ICLR 2026
The Fourteenth International Conference on Learning Representations

Grounding or Guessing?

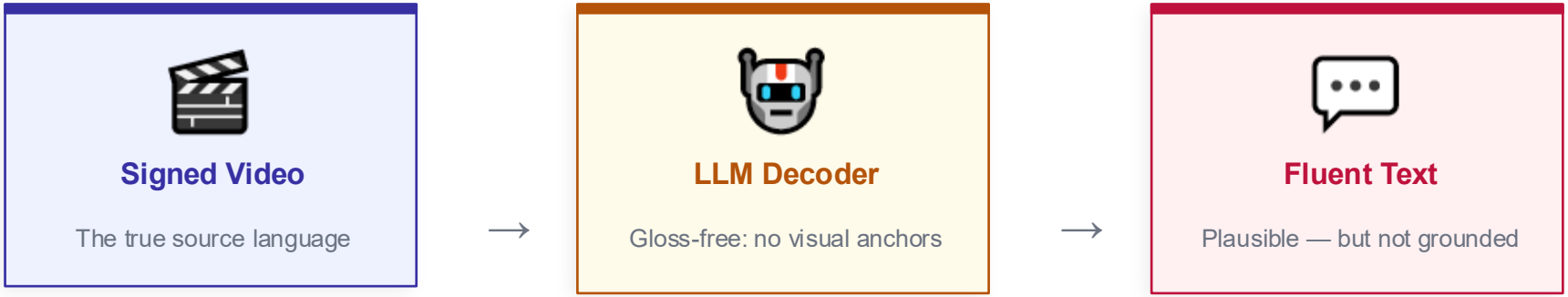
Visual Signals for Detecting Hallucinations in Sign Language Translation

Yasser HAMIDULLAH · Koel Dutta Chowdhury · Yusser Al Ghussin · Shakib Yazdani · Cennet Oguz · Josef Genabith · Cristina España-Bonet

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

In SLT, the video IS the source language — but LLMs tend to ignore it.



Hallucination in SLT = *when content typically signed is instead guessed not grounded on the signed video*

At each decoding step, the model does one of two things:

Token derived from the signed video

Token predicted from language priors alone

HALLUCINATE

Content absent from the video entirely

Our goal: *assign a reliability measure* $r_t \in [0,1]$ *to each token high when grounded, low when guessed.*

How much does each token actually rely on the video?

Input: video

3 passes

Clean

No video

Mismatched

https://signllm.github.io

Visual signal Extraction

Feature-Based (h)

Hidden-state angle

Cross-attention

Counterfactual (g)

