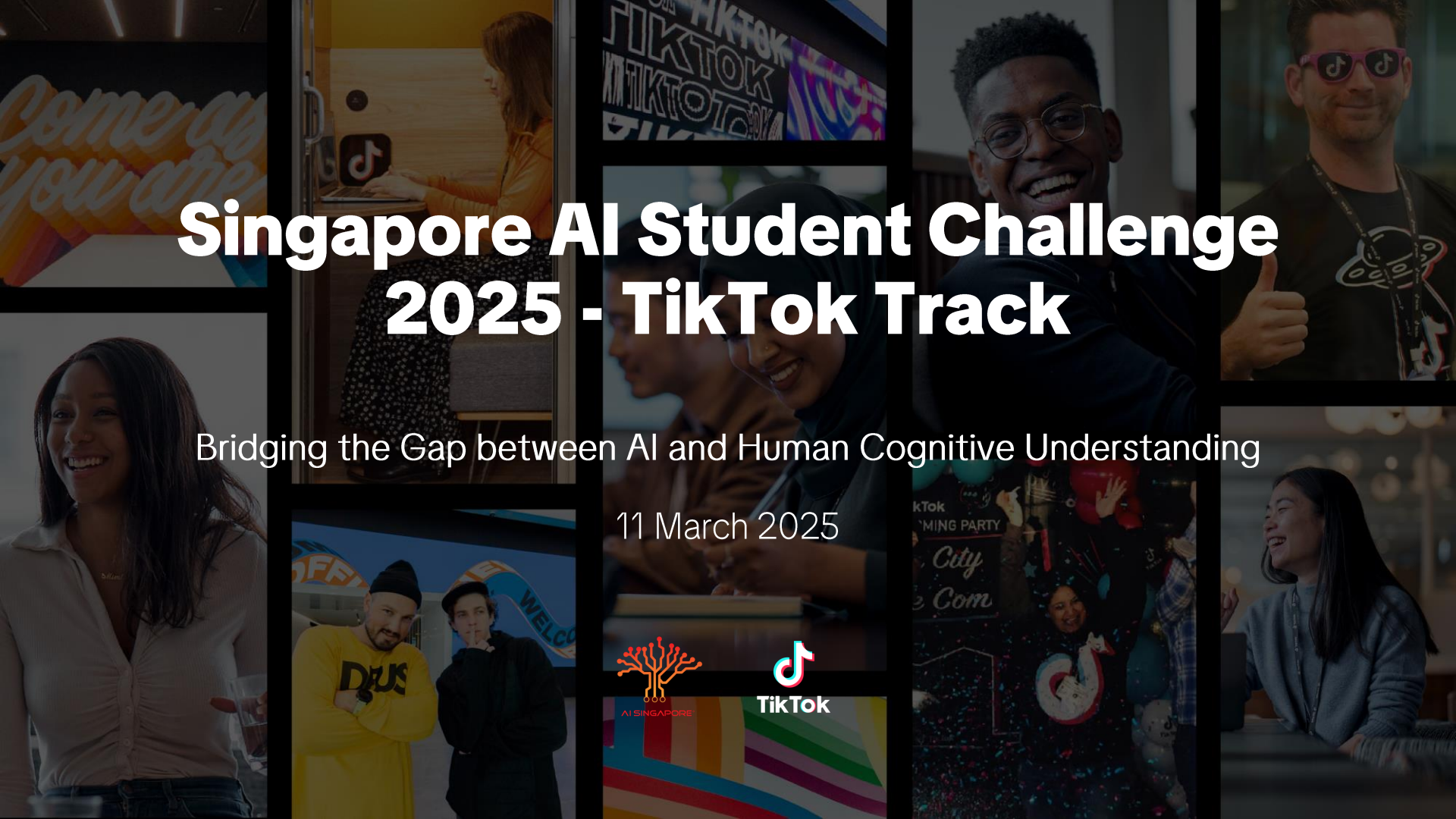


Singapore AI Student Challenge 2025 - TikTok Track

Bridging the Gap between AI and Human Cognitive Understanding

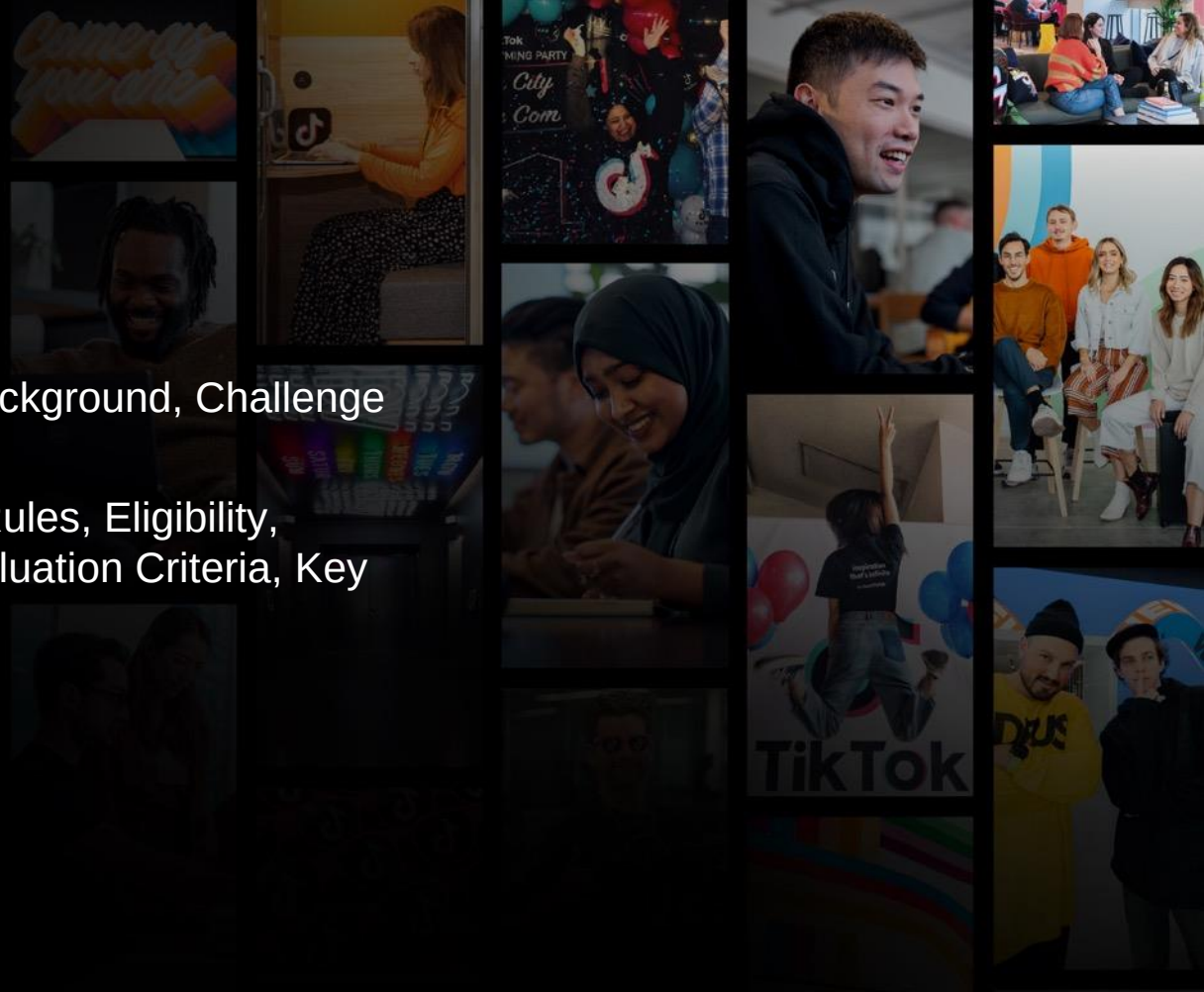
11 March 2025

Agenda

Introduction to TikTok

Contest Introduction (Project Background, Challenge Content, Example Analysis)

Contest Process (Competition Rules, Eligibility, Submission Requirements, Evaluation Criteria, Key Dates, Prizes)



