Technical Summary of Perception Tokens Enhance Visual Reasoning in Multimodal Language Models

Part I: Problem Formulation, Methods, and Evidence

Kai-Yu Lu

2025/10/16

1 Research Problem and Motivation

Multimodal language models are effective on high level tasks that rely on language priors, yet they remain weak on mid level and low level visual perception tasks that require explicit intermediate visual structure, such as relative depth reasoning and instance counting. Standard finetuning does not generalize well across such tasks, and tool using approaches that call external depth estimators or detectors impose latency and memory overhead. This work proposes to expand the token vocabulary with *perception tokens* that represent intrinsic image representations, and to train the model to generate and to reason over these tokens as intermediate steps, thereby improving visual reasoning performance without external tools.

2 Related Work

The study distinguishes end to end multimodal language models that integrate images through cross attention or visual instruction tuning from tool using systems that route queries to external vision modules. Prior any to any models can generate masks, keypoints, or depth maps, yet they do not explicitly reason over their own visual generations to solve downstream perception tasks. The present approach augments the vocabulary to include discrete visual tokens and conditions chain of thought style reasoning on these tokens, thereby combining interpretability and single model inference.

3 Dataset Construction

Depth related data

Depth maps are generated on ADE20K images using a high quality depth estimator. A vector quantized variational autoencoder with a codebook size of 128 tokenizes each 320×320 depth map into a 10×10 grid that yields a 100 token sequence wrapped by `DEPTH_START` and `DEPTH_END`, adding 130 depth related tokens to the vocabulary. Three subsets are prepared: depth generation data for 20 000 ADE20K images, chain of thought data for 500 images with two to five markers, and direct labeling data using the same 500 images.

Counting related data

For object counting, images are taken from LVIS. Bounding boxes are represented with discrete coordinate tokens after resizing images to 336×336, which introduces 336 coordinate tokens `PIXEL_0` to `PIXEL_335`.

Three subsets are used: 5 000 LVIS images for bounding box prediction, 250 images for chain of thought prompts, and 250 for direct labeling.

Benchmarks

Relative depth is evaluated on BLINK and on a curated harder set named HardBLINK with 3, 4, and 5 markers placed at mid height. Counting is evaluated on CVBench counting, SEED Bench counting, and BLINK counting. Multiple choice options are removed, and exact outputs are required.

4 Query Protocol and Task Definitions

- **Relative depth estimation**: given an image with marked points, predict which point is closest to the camera. Metric is accuracy.
- **Object counting**: given an image and a target category, predict the exact number of visible instances. Metric is accuracy.

Prompts for chain of thought require producing perception tokens first, then deriving the final answer. Direct labeling prompts require producing only the final answer.

5 Modeling Approach

The approach, named AURORA, augments LLaVA 1.5 13B with an expanded vocabulary $V' = V \cup V_{\text{aux}}$, where V_{aux} contains perception tokens for pixel level depth and structured bounding boxes. Training follows a progressive curriculum that mixes three data regimes per task: atomic token generation, chain of thought prompts, and direct labeling. Constrained decoding ensures valid token sequences during inference.

Numbered Formulae and Explanations

Specialist to generalist distillation

$$\ell_{\text{dist}} \;=\; \min_{M} \Big(-\sum_{i} q_i \log p_{M(i)} \Big). \tag{1}$$

Symbols and background. q_i is the target probability that a frozen specialist assigns to token i in its own vocabulary V_{spec}. $M : V_{\text{spec}} \to V_{\text{aux}}$ is a one to one mapping from specialist tokens to auxiliary perception tokens. $p_{M(i)}$ is the model probability for the mapped auxiliary token. Cross entropy distillation transfers specialist knowledge. **Intuition.** Aligning distributions allows the language model to generate accurate perception tokens. **Role.** This term teaches the prediction of auxiliary tokens used later in reasoning.

Optional reconstruction

$$\ell_{\text{rec}} \;=\; \|g(t) - f\|_2^2. \tag{2}$$

Symbols and background. $t \in V_{\text{aux}}$ is an auxiliary token. g is a lightweight decoder that maps tokens back to a feature space such as depth. f is the target feature. **Intuition.** Tokens should decode to meaningful specialist features. **Role.** Improves interpretability and can help fidelity. The study notes it is optional under resource constraints.

