LREC 2026 Palma

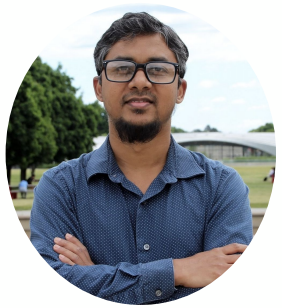
QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

YORK UNIVERSITÉ UNIVERSITY

# Multilingual and Multimodal LLMs in the Wild: Building for Low-Resource Languages


**Firoj Alam**
Senior Scientist, QCRI

**Shammur Chowdhury**
Scientist, QCRI

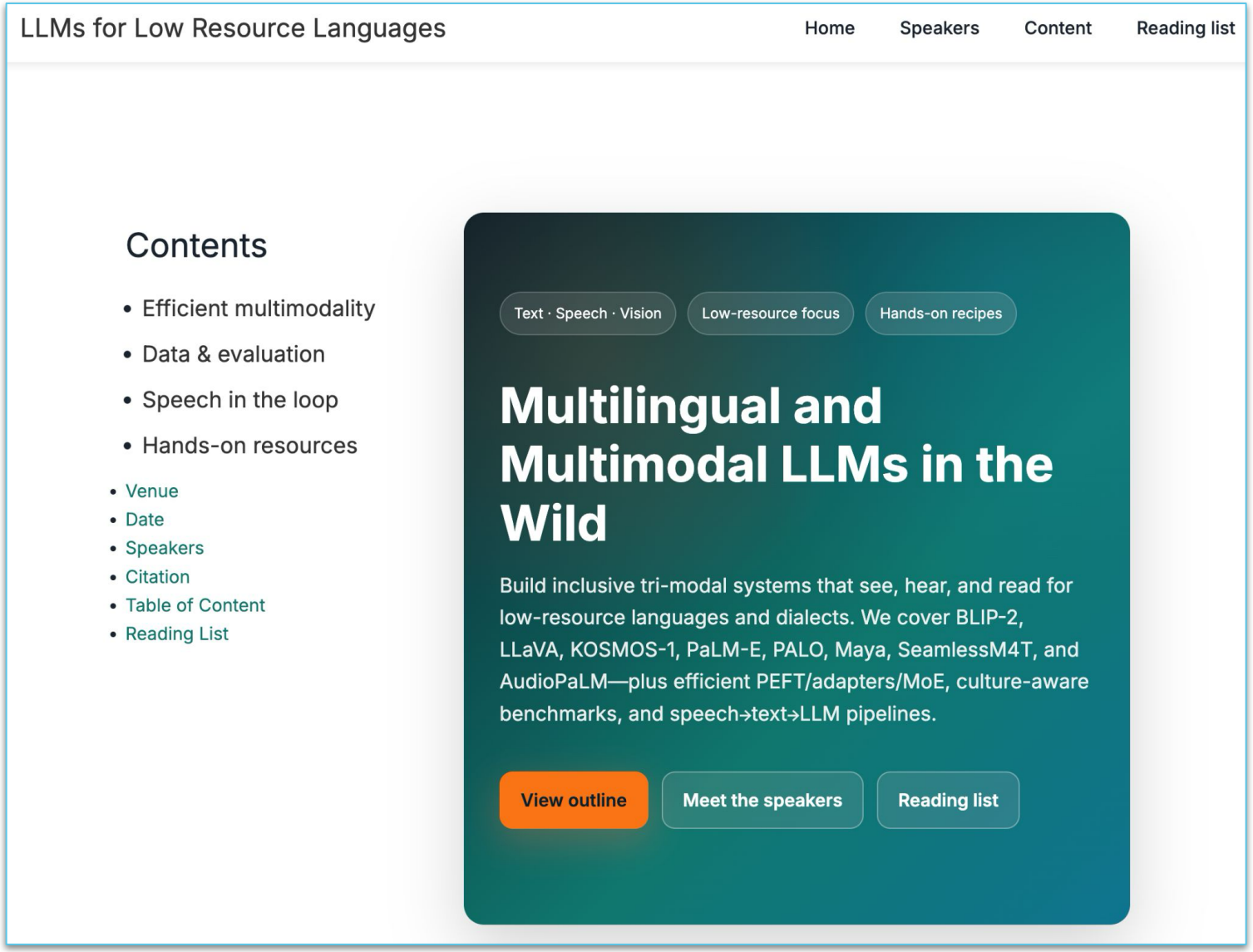


**Enamul Hoque Prince**
Associate Professor,
York University



**LREC 2026, 12th May, 2026**

# Tutorial Resources




https://mm-llms-in-the-wild.github.io

# Tutorial Outlines

**Introduction**
(low-resource and challenges)

**Multilingual and Multimodal Models**
(Capabilities for low-resource languages)

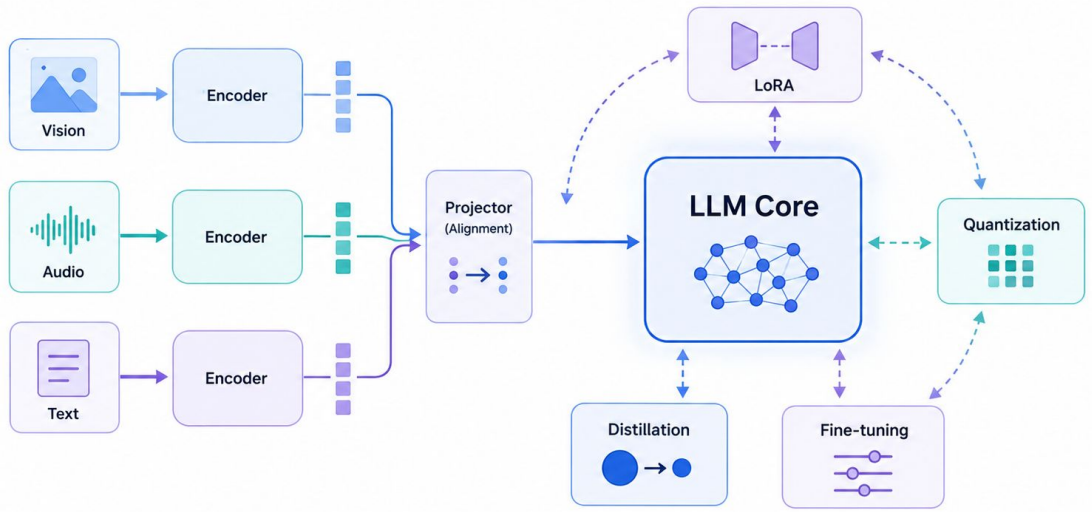
**Multilingual & Multimodal Resource Development**
(Low-cost pipeline, training, and benchmarking datasets)

