





Luigi Seminara

Antonino Furnari

Lorenzo Torresani

RECIPE: Procedural Planning via Grounding in Instructional Video

NeurIPS 2026

@luseminara.bsky.social

@Gigii_Gii

luigi-seminara

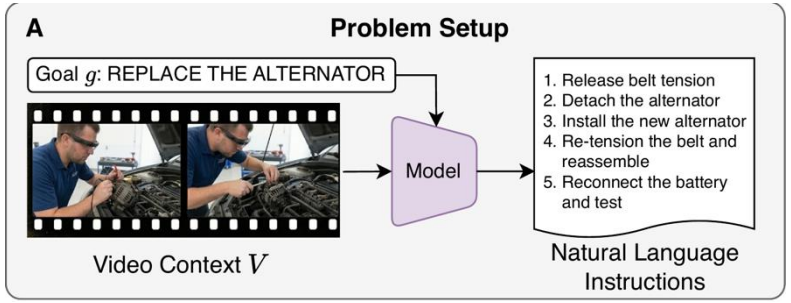
Open-Vocabulary Visual Planning

Given a partial video and a goal, generate the remaining steps in **natural language**.

CLOSED-TAXONOMY PLANNING

- **Closed set of key-steps** (e.g. 105 CrossTask actions)
- **Clean annotated sequences**: small, costly, narrow in domain
- Predictions are labels from a fixed taxonomy
- Works well when the step vocabulary is known and annotated

OPEN-VOCABULARY PLANNING

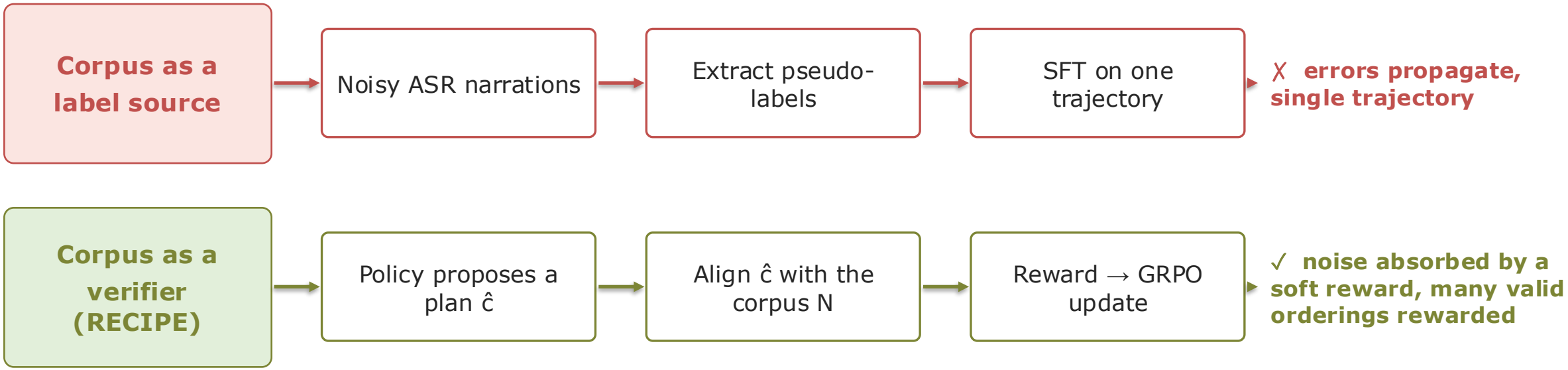


- Free-form steps conditioned on what is visible → **no fixed taxonomy**
- Largest knowledge source, HowTo100M (1.2M videos), is **noisy narration**: pseudo-labels propagate errors
- Many valid orderings for the same goal

Can we exploit noisy instructional video without extracting labels from it?

Key Insight: Verify, Don't Extract

Extracting clean step labels from noisy video is **hard**. Verifying whether a generated step sequence is temporally grounded in ASR transcripts is **cheap** and scales to millions of videos via precomputed text embeddings.



Density-estimation principle: the history localizes the relevant region of the corpus; the reward measures how well the completion aligns with it. The data *itself* is the model of plan validity.

