TPRL

Adaptive Visual Token Pruning in LVLMs

via Language-Guided Reinforcement Learning

MULTI-STEP TRAINING → ONE-SHOT INFERENCE

66.7%	65.3%	99.8%
VISUAL TOKENS REMOVED	FLOPs REDUCTION	RELATIVE ACCURACY · LoRA

Results on LLaVA-1.5-7B with approximately 192 retained tokens

MOTIVATION

The efficiency–adaptivity dilemma

Vision-language models process many redundant image tokens.

01 Heuristic pruning

FAST, BUT MOSTLY STATIC

Uses attention scores or visual feature similarity to rank tokens.

Limitation: token importance can change after earlier tokens are removed.

02 Iterative pruning

ADAPTIVE, BUT COSTLY

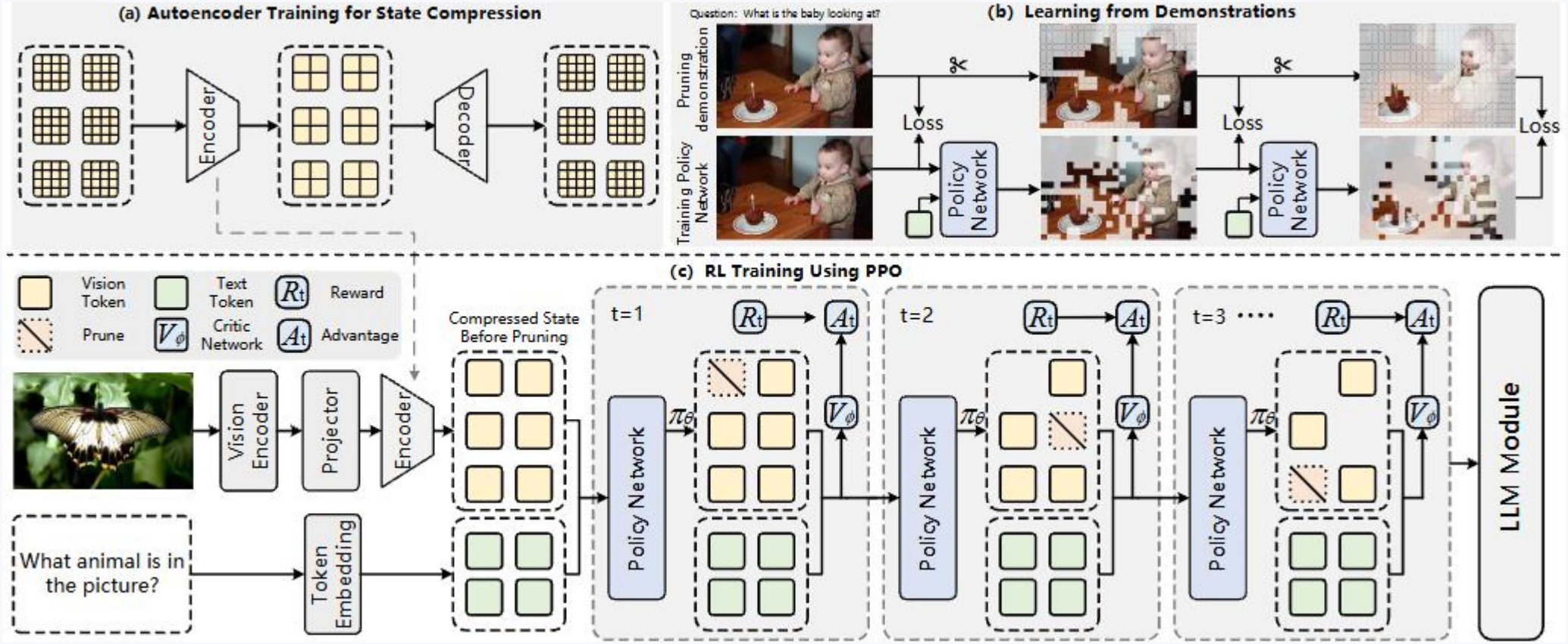
Models evolving token relationships with repeated decision steps.

Limitation: multiple policy calls may offset the latency savings.

TPRL: learn sequential dependencies offline, then select tokens once online.