**Architectures and Efficient Training**
(Adapter-based and parameter-efficient methods)

**Speech-centric LLMs**
(State-of-the-art models and curated resources)

**Evaluation, Benchmarking Tools, Analysis**
(Evolution of Evaluation metrics, benchmarking tools and resources)

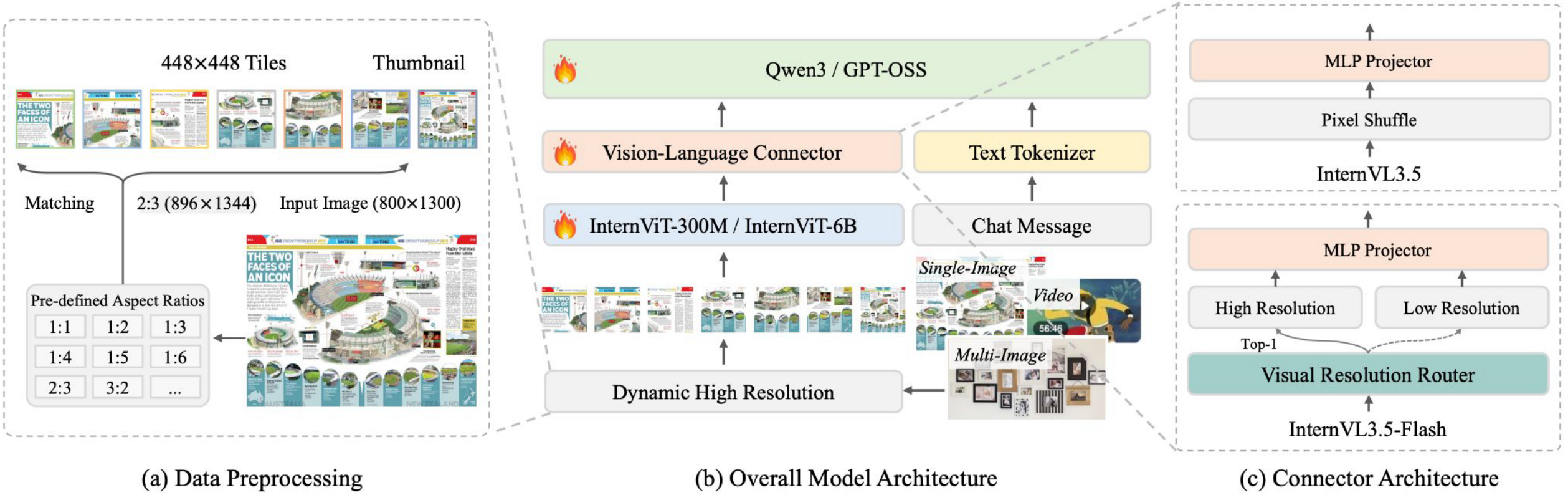
# Architectures & Efficient Training



# InternVL3.5

https://huggingface.co/OpenGVLab/InternVL3_5-8B

- InternVL3.5 adopts the “ViT–MLP–LLM” paradigm



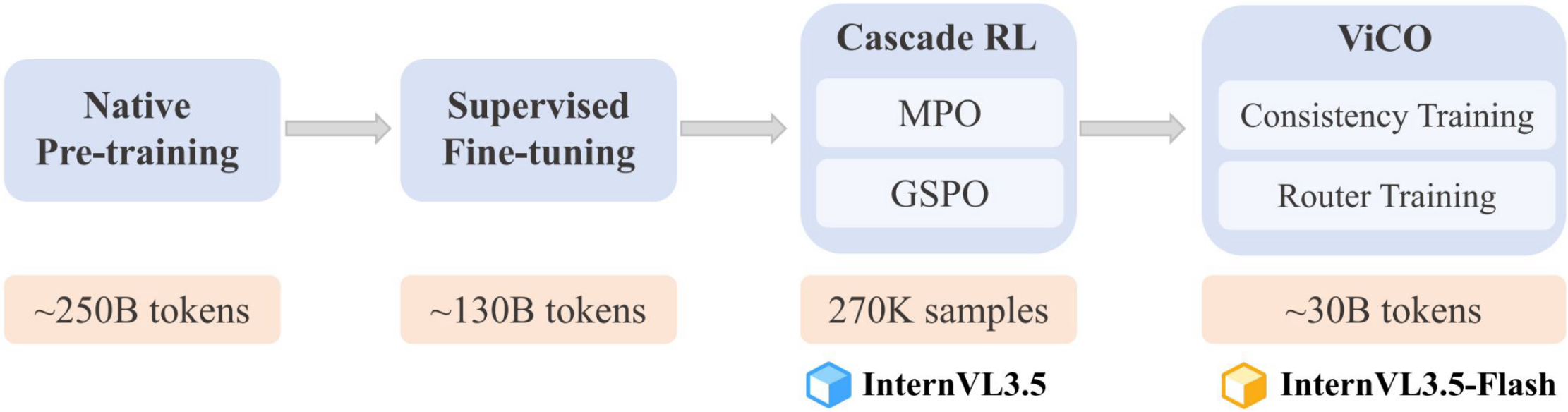

(a) Data Preprocessing (b) Overall Model Architecture (c) Connector Architecture

Wang et al., InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency, 2025

# InternVL3.5

https://huggingface.co/OpenGVLab/InternVL3_5-8B

- InternVL3.5 adopts the “ViT–MLP–LLM” paradigm




Wang et al., InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency, 2025

# Aya Vision

https://huggingface.co/docs/transformers/en/model_doc/aya_vision

- **A balanced mix matters** ; the final recipe uses 35% multilingual multimodal data and 65% English data for cross-lingual transfer.
- **Cross-modal model merging** helps recover text-only performance while improving multimodal generation.
  - Add vision while reducing catastrophic forgetting of text-only multilingual capabilities.
- **Data improvements produced the largest gain** : better recaptioning, filtering, and translation quality improved multilingual vision win rates substantially.

Dash et al. Aya Vision: Advancing the Frontier of Multilingual Multimodality, 2025

# Aya Vision: Training

https://huggingface.co/docs/transformers/en/model_doc/aya_vision

## Late-Fusion Vision-Language Design

Vision Encoder — Vision-Language Connector — Large Language Model (LLM)

### Vision Encoder: SigLIP2-SO400M

- **Aya-Vision-8B:** patch14-384
- **Aya-Vision-32B:** patch16-512

### Language Model

- **8B:** Command-R7B-based model post-trained with Aya Expanse recipe.
- **32B:** Aya-Expanse-32B.

### Image Processing

Resize to supported resolution, split into up to 12 tiles, plus thumbnail.

### Connector

2-layer MLP with SwiGLU activation.

### Token Reduction

Pixel Shuffle downsamples image tokens by 4x.