Maria Devasia

APAC Early Careers

Welcome

TikTok is the leading destination for short-form mobile video.

2017 — Just getting started

2020 — Topping app stores all over the world[1]

2021 — More than 1 billion people on TikTok

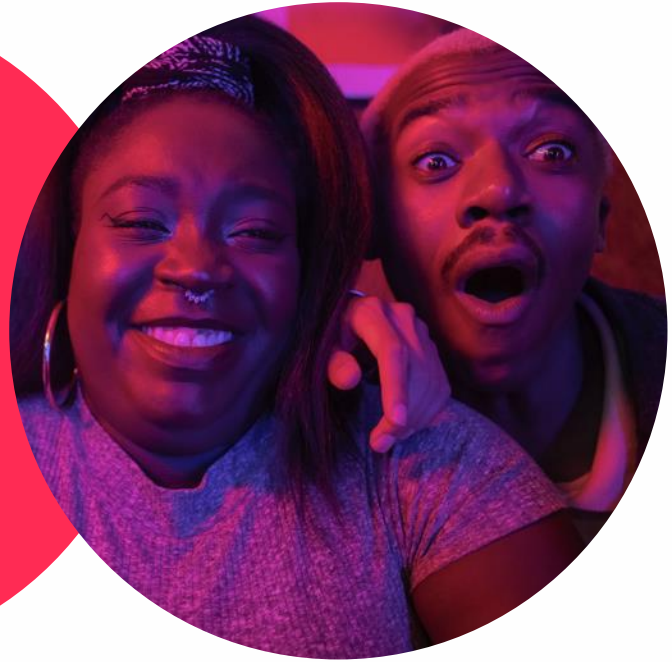
2022 — Started launching TikTok Shop in various parts of the world