RECIPE = Reward from Empirical Corpus of Instructional Procedure Examples

Worked Example: Grounding and Reward

① Rollouts

Goal: **Replace the alternator**

History h (from the video)
1. Disconnect the battery
2. Locate the alternator
3. Remove obstructing components

π_θ *samples G = 4 continuations (2 shown)*

Candidate $\hat{c}^1$ ✗ copies the history
4. Remove obstructing components
5. Remove obstructing components
6. Release belt tension
7. Reconnect the battery

Candidate $\hat{c}^2$ ✓ continues the plan
4. Release tension from belt
5. Remove alternator
6. Replace alternator
7. Re-tension the belt
8. Connect the battery

② Grounding in the corpus

Corpus N · HowToCaption
25M captions from 1.2M videos, text only. Scored: $s_{hist} = h$ and $s_{full} = h \oplus \hat{c}$

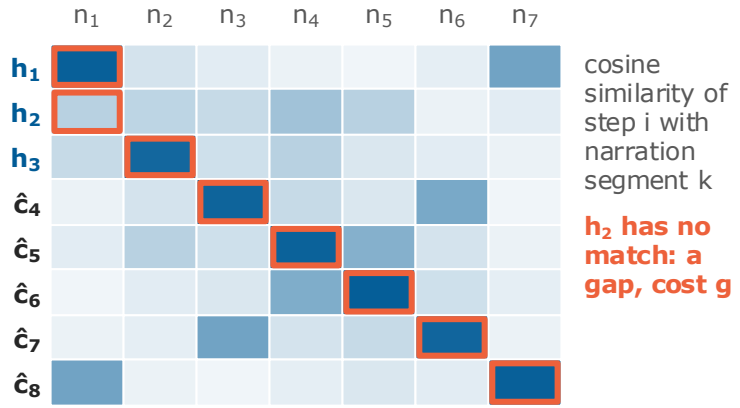
Stage 1 · retrieval, keep the top-K = 25

✓ *"…disconnect the negative terminal… loosen the tensioner… unbolt the alternator…"*

✓ *"…slip the belt off… swap in the new alternator… reconnect the battery…"*

✗ *"…clamp the jumper cables to the dead battery…" (not retained)*

Stage 2 · global alignment (NW) of $h \oplus \hat{c}^2$



③ Reward

	A(s, N)	ρ(ĉ)	$R_{base}(\hat{c})$
$s_{hist} = h$	0.40	–	–
$h \oplus \hat{c}^1$	0.43	0.05	**−0.10**
$h \oplus \hat{c}^2$	0.73	0.55	**0.73**

Grounding score: best alignment over the retained narrations

$\rho = (A_{full} - A_{hist}) / (1 - A_{hist})$
progress beyond what the history already explains

Gate $\tau = 0.10$:
$\hat{c}^1$: $\rho < \tau \rightarrow$ penalty $\text{clip}(\alpha(\rho - \tau), -1, 0)$, $\alpha = 2$
$\hat{c}^2$: $\rho \geq \tau \rightarrow$ reward $A(s_{full}, N)$

GRPO: $\hat{A} = (R - \mu) / \sigma$ within the group
$\hat{c}^2$: $\hat{A} > 0 \rightarrow$ made more likely
$\hat{c}^1$: $\hat{A} < 0 \rightarrow$ made less likely

The corpus decides: progress the narrations support is rewarded, copying the history is penalized.

Toy example: narrations, similarities and scores are illustrative; $\tau = 0.10$ and $\alpha = 2.0$ as in the paper.

Open-Ended Evaluation Protocol

Open-ended plans preclude exact match → reference-based **LLM-as-judge** (Llama-3.3-70B-Instruct). Input: goal, textual history, reference continuation, prediction → six integer scores in [0, 5].

1 Logical progression
Strict cause-and-effect: every step has its physical prerequisite in earlier steps.

2 Temporal alignment
Correct chronology; penalizes insertions, reorderings, temporal drift.

3 Spatial grounding
Coherence with the visible scene; penalizes hallucinated objects or locations.

4 Continuation
Smooth hand-off from the history; penalizes repeating it or skipping bridging steps.

5 Clarity
Precision and readability; penalizes verbose phrasing that buries the instruction.

6 Semantic alignment
Matches the reference's intent and overall plan; penalizes off-path actions.

Penalty rules

- Fluff: caps Semantic Alignment, Clarity, holistic score at 2
- Fatal failure: empty / no actionable step → all 0

Macro rubric score (%)

- Score / 30 per example, averaged per dataset, then across the datasets of a split (Score@1)

Not saturated

- GPT-5.4-mini 56.3 / 69.7, Gemini-3.1-Flash 55.5 / 69.0 (In / ZS)

Released as a benchmark for open-vocabulary planning, with an online evaluation server.