FRAMEWORK

Three stages of TPRL training



TRAINING Sequential policy exploration ⟶ **INFERENCE** One-shot token mask

CORE METHOD

Sequential optimization, single-pass deployment

DURING TRAINING

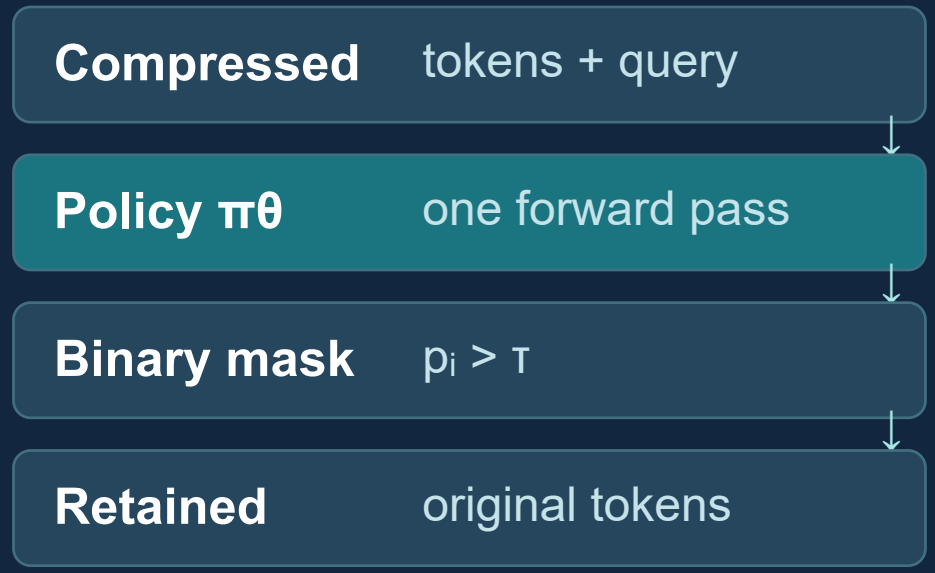
Language-guided MDP

State	$s_t = (Z_t, q)$
Action	Keep or prune each token
Update	Update the retained token set

Reward Δ Task score + Token reduction

DURING INFERENCE

One-shot token selection



The LLM receives original visual features, not compressed policy embeddings.

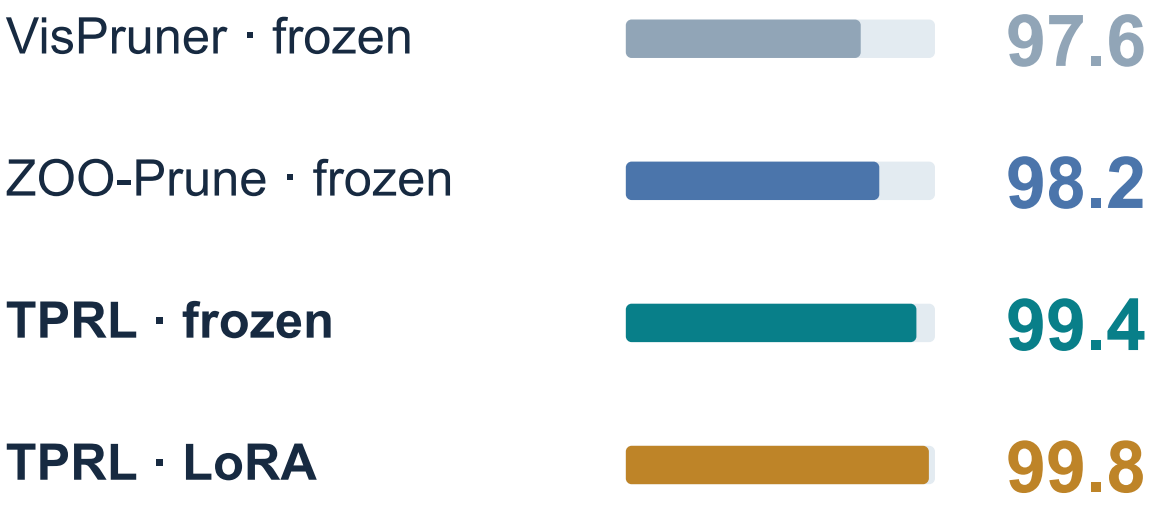
EXPERIMENTS

Near-lossless accuracy at 192 visual tokens

LLaVA-1.5-7B · 66.7% of the 576 visual tokens removed

Avg. relative accuracy (%)

At the same 192-token budget



* The LoRA result uses LLM fine-tuning.

Efficiency of TPRL · LoRA

65.3% — **Fewer FLOPs**
9.6 T → 3.3 T

38.8% — **Lower latency**
57.8 ms → 35.4 ms

65.3% — **Less cache storage**
302.4 MB → 104.8 MB

99.8% of full-token accuracy

ABLATION & VISUALIZATION

Why sequential training matters

192 tokens · Frozen LLM · Average relative accuracy

94.4%	96.3%	97.6%	99.4%
One-shot distillation	One-step RL	VisPruner teacher	**Full TPRL (3 steps)**



Paper Fig. 3: Progressive pruning preserves task-relevant image regions.