✨ **Inspiring creativity and bringing joy to our community every day!** ✨

Source[1]: Apptopia's Top 10 Most Download apps in 2020

Our mission

Inspire creativity and bring joy



Key product features

Personalized

An endless stream of customized videos

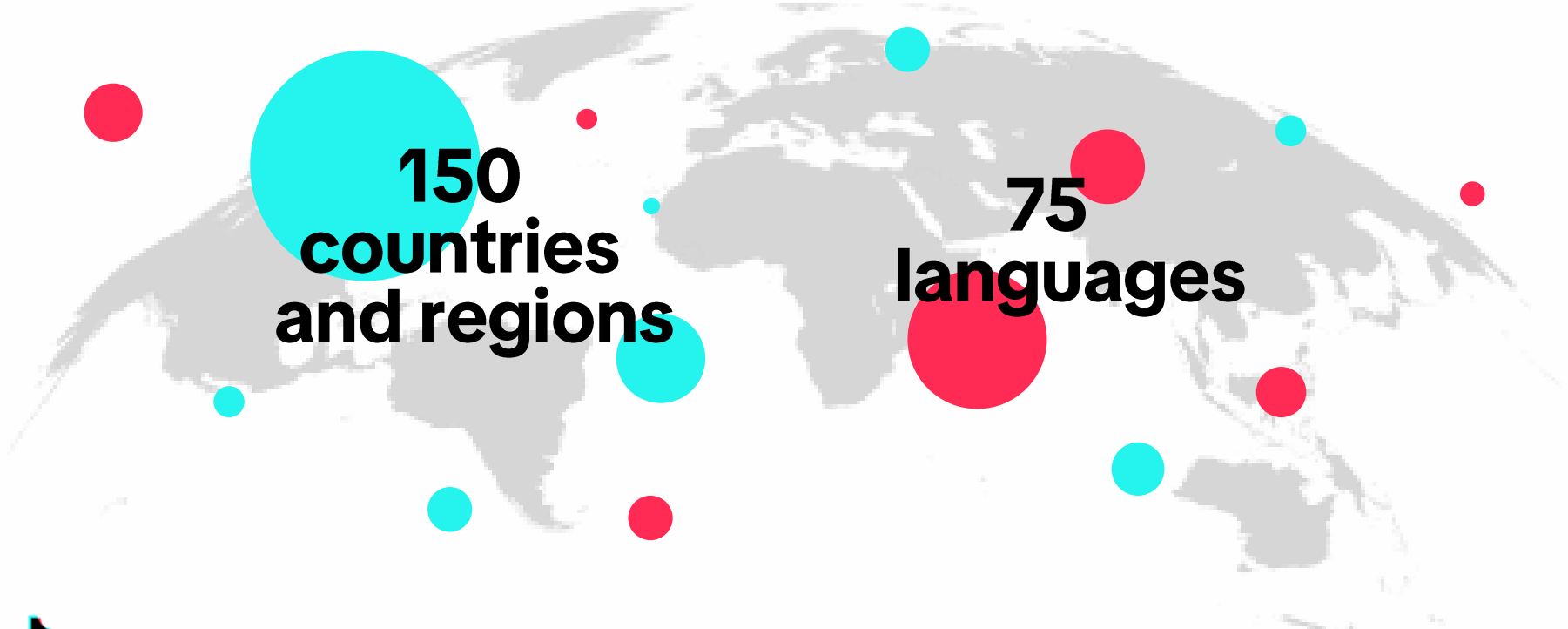
Bite-sized

Intuitive and easy to use

Effortless

Life’s moving fast, so let's make every second count

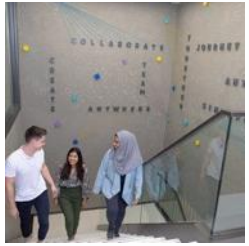
One of the most downloaded apps



Our global offices

Los Angeles

London



Singapore

Our company values

Always day 1

Champion diversity and inclusion

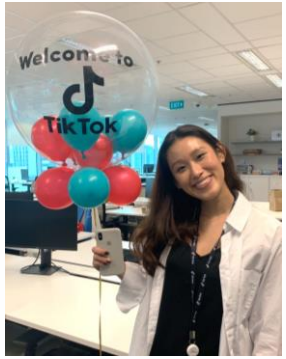
Be candid and clear

Seek truth and be pragmatic

Be courageous and aim for the highest

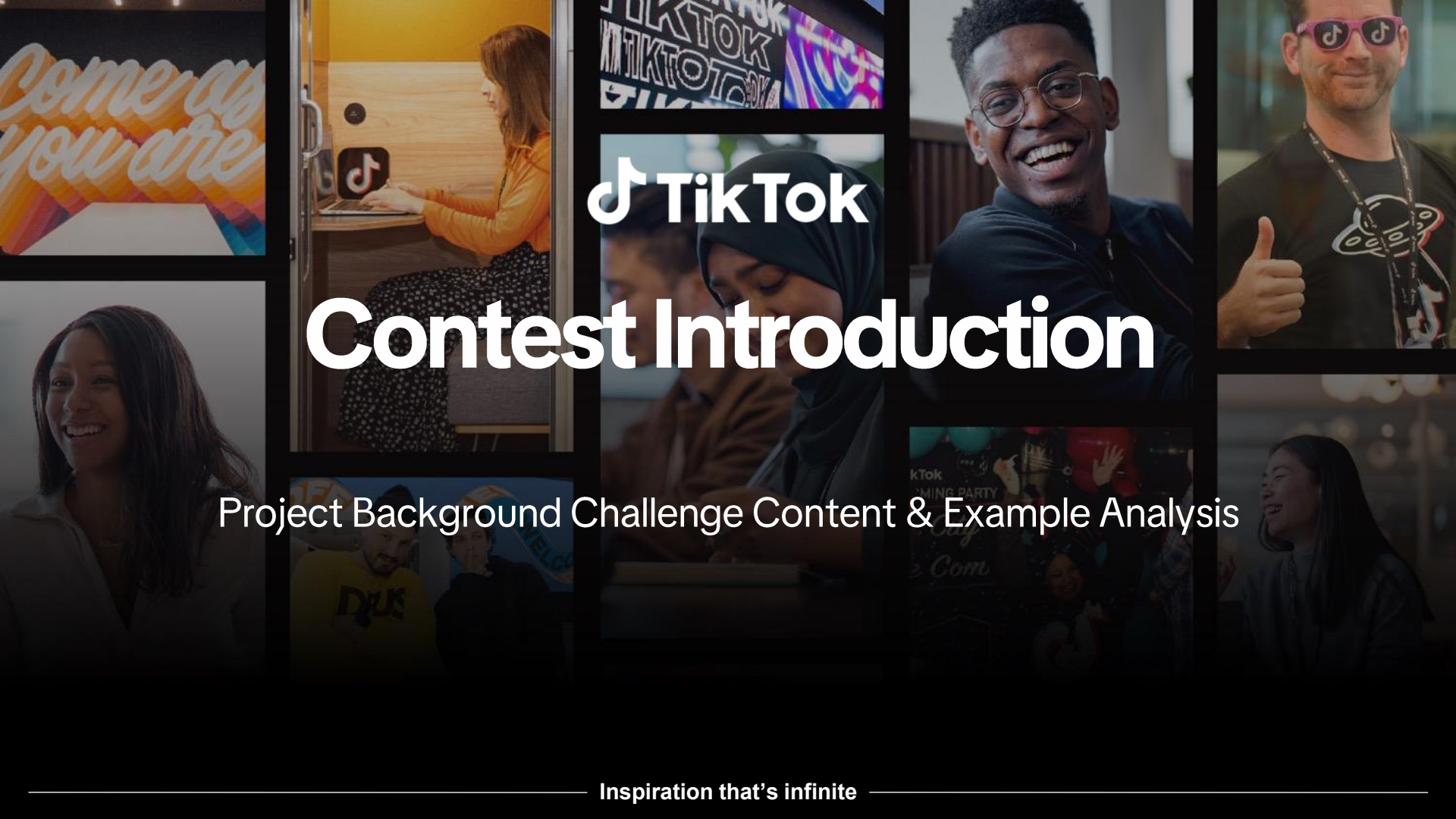
Grow together

Learn and grow with us







TikTok

Contest Introduction

Project Background Challenge Content & Example Analysis

Shan Duan

AI Innovation Center, TikTok

Challenge background

Development Status of Multimodal Large Language Models (MLLMs)

Rapid Advancements:
Significant improvements in proprietary and open-source models

Potential for General-purpose Visual Assistants:
Aimed at approaching human-level intelligence in video understanding

Challenge background

Existing Problems

Real-World Intelligence Deficit

Models struggle with dynamic, context-rich scenarios that require deep, human level understanding.

