
# Open Source LLM Speculation (LLaVA-OneVision-1.5 (4B or 8B)) BBS Project

Kazi Islam

Version 1.0

Prepared for Prof David Smith

Date [10-15-2025]

# Page of Content

# 1. Abstract

This document presents the exploratory phase of developing an open-source image-and-text recognition framework for the Balanced Blended Space (BBS) puppet show, focusing specifically on evaluating LLaVA-OneVision-1.5 (4B or 8B) as a candidate vision-language model. The goal is to determine whether this model can effectively learn from paired character and prop images with descriptive captions to recognize and label new photos in the show's dataset. As part of the speculative and discovery stage, this study examines LLaVA-OneVision-1.5 (4B or 8B) model's architecture, permissive licensing, fine-tuning feasibility, and multimodal reasoning capabilities. We assess its suitability for small-to-medium creative projects by testing its responsiveness to visual prompts, text alignment accuracy, and generalization to unseen scenes. The findings will inform a practical foundation for integrating open-source large language models (LLMs) into artistic and educational media contexts, guiding future development of a lightweight, reproducible recognition model tailored to the BBS project's storytelling environment.

# 2. Objective

## 2.1 Primary objective:

Determine whether LLaVA-OneVision-1.5 (4B and 8B, Instruct) is a suitable, low-cost, fully open-source model for BBS image-and-text recognition, and select the checkpoint (4B vs 8B) that best balances accuracy, latency, and hardware cost for deployment.

### 2.1.2. Specific objectives:

1. **Model Selection (4B vs 8B).**
   Compare accuracy vs. efficiency (VRAM, latency, throughput) to recommend 4B (speed/cost) or 8B (higher capacity) for the final pipeline.

2. **Cost & Resource Envelope.**
   Record hardware requirements and run-time costs on commodity GPUs/CPUs; verify that the model's **native-resolution training and optimized MegatronLM stack** translate into practical, budget-aware inference/fine-tuning for a small team.

3. **Openness & Compliance.**
   Confirm Apache-2.0 licensing and the availability of fully open assets (datasets used for SFT/mid-training, code, configs) to ensure the pipeline is legally and operationally reusable in educational/creative contexts.

# 3. Background

LLaVA-OneVision-1.5 is a fully open-source family of large multimodal models (LMMs) designed to read images and text together. A key design choice is training on native-resolution images, which improves fine detail recognition without upscaling/downscaling artifacts. The release includes a complete, reproducible setup which is curated mid-training and instruction-tuning datasets, training recipes/configurations, checkpoints, and logs, so teams can study, adapt, and re-train the models end-to-end. The family emphasizes state-of-the-art accuracy at substantially lower cost, enabled by an efficient training stack.

## 3.1. Common Features of Both Models

- **Data quality & coverage:** Rigorously filtered, concept-balanced caption data and broad instruction-tuning tasks; high data efficiency (on the order of ~64B tokens reported for pretraining stages).
- **Training efficiency:** An optimized pipeline targeting cost-effective scaling and practical fine-tuning on modest hardware budgets.
- **Openness:** Permissive licensing and full transparency (code, configs, checkpoints, and metrics) to enable verification and reuse.
- **Performance focus:** Strong results across diverse multimodal benchmarks; the series is frequently reported to outperform competing open models of similar size on many evaluations.

# 4. Evaluation of Both Models

| Criterion | 4B-Instruct | 8B-Instruct | What it means for BBS |
|---|---|---|---|
| **Required hardware (minimum storage for weights)** | **≈ 9.6 GB** for model files. Weights-only VRAM footprint ~8–10 GB; practical runtime usually ≥10–12 GB at FP16 (lower with 8-/4-bit quantization). (Hugging Face) | **≈ 17.2 GB** for model files. Weights-only VRAM footprint ~16–18 GB; practical runtime often ≥18–24 GB at FP16 (lower with 8-/4-bit quantization). (Hugging Face) | 4B fits easier on a single consumer GPU; 8B may require a higher-VRAM card or quantization. |