Is the LLM Judge Aligned with Humans?

ρ = 0.900

system-level Spearman with human rankings (both splits)

82.4%

ε-calibrated pairwise accuracy (tie-aware)

70.9%

pairwise agreement: judge's top pick = human's

Human study

- **178** examples (103 in-domain, 75 zero-shot), **13** human raters
- Raters see goal, history and reference, and pick the preferred continuation among Base, SFT-Ann., SFT-HT100M, RECIPE (SFT→RL), RECIPE-RL
- In 56.7% of queries the margin to the nearest alternative is ≤ 1/30 → many effective ties; differences ≤ ε = 1/30 are treated as ties (Deutsch et al., EMNLP 2023)

Split	Pairwise agreement	ε-calibrated pairwise acc.
All	70.9%	**82.4%**
In-domain	68.6%	80.6%
Zero-shot	74.2%	85.0%

The judge tracks human preferences: a reliable proxy for open-ended plan quality.

Main Results

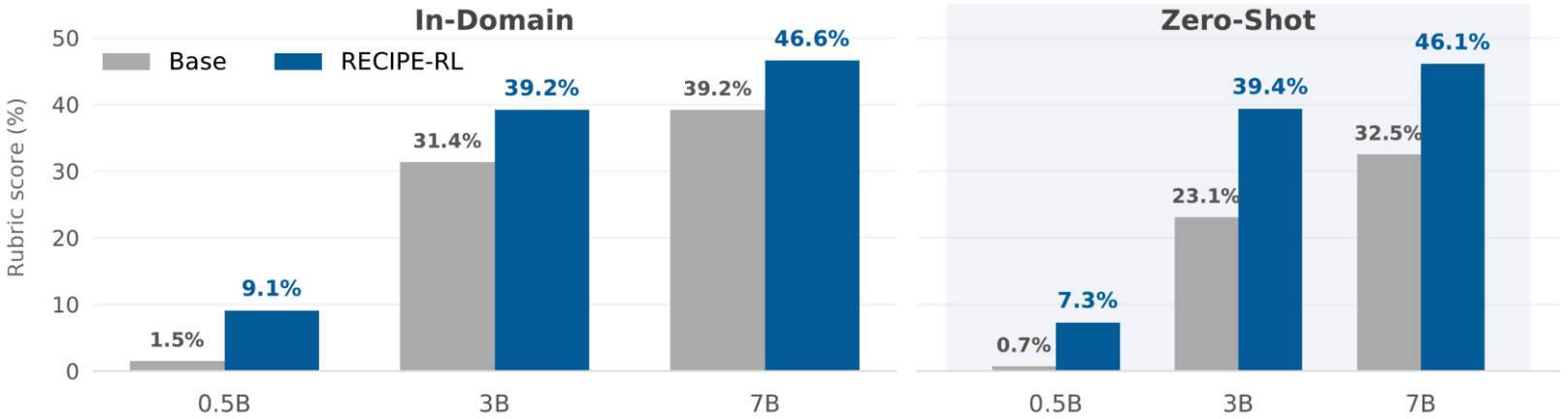


Figure 2 – macro rubric score (%) from the six-criterion LLM judge; Socratic configuration

+7 to +8 in-domain, at every scale

up to +16 zero-shot (3B: 23.1 → 39.4)

46.6 / 46.1 7B: in-domain / zero-shot

	In-domain					Zero-shot		Macro avg.	
Model	**CC4D**	**COIN**	**CrTk**	**EgPL**	**EgPER**	**EgEx**	**NIV**	**In**	**ZS**
Qwen2.5-0.5B Base	1.9	1.9	1.7	0.4	0.9	0.4	0.9	1.5	0.7
Qwen2.5-0.5B RECIPE-RL	**5.0**	**12.0**	**9.9**	**9.5**	**6.7**	**5.8**	**9.5**	**9.1**	**7.3**
Qwen2.5-3B Base	21.4	35.2	35.9	33.1	18.8	25.3	25.3	31.4	23.1
Qwen2.5-3B RECIPE-RL	**33.1**	**42.7**	**38.6**	**42.5**	**30.3**	**36.9**	**51.0**	**39.2**	**39.4**
Qwen2.5-7B Base	32.2	42.4	40.8	41.6	23.7	30.4	43.5	39.2	32.5
Qwen2.5-7B RECIPE-RL	**39.1**	**46.7**	**48.0**	**52.4**	**38.4**	**44.0**	**55.9**	**46.6**	**46.1**

RECIPE-RL improves every benchmark at every scale (Table 16).