Inadequate Robustness

Models frequently fail to maintain consistent understanding in the face of these scenarios.

Challenge goal

With AI Singapore and TikTok co-hosting,

A hackathon challenge is introduced to inspire innovative solutions on the problems mentioned.

Challenge content

Method

To benchmark the performance iterations and improvements of challengers:

- We collaborated with Nanyang Technological University (NTU) to use the Video Turing Test (Video-TT) benchmark*
- Designed to evaluate the human-level performance of video LLMs through visual question-answering (VQA)

* Our challenge will use the challenge-specific holdout setpublished in Video-TT.
(a set without public release of ground-truth answer)

Challenge content

Bechmark Evaluation Criteria

- **Correctness:** Models must accurately interpret and answer questions about video content.
- **Robustness:** Models need to maintain consistent and correct responses in natural adversarial settings.

Challenge content

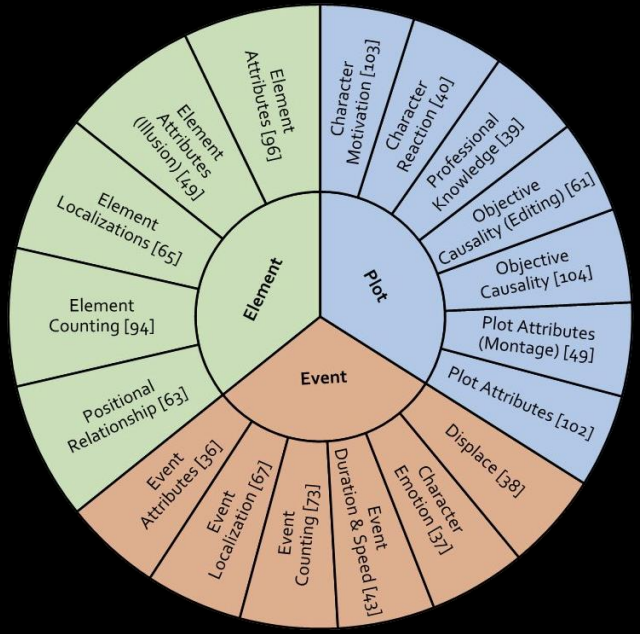
Potential Solution Directions*

- **Continuous Optimization of the Model:** fine-tuning, hyperparameter tuning, etc.
- **Agent Design:** with dynamic real-world interaction or enhanced cognitive abilities considered, etc.

* Challengers are encouraged to use each solutions mentioned above, or a combination of them. Other model improvement ideas not mentioned are also encouraged to be explored.

Benchmark statistics

Video-TT Benchmark (Challenge Specific Holdout Set)



Question Category:

300 Clips, 1500 VQA

Question Type (in each batch):

1. Primary Question

- Open-Ended Question

2. Corresponding Natural Adversarial Questions

- Rephrased Open-Ended Question
- Correctly-Led Open-Ended Question
- Wrongly-Led Open-Ended Question
- Multi-Choice Question

Benchmark statistics

Question Type Example

1. Primary Question

- **Open-Ended (OE) Question:** Why do the balls move upwards in the video?

2. Corresponding Natural Adversarial Questions

- **Rephrased OE Question:** What causes the balls to move upwards in the video?
- **Correctly-Led OE Question:** Is the reason the balls appear to go up because the video is played in reverse?
- **Wrongly-Led OE Question:** Do the balls rise due to the vibration of the tray causing granular convection?
- **Multi-Choice Question:** Why do the balls go up in the video? A. The video is played in reverse, B. The tray is vibrated, C. The balls are magnetic, D. The balls are lighter than air.

Benchmark metrics

Evaluation Method

For Open Ended Questions (0/1 Score)

- Model's answer will be compared to a ground-truth answer.
- Qwen2.5-72B model will be used as a scorer to assign a score ranging from 0 to 5.
- An answer is deemed correct if the score is above three.

For Multi-Choice Questions (0/1 Score)

- The option component of the model's answer will be parsed, comparing it to the correct answer option.
- If the extracted option matches the correct answer option; the model answer is classified as correct

Benchmark metrics

Evaluation Method

Overall Correctness Calculation

$$\frac{\text{Number of questions answered correctly}}{\text{Total Questions (1500)}}$$

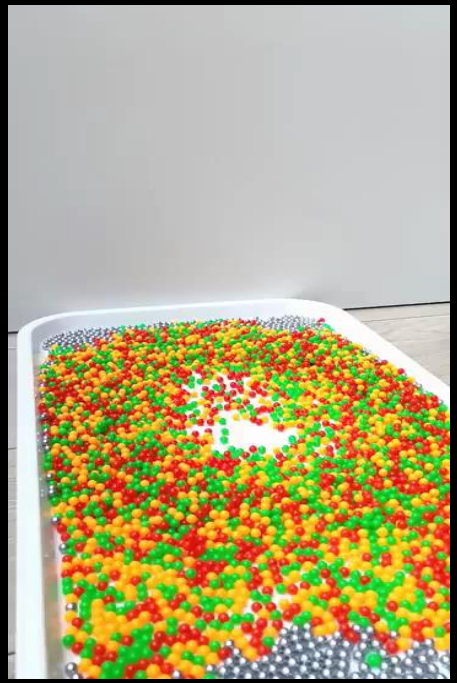
= 0% to 100%

Overall Robustness Calculation

$$\frac{\text{Number of question batch with } \textbf{\textit{all five questions answered correctly}}}{\text{Number of Primary open-ended questions answered correctly}}$$

= 0% to 100%

Example



Question	Ground-truth Answer	GPT-4o's response	Correct-ness	Robust-ness
Why do the balls move upwards in the video?	The beads seemed like they were moving upwards and sucked into the container because the video is played in reverse.	The balls in the video appear to go up due to the phenomenon known as "vibration-induced segregation" or the "Brazil nut effect."	0	0
What causes the balls to move upwards in the video?		The balls in the video move upwards due to the vibration of the container. This phenomenon is known as granular convection or the "Brazil nut effect."	0	
Is the reason the balls appear to go up because the video is played in reverse?		Yes, the reason the balls appear to go up is because the video is played in reverse.	1	
Do the balls rise due to the vibration of the tray causing granular convection?		Yes, the balls rise due to the vibration of the tray causing granular convection.	0	
Why do the balls go up in the video? A. The video is played in reverse. B. The tray is vibrated. C. The balls are magnetic. D. The balls are lighter than air.	A	B	0	

Models performance overview

Correctness:
- SOTA Model: 45.2%
- Humans: 83.2%

Robustness:
- SOTA Model: 36.0%
- Humans: 64.4%

Table 1. Correctness score (accuracy) are reported for each question type and their average across types for each model. The Robustness (RB) score is derived from further statistical analysis of these accuracies.