Probabilities

Probability Margins

Token-Level Fusion

$$r_t = \sigma(w \cdot h + w \cdot g + b)$$

Sentence-Level Fusion

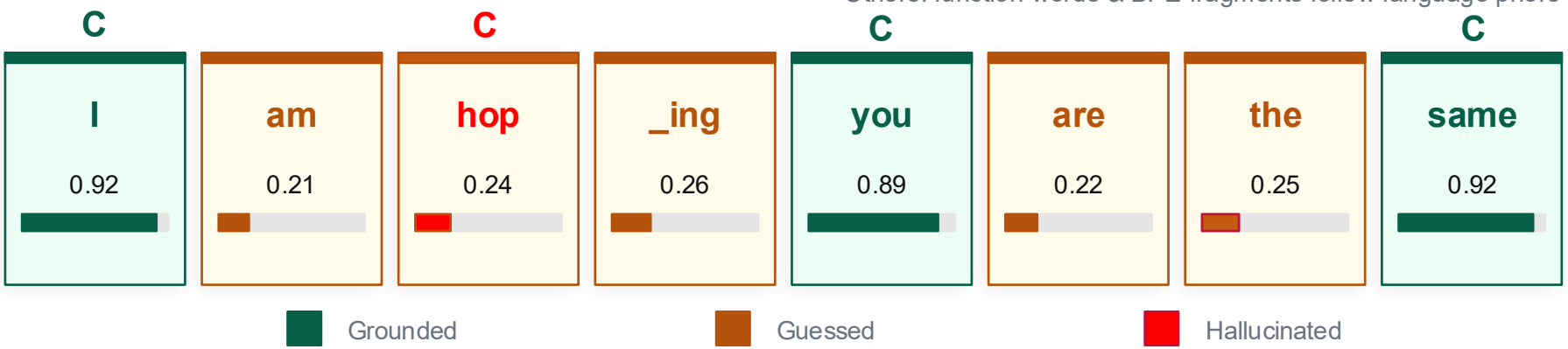
R = Aggregate over r_t

TOKEN-LEVEL EXAMPLE

Each token gets a reliability score — low scores flag hallucination risk.

Gold:	I think the same about you
Pred:	I am hop _ing you are the same

C: Content typically signed
Others: function words & BPE fragments follow language priors



Sentence level R = 0.24 → flagged as hallucinated

KEY RESULTS

Grounding alone already beats text-only signals.

97%	+0.20	≈0.99	ρ=0.72
Detection Accuracy	**AUC gain vs. Entropy**	**Average Precision**	**Spearman Correlation**
ACC @ 0.5 threshold	consistent across all settings	near-perfect recall	predicting CHAIR score

META = Grounding + Text signals → Best overall

Visual grounding captures what text-only signals miss — they are complementary, not redundant.

Three signals tell the same story.

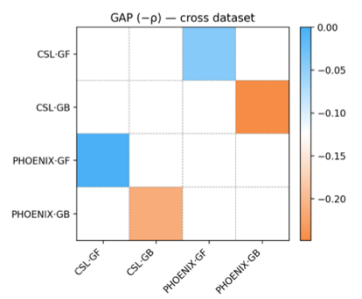
	Gloss-Free (GF)	Gloss-Based (GB)
1 **Log-Margin** 	Left-shifted — low margin even with real video	High-margin tail — video strongly boosts tokens
2 **Hidden-State Angle** 	0.20–0.35 — video barely changes hidden states	0.55–0.70 — strong shift when video is removed
3 **Confidence (−log p)** 	Near zero — over-confident, language-driven	Heavier tail — appropriate uncertainty

Formally proven: GF training raises hallucination rates by increasing weak visual use. (Prop. 1)

Reliability holds up across datasets, architectures, and degradations.

Cross-Dataset

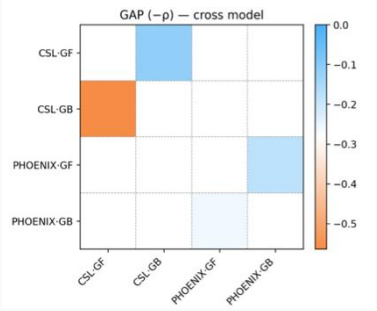
CSL-Daily ↔ PHOENIX



Weights transfer across datasets with only a small drop.

Cross-Architecture

Gloss-Free ↔ Gloss-Based

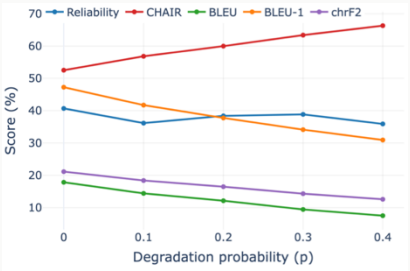


GF → GB: larger drop.

GB → GF: smaller GF weights generalize better.

Visual Degradation

Noise + Frame Dropping



As video quality degrades, reliability drops consistently.

CONCLUSION

"Did the video actually matter?"

That is what reliability asks — and answers.

- ✓Hallucination in SLT cannot be detected by text signals alone.
- ✓Reliability = token-level score quantifying dependence on video, via feature shifts and counterfactual margins.
- ✓Grounding and text only (uncertainty) => complementary signals.
- ✓Gloss-free models systematically under-use video => hallucinates more.
- ✓Framework generalizes across datasets and architectures