Same Corpus, Two Roles

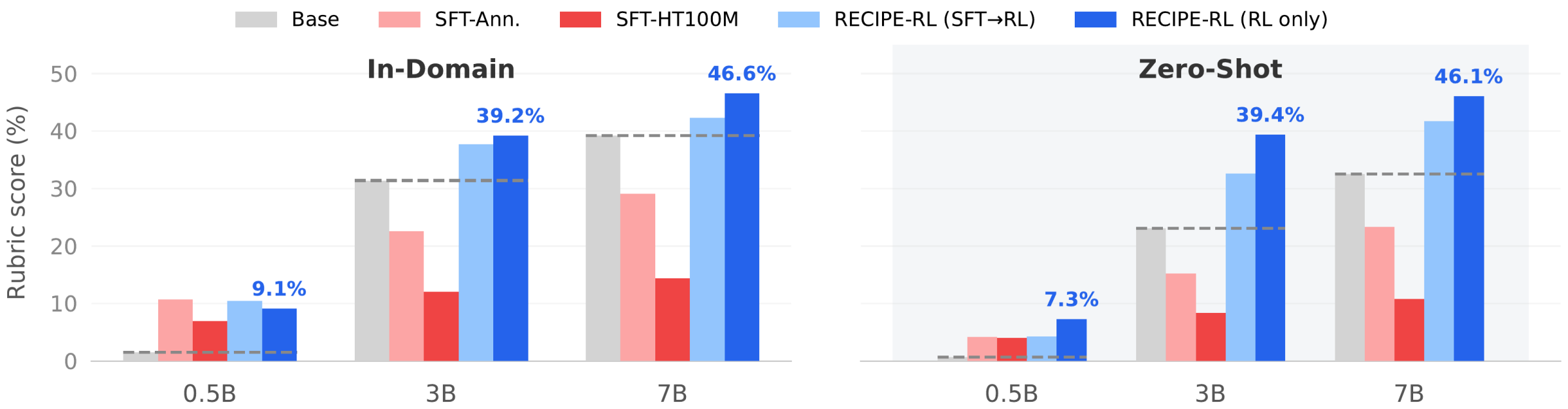


Figure 3 – both methods see HowTo100M: SFT-HT100M as pseudo-labels, RECIPE as a verifier

Labeling degrades

SFT-HT100M falls below Base at 3B and 7B: 31.4 → 12.1 in-domain at 3B

Verifying improves

Same noisy corpus, used as a verifier: +7.8 in-domain, +16.3 zero-shot at 3B

RL only is enough

RL only beats SFT→RL in 5 of 6 settings, with no SFT stage → default scheme

At 3B and 7B, both RECIPE variants outperform both SFT baselines.