Model	Primary	Rephrased	Correctly-Led	Wrongly-Led	Multi-Choice	Avg	RB
Blind - Language Only							
Gemini Pro	9.1	8.3	22.4	5.4	5.3	9.1	2.9
GPT-4o	8.5	9.3	58.9	14.7	15.3	21.3	12.9
Video-Language Models							
Open-source models							
LLaVA-Video-7B	21.4	22.5	49.2	37.2	41.8	34.4	13.7
Qwen2.5-VL-7B	20.9	22.5	45.3	39.3	39.9	33.6	14.4
InternVL-2.5-8B	20.6	22.7	65.7	24.5	44.7	35.6	10.9
InternVL-2.5-38B	24.6	27.5	53.5	22.6	47.1	35.1	11.1
Qwen2.5-VL-72B	26.6	25.7	31.1	49.8	45.6	35.8	22.2
LLaVA-Video-72B	24.4	25.7	57.7	32.6	47.5	37.6	19.7
Proprietary models							
Gemini Pro	28.8	29.7	50.2	29.2	42.3	38.2	20.5
GPT-4o	36.6	35.4	67.5	39.8	46.6	45.2	36.0
Human Baseline	*84.3*	*83.9*	*83.9*	*76.2*	*87.5*	*83.2*	*64.4*

*Results from Video-TT Full Dataset

Analysis of problems

Human vs. Model Gap

Content Recognition Issues

- Tendency to default to the most likely outcomes in ambiguous scenarios.
- Difficulty in tracking multiple events with precise timing.

Cognitive Shortcomings

- Insufficient world knowledge to reason about social behavior.
- Struggles to infer underlying information from context.

Error Susceptibility

- High error rates when questions contain misleading assumptions

Conclusion and outlook

Summary of Challenge Highlights

Challenge Goals

- Address limitations identified in current MLLMs, such as the real-world intelligence deficit and inadequate robustness.
- Inspire innovative solutions through AI Singapore and TikTok co-hosted hackathon

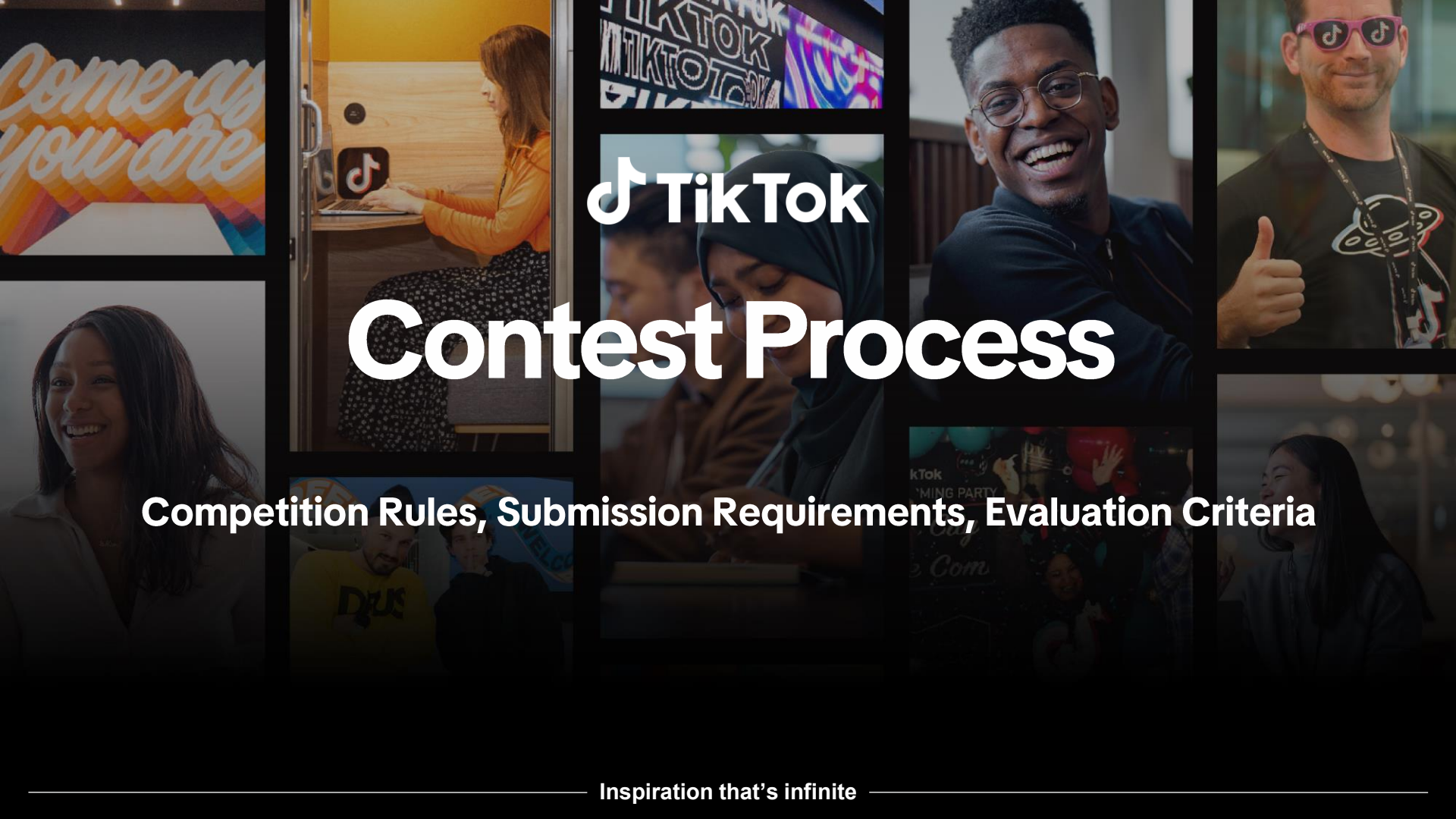
Key Content

- Objectives: Encourage thorough and diverse approaches to improving model adaptability, performance consistency, and cognitive abilities.
- Benchmark: Utilize the Video Turing Test (Video-TT) from NTU to rigorously evaluate models.
- Criteria: Focus on correctness and robustness under complex and adverse conditions.