Curriculum sampling distribution

$$p(d_t, s) \;=\; \frac{\exp\big(-d_t/\tau(s)\big)}{\sum_{i=1}^{m} \exp\big(-d_i/\tau(s)\big)}. \tag{3}$$

Symbols and background. d_t denotes the difficulty of task t, s is the training step, m is the number of difficulty levels. A temperature scaled softmax shifts sampling from easy to hard tasks.

Intuition. The model first learns the new symbols, then learns to reason with them.

Role. Reduces catastrophic forgetting while improving reasoning.

Temperature schedule

$$\tau(s) \;=\; \frac{\tau_0}{1 + \lambda \cdot s/S}. \tag{4}$$

Symbols and background. τ_0 is the initial temperature, λ is the annealing rate, S is total training steps. **Intuition.** Gradually lowers temperature to emphasize harder tasks later. **Role.** Implements the progressive curriculum in Equation (3).

Tokenization Schemes

Pixel level depth tokens. A VQVAE compresses depth maps into 100 code indices with two delimiters, yielding 130 depth related tokens.

Structured box tokens. After resizing to 336×336, each pixel coordinate is assigned a unique token. Bounding boxes are encoded as four token tuples (x_1, y_1, x_2, y_2).

Training and Inference

The vision backbone is frozen. LoRA is applied to the language model. The token embedding layer and the language model head are expanded and trained to include new tokens. Inference uses deterministic decoding and constraints to ensure valid depth token blocks and valid box tuples.

6 Empirical Results

Benchmarks and Metrics

Relative depth accuracy is reported for BLINK and for HardBLINK with 3, 4, and 5 markers. Counting accuracy is reported for CVBench counting, SEED Bench counting, and BLINK counting. Multiple choice options are removed to reduce language priors.

Main Findings

Relative depth. On BLINK relative depth, the model improves by 6.4 percentage points over a fine tuned baseline without perception tokens and maintains larger margins as the number of markers increases in HardBLINK.
Counting. On counting tasks the improvements are 10.8 percentage points on BLINK, 11.3 percentage points on CVBench, and 8.3 percentage points on SEED Bench compared with fine tuning without perception tokens.

Interpretability. Decoding depth tokens yields depth maps that pass programmatic relative depth checks, enabling human inspection of intermediate evidence.

Ablations and Generalization

Removing either chain of thought step degrades relative depth performance on HardBLINK. Using perception coordinate tokens outperforms using standard text numerals for coordinate encoding in counting. The optional reconstruction loss provides marginal gains relative to distillation. Cross task generalization from BLINK style depth training to CVBench depth shows an advantage over baselines.

7 Summary

The study introduces perception tokens and the AURORA training framework to integrate pixel level and structured visual tokens into the vocabulary of a multimodal language model. A progressive curriculum teaches both prediction and reasoning with these tokens. The approach improves relative depth and counting accuracy while preserving interpretability and avoiding external tools. Reported results use LLaVA 1.5 13B with an expanded vocabulary that adds 130 depth tokens and 336 coordinate tokens.

Technical Appendices for Reproducibility

- *Model and hyperparameters*: LLaVA 1.5 13B backbone with frozen vision encoder. LoRA applied to the language model. Embedding and language model head expanded to include new tokens. Ten epochs of fine tuning with cross entropy next token prediction and deterministic decoding at temperature zero during inference.
- *Token budgets*: 130 depth related tokens and 336 coordinate tokens added to the original vocabulary.
- *Data mixers*: Depth generation 20 000 samples. Depth chain of thought and direct labeling each built from 500 ADE20K images with 2 to 5 markers. Counting bounding box prediction 5 000 LVIS images. Counting chain of thought and direct labeling each built from 250 LVIS images.
- *Curriculum*: Shift from atomic token generation toward chain of thought and direct labeling across epochs according to Equations (3) and (4).
- *Evaluation*: Accuracy for relative depth and for counting with options removed to reduce multiple choice bias.