NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2026
Sydney

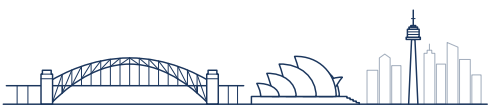
Visual Anchoring for Scenario-Guided Forecasting

Patara Trirat[1] Jay Heo[1] Heejun Lee[1,2] Sung Ju Hwang[1,2]

[1]DeepAuto.ai [2]KAIST

DeepAuto.ai

KAIST



Motivation

An analyst asks a multimodal LLM (MLLM) to forecast the dollar index under *“the Fed hikes 75 bp in Q3 and holds.”* In a stress test, the scenario must shape every step.

Does the scenario shape the forecast for twelve months, or for twelve days?

```
You are a forecasting assistant.
Given the historical time series
and the forward-looking scenario
below, forecast the next
30 values.

Scenario: If central bank
guidance remains collectively ...

Observed history (pre-anchor, most
recent last):
step;value
1;114.5676
...
90;114.5619

Produce exactly 30 numerical
values, one per line, numerical
only. No commentary.
```

The text-only prompt, shortened.

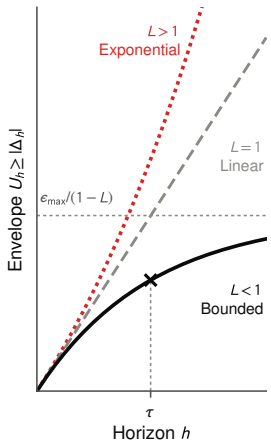
Contributions

- **A law.** Scenario drift obeys a closed-form envelope with two fitted numbers, $\epsilon_{\max}$ and L, both identifiable from paired runs.
- **A failed fix.** Restating the scenario as text every 20 steps enlarges the drift on all three open-weight models.
- ***Visan*.** One chart of the history, with no training, lowers WIT error on four of five MLLMs, significantly on three.

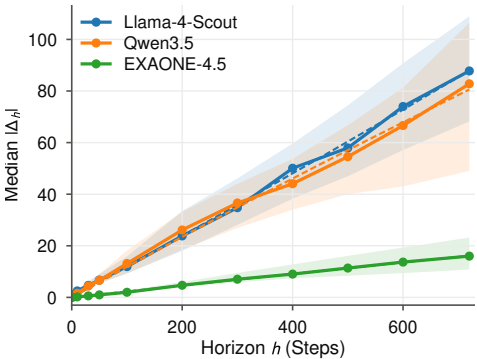
How Drift Adds Up

Two runs of one model and seed differ only in the scenario. Their gap is the drift Δ_h.

Each forecast value feeds the next step. The scenario adds at most $\epsilon_{\max}$ at each step, and the gap carries forward with a gain of at most L. So drift is bounded by a geometric sum that levels off if $L < 1$, grows linearly if $L = 1$, and grows exponentially if $L > 1$.



Drift Compounds Through All 720 Steps



Model	$\hat{\epsilon}_{\max}$	$\hat{L}$
Llama-4-Scout	0.118 [0.092, 0.156]	1.0001 [0.9995, 1.0004]
Qwen3.5	0.120 [0.097, 0.154]	0.9998 [0.9977, 1.0003]
EXAONE-4.5	0.023 [0.020, 0.031]	0.9999 [0.9987, 1.0004]

The fitted $\hat{L} \approx 1$ on all three models, so drift keeps compounding through all 720 steps. Attention tracks the drift on Llama-4-Scout only, not on Qwen3.5 or EXAONE-4.5.

Restating the Scenario Makes Drift Worse

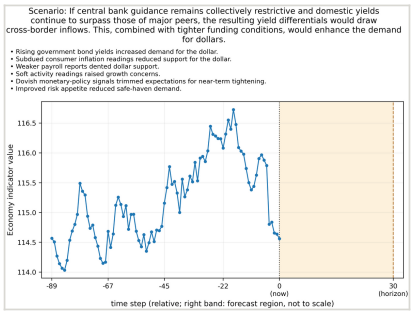
The envelope restarts within each chunk of consecutive steps, which suggests a cheap fix, restating the scenario every 20 steps.

	Test 1		Test 2	
Model	**Text-Only**	**Re-Injection**	**Random Text**	**True Scenario**
Llama-4-Scout	4.33	**31.00**	3.79	**23.00**
Qwen3.5	4.09	**18.41**	1.49	**12.76**
EXAONE-4.5	4.34	**12.35**	5.00	**14.95**

Median maximum excursion within a chunk of 20 steps.

Restating the scenario enlarges the drift within each chunk on all three models, while random text of the same length shrinks it.