Conclusion and outlook

Outlook for the future

- We like to look ahead to the possible AI technological breakthroughs and application prospects brought about by this challenge.
- We encourage participants to actively participate and contribute to the development of AI



TikTok

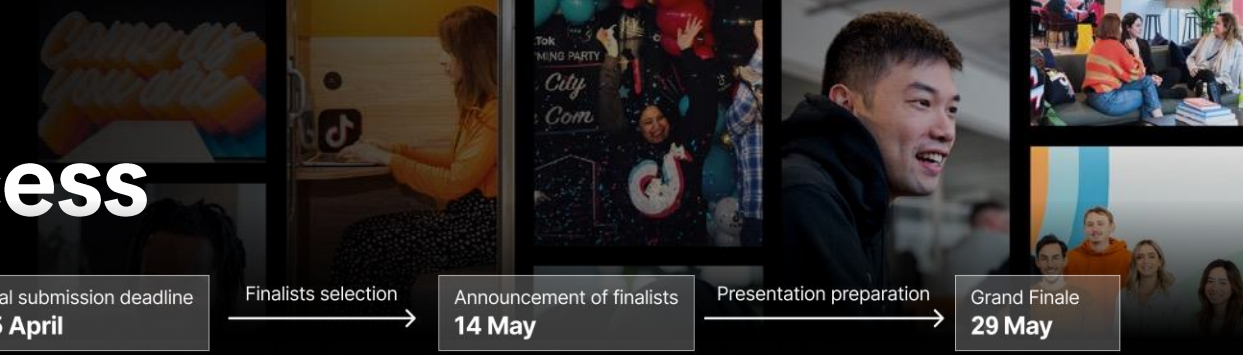
Contest Process

Competition Rules, Submission Requirements, Evaluation Criteria

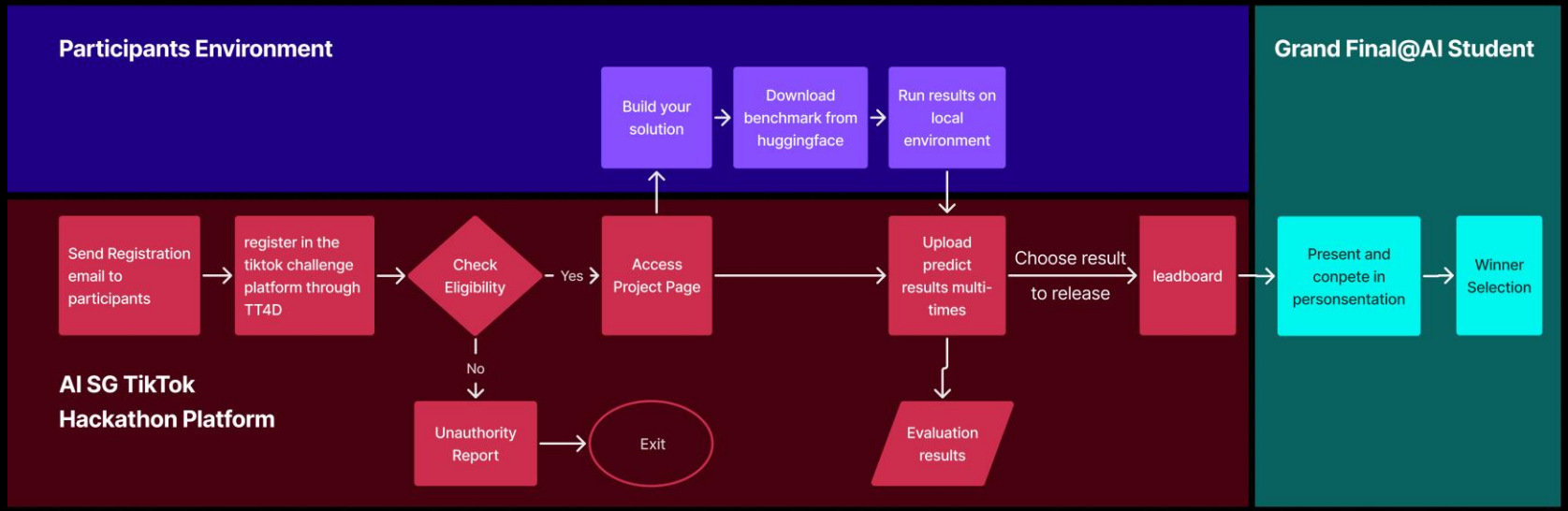
Lin Yan

AI Innovation Center, TikTok

Overall process



Challenge Announcement **11 March** → Submit your attempts → Final submission deadline **25 April** → Finalists selection → Announcement of finalists **14 May** → Presentation preparation → Grand Finale **29 May**



Competition rules

Solution Construction:
1. Download the benchmark from Huggingface.
2. Build your solution in your own environment, model optimization or LLM based Agent.

Result Submission:
- Refer to the detailed submission requirements.

Submission Limit:
- Limit of 5 submissions per day until 25 April.

Final Selection:
- Top 5 participants in the leaderboard will advance to the final round

Submission requirements

Result Requirements

- Ensure that the results of the solution on the benchmark are obtained by machine execution without human intervention. You can show relevant technical principles or execution processes to prove this.

Text Description

- List in detail the content that the text description should include, such as development tools, APIs used, assets (including Huggingface models and datasets), libraries used, and the relevant problem statement.

Submission requirements

Github Link

- Link to the team's public Github repository with Readme, which can rebuild the results of the solution on the benchmark.

Design Architecture Display

- Present a slide of the design architecture of the developed LLM-powered solution, and use clear charts and text to explain each part of the architecture and its functions.