Robustness to Weak Supervision

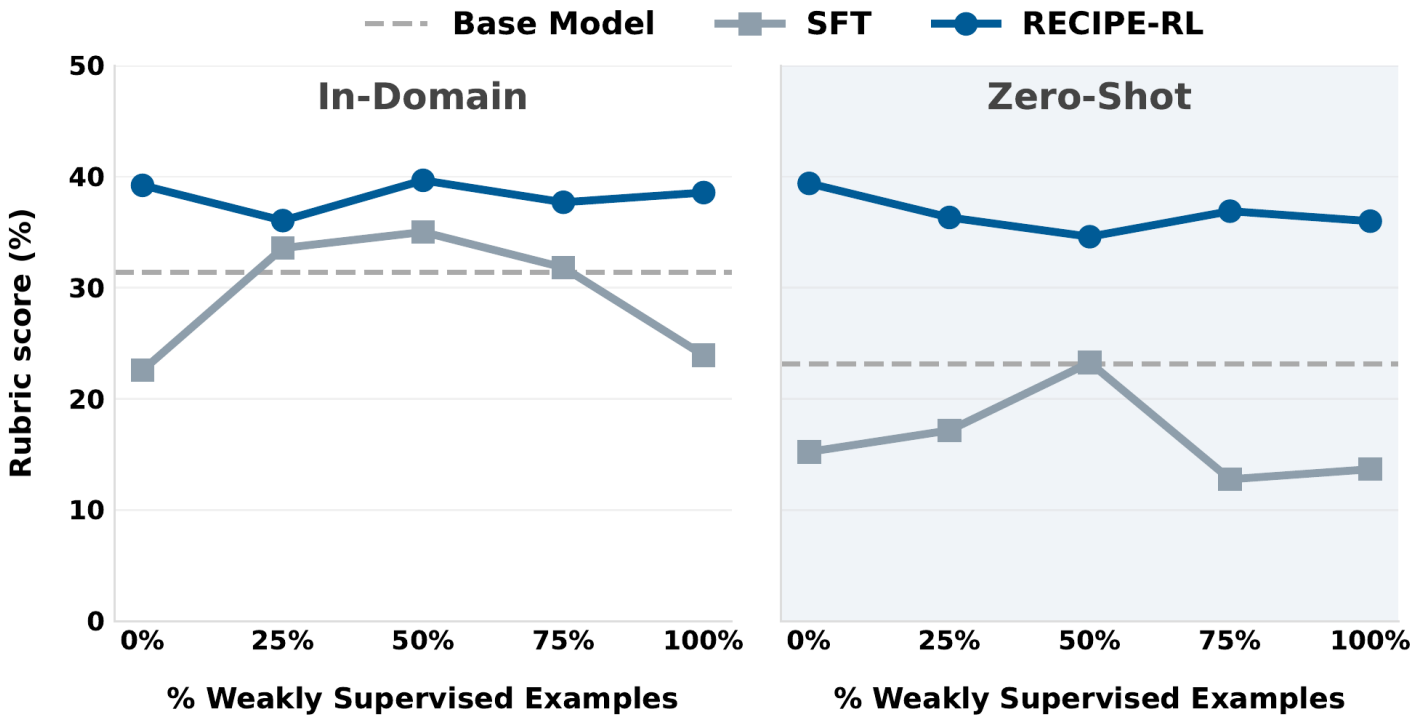


Figure 4 – Qwen2.5-3B; x-axis: % weakly supervised training examples

Per-example mixing: each training example independently uses the human trace or the VLM pseudo-trace, fixed across epochs.

Annotation-free ≈ annotated (Table 2)