Dash et al. Aya Vision: Advancing the Frontier of Multilingual Multimodality, 2025

# Aya Vision: Training

https://huggingface.co/docs/transformers/en/model_doc/aya_vision

## 1. Training Process

**Vision-Language Alignment**

- Train only vision-language connector.
- Vision encoder & LLM are **frozen**.
- **Data:** LLaVA-Pretrain + 14% multilingual data.
- Steps: 9.7k (8B); 19k (32B).

## 2. Visual Instruction SFT

- **Train:** Connector and LLM.
- **Frozen:** Vision encoder remains frozen.
- **Batch Size:** 128.
- **Training Length:** 31k iterations.
- **Optimization:** Sequence packing to length 8192.

## 3. Cross-Modal Merging

- **Training-free** merge after multimodal training.
- Interpolates text-only and multimodal LLM weights.
- Vision encoder and connector are inherited from the multimodal model.

Dash et al. Aya Vision: Advancing the Frontier of Multilingual Multimodality, 2025

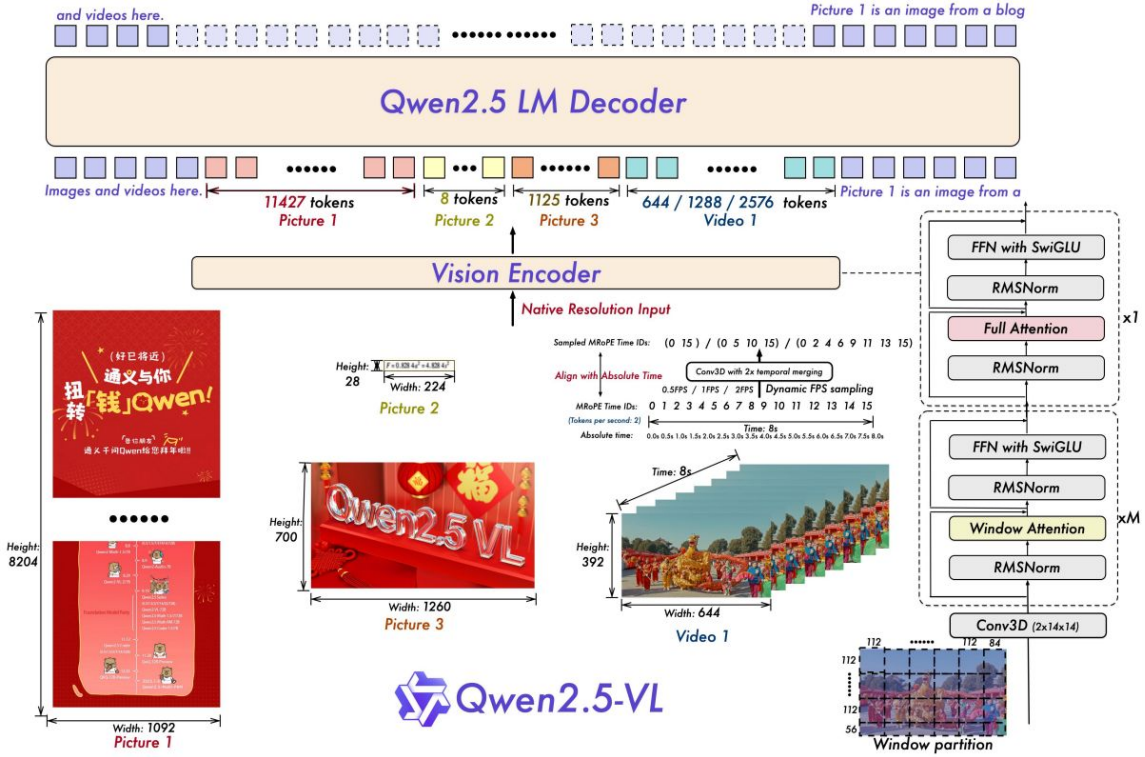
# Fanar-Oryx-V1: Training

https://api.fanar.qa/docs

**Base model: Qwen2.5-VL**

**Type:** Decoder model
**Size:** 7B

- **Vision Encoder:** Dynamic-resolution ViT with *window attention*
- **Fusion Module:** Projects vision tokens into LLM embedding space
- **Positional Encoding:** M-RoPE (Multimodal Rotary Positional Embeddings) for spatial & temporal alignment




**Input Support:** Images + videos + documents (layout/structured content)
**Licence:** Apache 2.0



FANAR TEAM et al., Fanar 2.0: Arabic Generative AI Stack, 2026

# Fanar-Oryx-V1: Training Recipe

https://api.fanar.qa/docs

**Pre-training (Qwen2.5-VL-7B)**

- **Objective:** Multimodal pre-training to align **vision, language, and layout** understanding.
- **Scale:** ~4.1T tokens (mixed text–image–video corpus) → 1.2T in QWEN2
- **Data composition:**
  - Interleaved **image–text pairs** (captioning, reasoning, OCR).
  - **OCR:** French, German, Italian, Spanish, Portuguese, **Arabic**, Russian, Japanese, Korean, and Vietnamese
  - **Video clips** with temporal annotation for event grounding.
  - **Document layouts** (forms, tables, charts, handwritten content).
  - **Object-grounding datasets** with bounding boxes / point annotations.



