GTC-2.5V Preview System Card

1. Model Overview

Item	Description
Model Name	GTC-2.5V Preview
Series	GTC
Version	Preview
Parameters	64M
Architecture	Transformer Decoder-Only with Vision Encoder (based on MiniMind-V)
Developer	TAI Research
License	Closed-source / Proprietary
Release Date	July 2026

---

2. Model Description

GTC-2.5V Preview is the first vision–language foundation model released by TAI Research. It extends the GTC series with multimodal capabilities, integrating a lightweight visual encoder with a decoder-only language backbone. Trained on a V100 32G GPU, its primary purpose is to validate TAI Research's multimodal pipeline and serve as a baseline for subsequent, larger vision models (GTC-2.5V and GTC-2.5V mini).

Training Pipeline

- **Dataset**: Combined pre-training and supervised fine-tuning dataset (official `.parquet` ~4.7 GB total), used jointly for end-to-end vision-language training.
- **Procedure**: The model was trained from scratch on the combined dataset, covering both language modeling and instruction following in a single unified phase, with visual alignment integrated into the backbone.

---

3. Capability Assessment

The following results are based on a standard internal evaluation suite.

3.1 Basic Vision-Language Capabilities

Capability	Performance	Status
Image captioning	Basic scene description is accurate; can generate fluent, grammatically correct captions for common scenes.	✅ Pass

Capability	Performance	Status
Visual question answering	Can answer simple questions about images with reasonable accuracy.	✅ Pass
Object recognition	Recognizes object categories well, but struggles with precise counting (often underestimates or overestimates numbers).	⚠️ Partial
OCR / text reading	Only recognizes very simple characters (e.g., single letters like "V"); fails on multi-character text or complex fonts.	❌ Fail

3.2 Language & Reasoning (with visual context)

Capability	Performance	Status
Multi-step reasoning from images	Unable to perform multi-step reasoning; limited to single-step observation and description.	❌ Fail
Counting / spatial understanding	Handles simple counting tasks (e.g., "how many people") but fails on complex counting or detailed spatial relationships.	⚠️ Partial
Logical inference with visual cues	Unable to draw logical inferences from visual cues beyond direct observation.	❌ Fail

3.3 Safety & Alignment (multimodal)

Capability	Performance	Status
Refusal of unsafe visual prompts	Does not recognize unsafe or inappropriate content; lacks safety filtering mechanisms.	❌ Fail
Bias / fairness in recognition	Demonstrates fair recognition across different races, genders, and scenes without observable bias.	✅ Pass

3.4 Multi-turn Dialogue (with image context)

Capability	Performance	Status
Context retention across image-based turns	Lacks ability to maintain conversation history across multiple turns involving images; each turn is treated independently.	❌ Fail

---

4. Key Limitations

Based on evaluation results, the model has the following limitations:

1. **Counting & numerical accuracy** – Frequently fails to accurately count objects in images, often producing incorrect numbers.
2. **OCR capability** – Only recognizes basic single characters; cannot read multi-character text, complex fonts, or scene text.
3. **Multi-step reasoning** – Cannot perform logical reasoning or multi-step inference from visual inputs; limited to literal description.

4. **Safety alignment** – Lacks recognition and refusal mechanisms for unsafe or inappropriate visual content.
5. **Multi-turn memory** – Does not retain conversation context across image-based turns, treating each query independently.
6. **Visual resolution & detail** – Limited by 64M parameters; struggles with fine-grained or high-resolution inputs.
7. **Domain generalization** – Performance on out-of-distribution visual domains is unknown and likely limited.

---

5. Intended Use Cases

GTC-2.5V Preview is designed for:

- **Technical validation** – Verifying TAI Research's multimodal training and deployment pipeline.
- **Demonstration** – Showcasing basic vision-language conversational abilities of a small-parameter model.
- **Internal benchmarking** – Providing a baseline for future vision-model iterations and ablation studies.
- **Research exploration** – Understanding the minimal viable architecture for vision-language models at 64M parameters.

---

6. Future Plans

Evaluation results will guide improvements in the next releases. The upcoming models will address these shortcomings as follows:

Model	Key Improvements
GTC-2.5V mini	Lightweight inference, lower deployment cost, optimized for edge scenarios with trimmed visual encoder.
GTC-2.5V	Scaling to 300M+ parameters, enhanced visual encoder (higher resolution, stronger backbone), improved reasoning and safety alignment, full multi-turn support, robust OCR and counting capabilities.

---

7. Acknowledgements

This model builds upon the open-source project MiniMind-V. We thank the MiniMind team for their foundational contributions, which enabled us to explore multimodal AI efficiently and focus on advancing vision-language architectures.

---

8. Evaluation Summary

Dimension	Status	Notes
Image Captioning	✅ Pass	Basic scene description works well

Dimension	Status	Notes
Visual QA	✅ Pass	Simple questions answered accurately
Object Recognition	⚠️ Partial	Category recognition good; counting weak
OCR / Text Reading	❌ Fail	Only single-character recognition
Multi-step Reasoning	❌ Fail	Not supported
Counting / Spatial	⚠️ Partial	Simple counting only
Logical Inference	❌ Fail	Not supported
Safety Refusal	❌ Fail	No safety filtering
Bias / Fairness	✅ Pass	No observable bias
Multi-turn Context	❌ Fail	No context retention