Visan Draws the History as One Chart



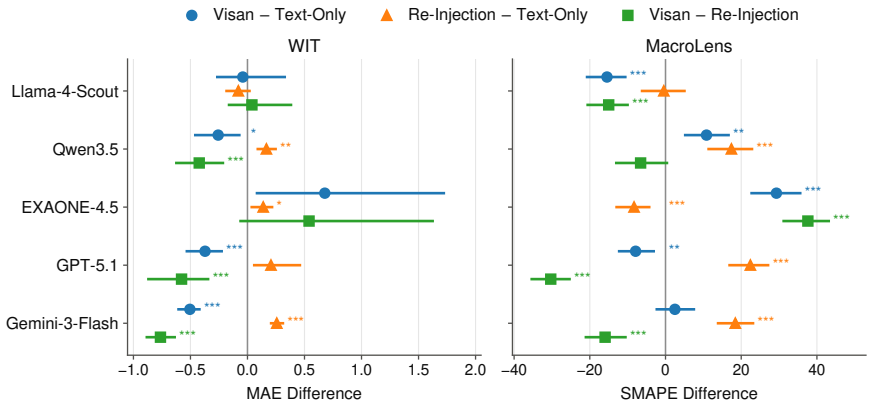

```
Use the attached chart as the sole input.
Forecast 30 consecutive future values
starting at step 1, one numeric value per
line, no commentary.
```

Visan draws the lookback, the scenario, and the historical context as one chart with a bare directive, and it needs no training.

The scenario is the chart's title, the context is its header, and the lookback is the curve, with the forecast region marked.

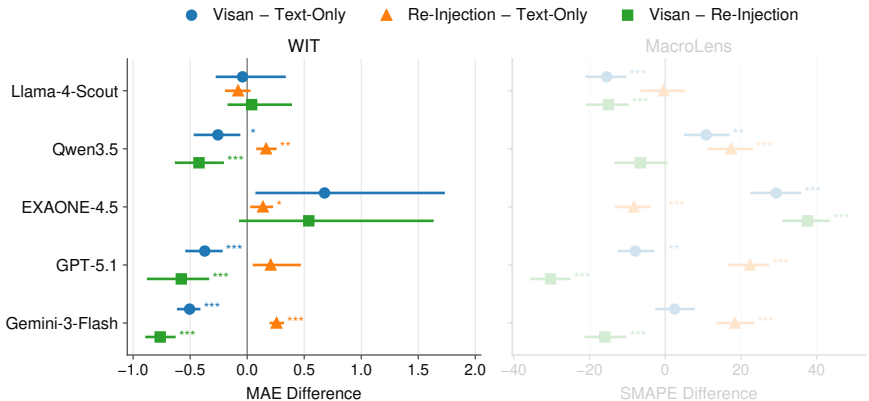
***Visan* helps only a model that can read numbers off a chart.**

Visan Reduces Forecast Error



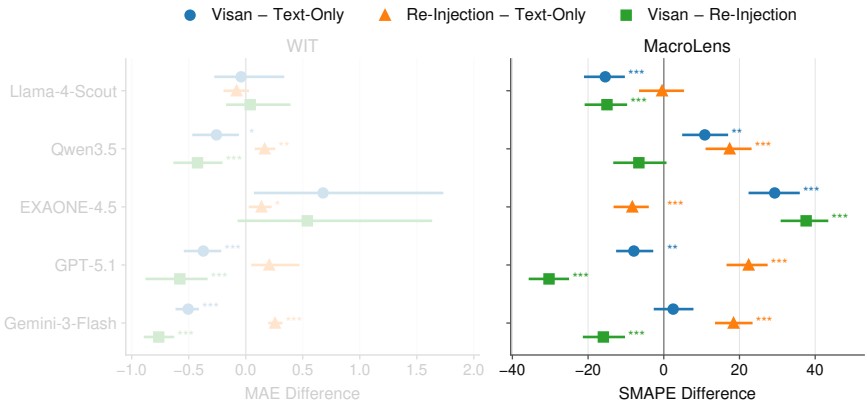
Paired contrasts with 95 % CIs, one row for each MLLM. Negative values favor the first prompt of each pair.

Visan Reduces Forecast Error



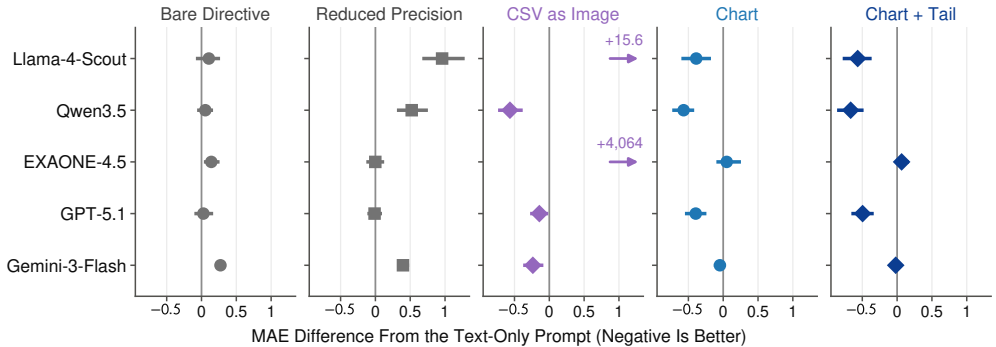
On WIT, *Visan* lowers MAE on four of five MLLMs, significantly on three, while text re-injection improves none significantly.

Visan Reduces Forecast Error



On the 252-step MacroLens task, *Visan* beats text re-injection on four of five MLLMs, and the text-only prompt on two.

Why It Works



Neither the directive nor the digit precision produces the gain, and the two chart variants are the only prompts that hurt no model.

Takeaway

Measure drift with paired runs, and draw the history instead of restating the scenario.

- Drift keeps compounding at $\hat{L} \approx 1$.
- Restating the scenario makes drift worse.
- One chart lowers error on most MLLMs.

deepauto-ai.github.io/visan