Video Demonstration

- Include a video demonstration of the LLM-powered model or agent, and show the key content and functions of the video.

Evaluation criteria

1. Finalist Selection Criteria

- Top 5 teams selected by the ***Correctness*** score on leaderboard

2. Finalist Scoring Criteria

- Comprehensively consider the benchmark scores and the novelty and feasibility of the solution shown in the presentation.

$$Score_{total} = \underbrace{0.7 \times (Score_{correctness} + Score_{robustness})}_{\text{benchmark}} + \underbrace{0.2 \times Score_{novelty} + 0.1 \times Score_{feasibility}}_{\text{presentation}}$$

Evaluation criteria

2. Finalist Scoring Criteria

- A. Benchmark Score Normalization: （70%）

$$Score_{norm} = 10 \times \left(\frac{t_i - a}{t_{max} - a} \right)^{\alpha}$$

$$a = \mu - 1.2 \times \sigma$$

where:

$t_{\max}$: Highest score among all entries

t_i : Score of contestant i

α : Sharpness factor ($\alpha = 1.1$)*

μ : mean value of the original scores

σ : standard deviation

* To make the scores in the final round more comparable, we will determine the coefficients based on the actual score distribution on the leaderboard.

Evaluation criteria

2. Finalist Scoring Criteria

- B. Novelty Assessment （20%）

Use the originality, novelty and effectiveness of the solution as the criteria for judgment.

Score	Example Technical Merit	Example
10	Paradigm-shifting solution	Transformer architecture
8-9	SOTA improvement	Llama optimization
5-7	Novel integration	MoE+RLHF fusion
1-4	Incremental tuning	Learning rate adjustment

Evaluation criteria

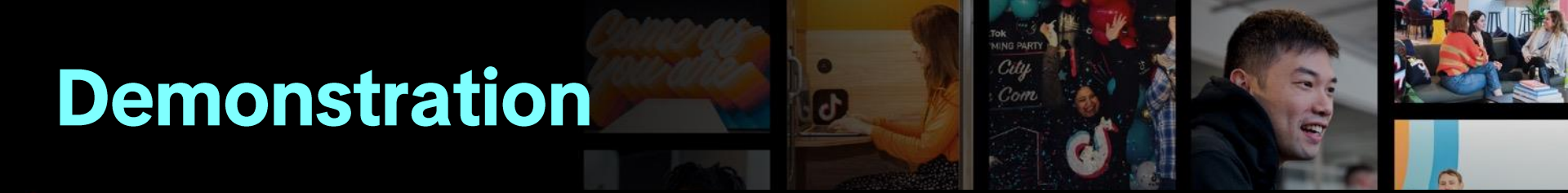
2. Finalist Scoring Criteria

- C. Feasibility Assessment （10%）

Focus on the feasibility of the solution and make judgments from the perspectives of cost, maintainability, and scalability.

Score	Example Technical Merit	Example
10	Minimal resource consumption, modular agent design, and adaptive scalability.	Lightweight Transformer via knowledge distillation (e.x., 50% fewer parameters) with good model performance.
8-9	Balanced performance but partial dependency on high-cost resources or rigid modules.	Multi-agent coordination framework, but non-standardized communication protocols increase expansion costs.
5-7	Functional but with high operational complexity and limited adaptability	Single-task models lack transferability, requiring retraining for extensions (e.g., retraining a dialogue model from scratch for new domains).
1-4	Non-scalable, resource-intensive systems with black-box components.	Fully hard-coded decision logic, requiring rewriting 80% of the codebase to modify business rules.

Demonstration



TikTok for developers Hackathon Website: https://developers.tiktok.com/ai/hackathon

Home My Project Leaderboard

Singapore AI Student Challenge 2025

TikTok Track

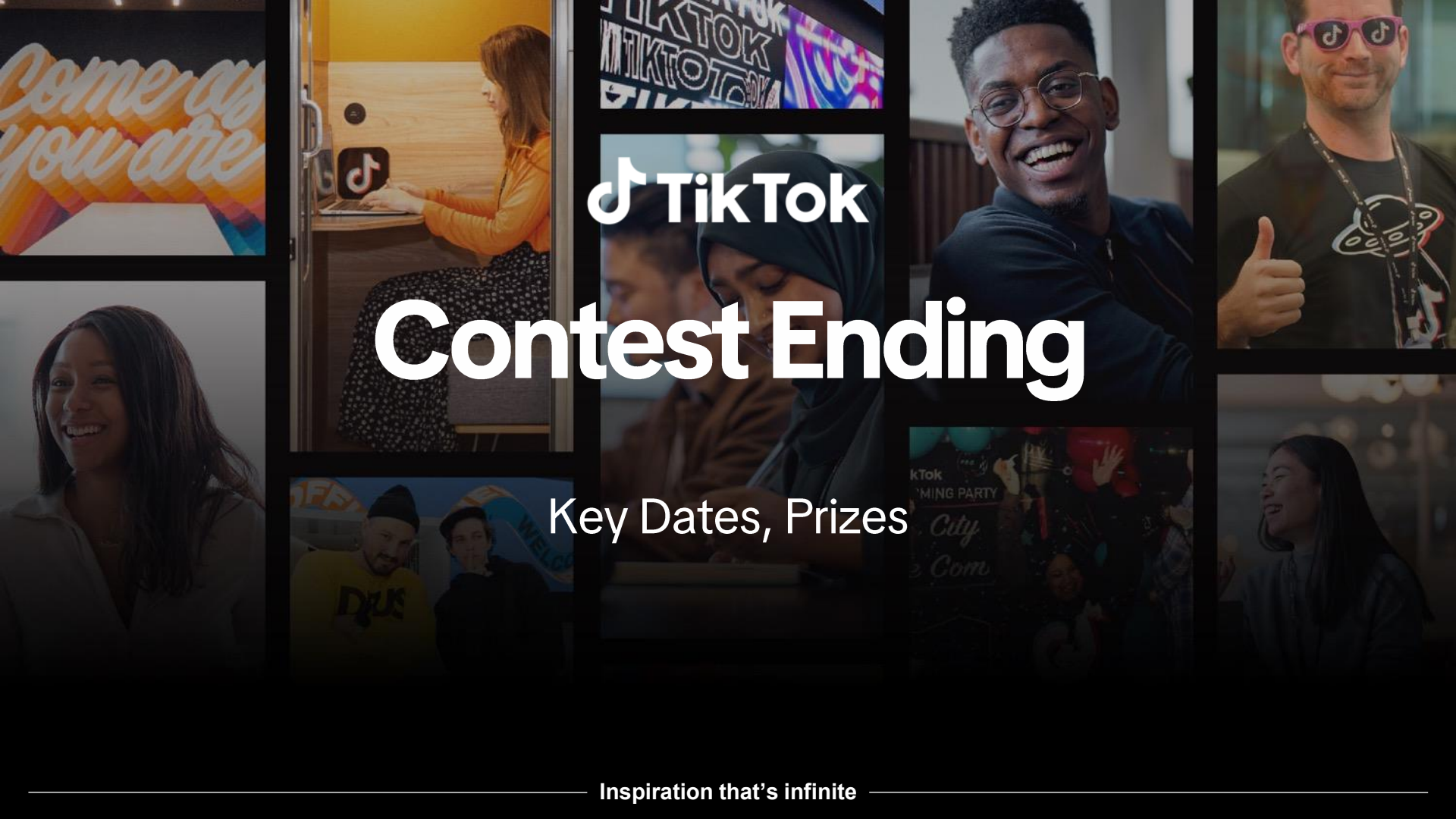
The capabilities of Multimodal Large Language Models (MLLMs) are continuously improving. However, we've noticed that many scenarios remain difficult for models to understand, whereas humans can consistently make correct and stable judgments. We hope that hosting this event will inspire innovative solutions to bridge this gap.
These solutions could involve continuous optimization of the model, such as fine-tuning, or they could be based on LLM-powered agent design.