| | | | |
|---|---|---|---|
| **Openness** | Permissive license with openly released code/recipes/checkpoints/logs for full reproducibility. (Hugging Face) | Same openness guarantees and artifacts. (Hugging Face) | Both are safe choices for an academic/creative pipeline and redistribution of configs/results. |
| **Efficiency** | Family emphasizes native-resolution training, optimized stack (MegatronLM; FP8/long-sequence; MoE options), and a **reported ≈ $16K** full-training budget. (Hugging Face) | Same framework; larger capacity trades some runtime efficiency for accuracy headroom. (arXiv) | Expect lower inference cost/latency with 4B; use 8B when accuracy gains justify extra compute. |
| **Accuracy (benchmarks)** | Competitive across diverse multimodal tasks; within the series, 4B trails 8B but is strong for many use cases. (arXiv) | Frequently **outperforms** smaller variants and is reported to beat peer open models on many evaluations. (arXiv) | For BBS recognition, start with 4B; move to 8B if complex scenes/text-image reasoning need higher accuracy. |
| **Latency** | **Lower** (faster responses) due to smaller parameter count and memory movement. | **Higher** (slower) relative to 4B under the same hardware/settings. | 4B is better for real-time-ish previews; 8B suits offline/batch labeling or when you can spare a few extra seconds. |

## 4.2. The two variants and when to use them

### 4.2.1. LLaVA-OneVision-1.5-4B (Instruct).

- **Profile:** Smaller parameter count for **lower VRAM**, faster inference, and easier deployment.
- **Best for:** real-time or near-real-time use, laptops/workstations with limited GPU memory, and wider device coverage.

- **BBS fit:** a practical default for on-device or single-GPU prototyping; solid for most single-character/prop shots and simple stage scenes.

### 4.2.2. LLaVA-OneVision-1.5-8B (Instruct).

- **Profile:** Larger capacity for **richer visual reasoning** and more robust text-image alignment on dense scenes.
- **Best for:** complex frames (multiple puppets/props, signage, occlusion), tougher caption grounding, and higher accuracy targets.
- **BBS fit:** the “headroom” option when incremental accuracy outweighs extra compute/latency.

Our Recommendation is using LLaVA-OneVision-1.5-4B for BBS and then moving to the LLaVA-OneVision-1.5-8B model if necessary.

# 5. References


1. lmms-lab. “LLaVA-OneVision-1.5-4B-Instruct — Model Card.” Hugging Face, 2025. Hugging Face+1
2. lmms-lab. “LLaVA-OneVision-1.5-8B-Instruct — Model Card.” Hugging Face, 2025. Hugging Face+1
3. An, X., et al. “LLaVA-OneVision-1.5: Fully Open Framework for Democratized Multimodal Training.” arXiv, 2025. arXiv+1
4. EvolvingLMMs-Lab. “LLaVA-OneVision-1.5 — Code Repository.” GitHub, 2025. GitHub+1
5. Hugging Face Papers. “LLaVA-OneVision-1.5: Fully Open Framework for Democratized Multimodal Training.” 2025. Hugging Face
6. Li, B., et al. “LLaVA-OneVision: Easy Visual Task Transfer.” arXiv, 2024. arXiv
7. lmms-lab. “LLaVA-OneVision-1.5-4B-Base — Checkpoints & Conversion Notes.” Hugging Face, 2025. Hugging Face
8. EvolvingLMMs-Lab. “LLaVA-OneVision-1.5 — Issues & Data Availability Updates.” GitHub Issues, 2025. GitHub
9. lmms-lab. “Chat Template & Config Commits (4B-Instruct).” Hugging Face, 2025. Hugging Face+1
10. lmms-lab. “LLaVA-OneVision-1.5-8B-Instruct — Discussion: Improve Model Card.” Hugging Face Discussions, 2025. Hugging Face

# 6. Version History

| Version | Created on | Created by |
|---|---|---|
| 1.0 | 10/15/2025 | Kazi Islam |
| 2.0 | 11/03/2025 | Kazi Islam |
| | | |