FANAR TEAM et al., Fanar 2.0: Arabic Generative AI Stack, 2026

# Fanar-Oryx-V1: Training Recipe

https://api.fanar.qa/docs

## Data

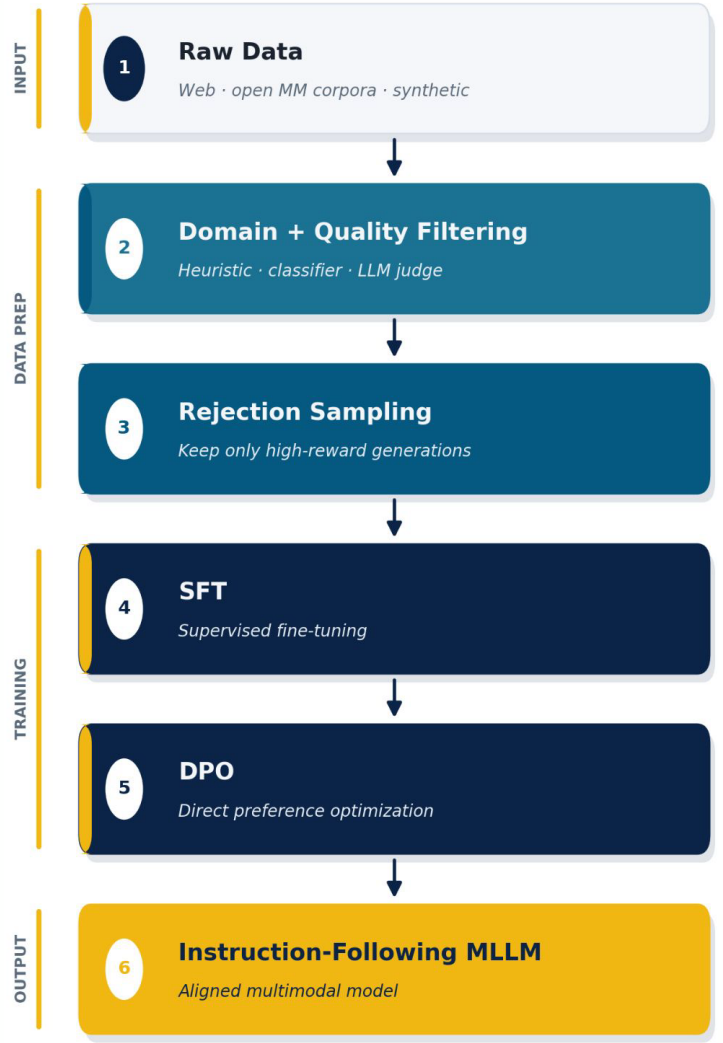
- ~2M SFT samples
- 50% text, 50% multimodal
- Chinese/English-centered, with multilingual coverage

**Domains:** VQA, Captioning, OCR/Docs, Grounding, Video, Coding, Math, Agent Tasks

## Key Choices

- Two-stage data filtering for quality and safety
- Correct reasoning examples retained for stronger CoT
- ViT kept frozen for efficiency and stability

**Better instruction following, cross-modal reasoning, and reliable responses.**






FANAR TEAM et al., Fanar 2.0: Arabic Generative AI Stack, 2026

# Other Recent Models Developed in Parallel

**PALO: A Polyglot Large Multimodal Model for 5B People**

https://arxiv.org/abs/2402.14818

- LLaVA-style vision-language architecture
- **Vision encoder:** CLIP ViT-L/336px
- **Language backbones:**
  - PALO-7B and PALO-13B use Vicuna.
  - MobilePALO-1.7B uses MobileLLaMA.
- LoRA fine-tuned

**AIN - The Arabic INclusive Large Multimodal Model**

https://arxiv.org/pdf/2502.00094 – Based on Qwen-2-VL-7B architecture

**Maya - An Instruction Finetuned Multilingual Multimodal Model**

https://arxiv.org/pdf/2412.07112

- Architecture is based on LLaVA 1.5.

# QA