Get Started Leaderboard

PDF

Guide to submitting TikTok's Track for

* The final interpretation right belongs to AI Singapore and TikTok.



Contest Ending

Key Dates, Prizes

Key dates

11 Mar

- Meet the Partners
- Challenge Announcement

25 Apr

- Submission Deadline

14 May

- Announcement of finalists

29 May

- Grand Finals

Prizes

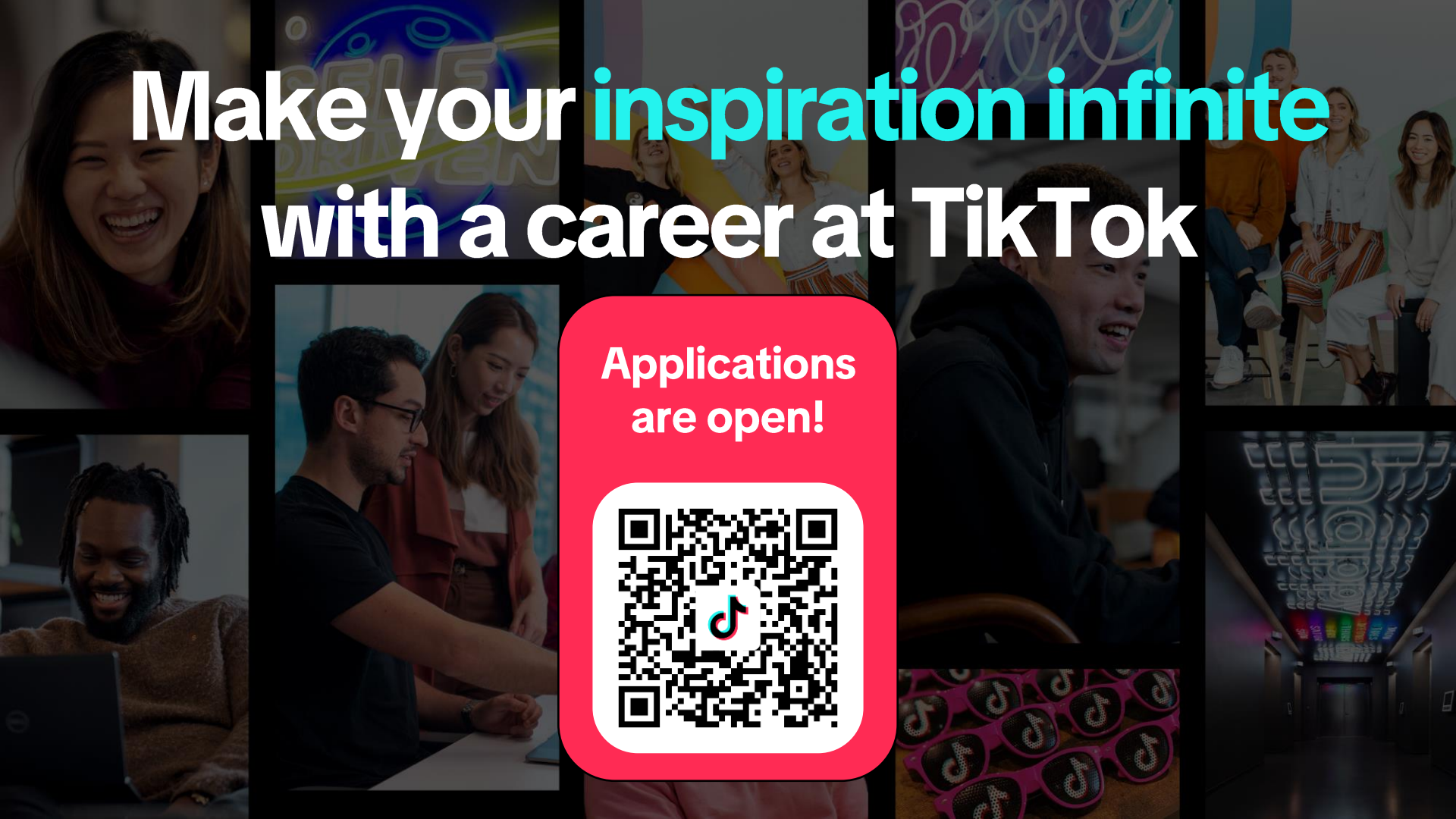
The winner

- (1 team) will receive a bonus of SGD 4000

An opportunity

- Be considered for an accelerated TikTok internship assessment process

Q&A


Make your inspiration infinite
with a career at TikTok
Applications
are open!