OpenLVLM-MIA: A Controlled Benchmark Revealing the Limits of Membership Inference Attacks on Large Vision-Language Models Supplementary Material

Ryoto Miyamoto
Waseda University
Tokyo, Japan
r-miyamoto@toki.waseda.jp

Xin Fan
Waseda University
Tokyo, Japan
fan_xin@fuji.waseda.jp

Fuyuko Kido
Waseda University
Tokyo, Japan
fkido@aoni.waseda.jp

Tsuneo Matsumoto
Hitotsubashi University
Tokyo, Japan
tsuneo.matsumoto@nifty.com

Hayato Yamana
Waseda University
Tokyo, Japan
yamana@yama.info.waseda.ac.jp

1. Model Training Details

1.1. OpenCLIP-LLaVA Training Process

1.2. Model Architecture

Our OpenCLIP-LLaVA follows the LLaVA architecture with all components trained on publicly available data. Table 1 shows the model configuration.

Table 1. OpenCLIP-LLaVA architecture configuration

Component	Details
Vision Encoder	OpenCLIP ViT-B-32
Language Model	Vicuna-7B v1.5
Projector	2-layer MLP with GELU
Input Resolution	224×224
Embedding Dim	768 → 4096
Max Tokens	2048

1.3. Training Phase 1: Projector Pretraining

Table 2 shows the hyperparameters for projector pretraining.

1.4. Training Phase 2: Instruction Tuning

Table 3 shows the LoRA and training parameters for instruction tuning.

1.5. Training Dynamics

Figures 1 and 2 show the training loss curves for our OpenCLIP-LLaVA model. These training dynamics are similar to those reported in the original LLaVA paper, indicating successful training. The pretraining phase exhibits rapid initial convergence followed by stable optimization, while the instruction tuning phase shows higher variance due to the diversity of instruction-following tasks. This similarity to the original LLaVA training curves suggests that our OpenCLIP-LLaVA, despite using only publicly available components and data, achieves comparable training quality and convergence patterns. These training dynamics provide confidence that our model serves as a valid benchmark for MIA evaluation on LVLMs.

Table 2. Projector pretraining hyperparameters

Parameter	Value
Data Settings	
Training Data	LLaVA-Pretrain (558K)
Data Source	BLIP, LAION, CC, SBU
Optimization	
Optimizer	AdamW
Learning Rate	1×10^{-3}
Weight Decay	0
Scheduler	Cosine Annealing
Warmup Ratio	0.03
Training Settings	
Epochs	1
Batch Size	32 per device
Max Sequence	2048 tokens
Compute Optimization	
Mixed Precision	BF16 + TF32
Gradient Checkpoint	Enabled
DeepSpeed	Zero Stage 2

Table 3. Instruction tuning LoRA and training parameters

Parameter	Value
LoRA Settings	
Rank (r)	128
Scaling (α)	256
Target Modules	Q, V projections
Dropout	0.05
Learning Rates	
LLM (LoRA)	2×10^{-4}
Projector Fine-tune	2×10^{-5}
Data Settings	
Training Data	LLaVA-Instruct-665K
Image Processing	Aspect ratio padding
Grouping	Dynamic batching
Training Settings	
Batch Size	8 per device
Gradient Accum.	2 steps (effective: 16)
Gradient Clipping	1.0
Checkpoint Save	Every 500 steps

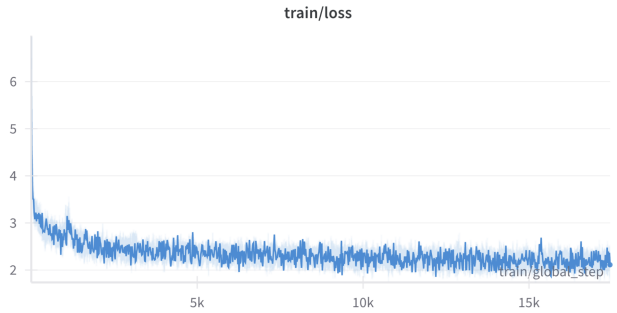

Figure 1. Training loss curve during the pretraining phase. The loss decreases rapidly in the initial steps (0-5k) from approximately 7.0 to 2.5, then stabilizes around 2.0-2.5 for the remainder of training (5k-15k steps), similar to the pattern observed in the original LLaVA training.

Figure 2. Training loss curve during instruction tuning phase. The loss exhibits higher variance compared to pretraining, fluctuating between 0.2 and 1.2 throughout the 40k training steps, similar to the original LLaVA's instruction tuning behavior.