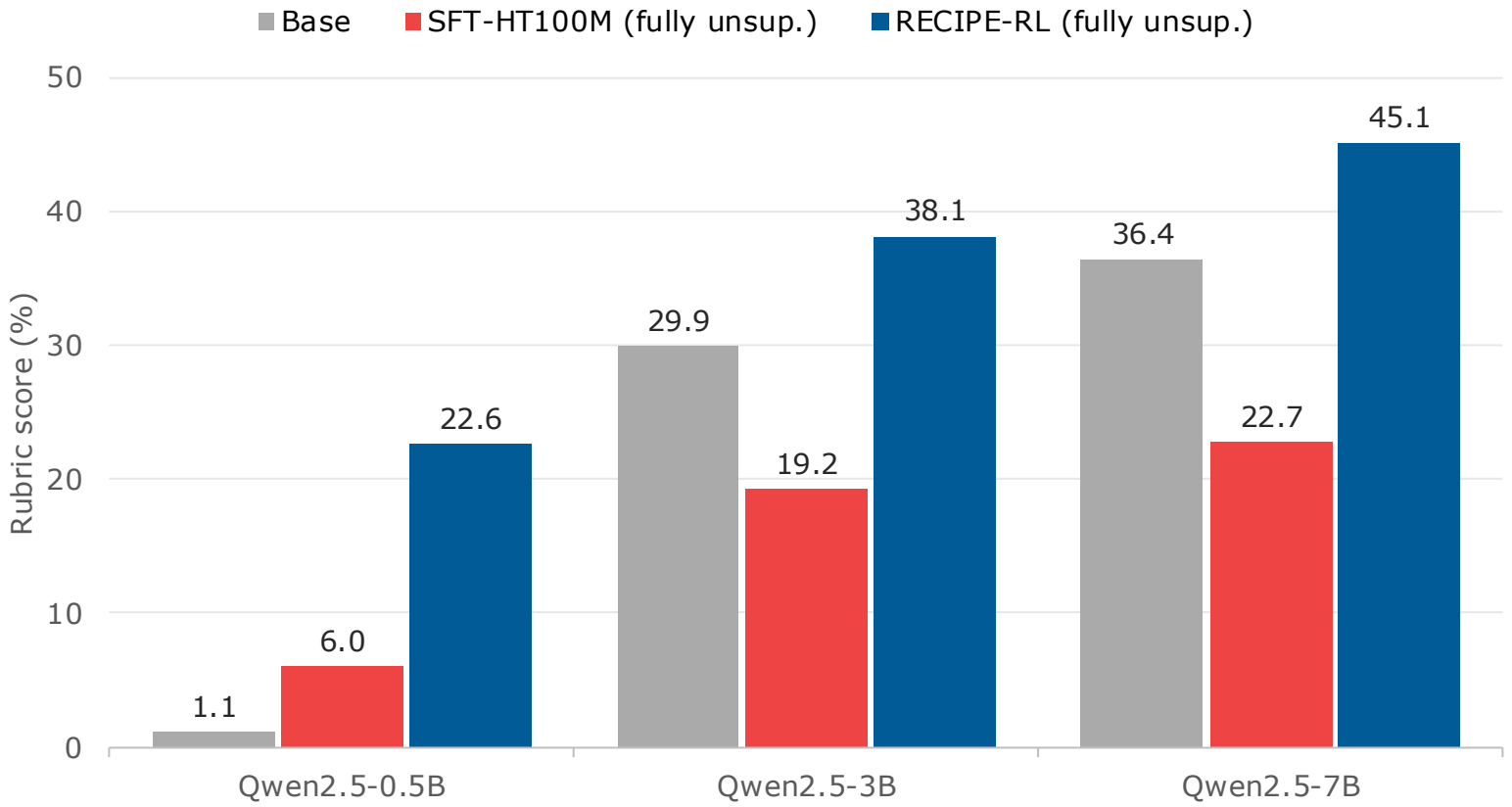
	Base		RL, 0%		RL, 100%	
Model	**In**	**ZS**	**In**	**ZS**	**In**	**ZS**
Qwen-0.5B	1.5	0.7	9.1	7.3	**11.9**	**9.1**
Qwen-3B	31.4	23.1	**39.2**	**39.4**	38.6	36.0
Qwen-7B	39.2	32.5	**46.6**	46.1	46.3	**46.3**

- RECIPE-RL is **flat**: 36–40 in-domain, 35–39 zero-shot, whatever the mix
- SFT peaks at intermediate mixes and **collapses** at the extremes; zero-shot below Base (12.8 at 25% annotated, 13.7 at 0% vs. 23.1)
- For RL a noisy pseudo-trace is just a prompt: the reward catches the mistakes SFT would lock in

The grounding reward removes the dependence on clean annotations.

Fully Unsupervised Training

No curated data: histories sampled from HowToCaption, reward against the same corpus → all 7 benchmarks are zero-shot



+21.5 over Base at 0.5B

+8.2 over Base at 3B

+8.7 over Base at 7B

SFT on the same data degrades 3B and 7B

Procedural logic can be extracted entirely from a noisy, uncurated corpus.

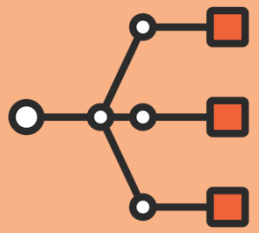
Summary



A noisy video corpus as a verifier: no pseudo-labels, no learned critic.



+7–8 in-domain, up to +16 zero-shot; robust to weak or no supervision.



Preserves the procedural variety that SFT collapses.

Project page: farsightlab.github.io/RECIPE

Università di Catania

Northeastern University
Khoury College of Computer Sciences



Luigi Seminara



Antonino Furnari



Lorenzo Torresani

RECIPE: Procedural Planning via Grounding in Instructional Video

NeurIPS 2026

@luseminara.bsky.social

@Gigii_Gii

luigi-seminara