CCMT 2026
Kunming, China
13–15 November

TD²: Text-Driven Decoding for Mitigating Image-Induced Hallucination in Low-Resource Multimodal Translation

Yi Chen Mieradilijiang Maimaiti† Pei Yu Jiaxin Wang
Shuoshuo Hou Zhexin Li Zhengyi Guo Wushouer Silamu
School of Computer Science and Technology, Xinjiang University, China

Overview & Motivation

Research Gap

- Image priors can override source meaning.
- Low-resource translation remains vulnerable to unsupported visual entities.

Challenges

- Image-induced hallucination
- Source–image conflicts
- Source faithfulness

Our Goal

Suppress source-agnostic image priors while retaining useful visual context.
TD² calibrates decoding logits without retraining or an auxiliary model.

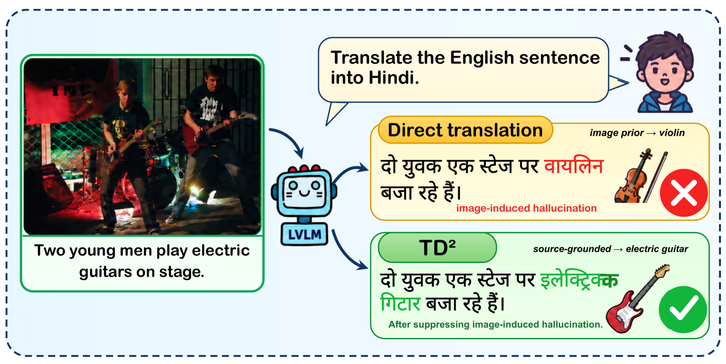


Fig. 1. TD² preserves the source entity.

METHOD: TEXT-DRIVEN DECODING

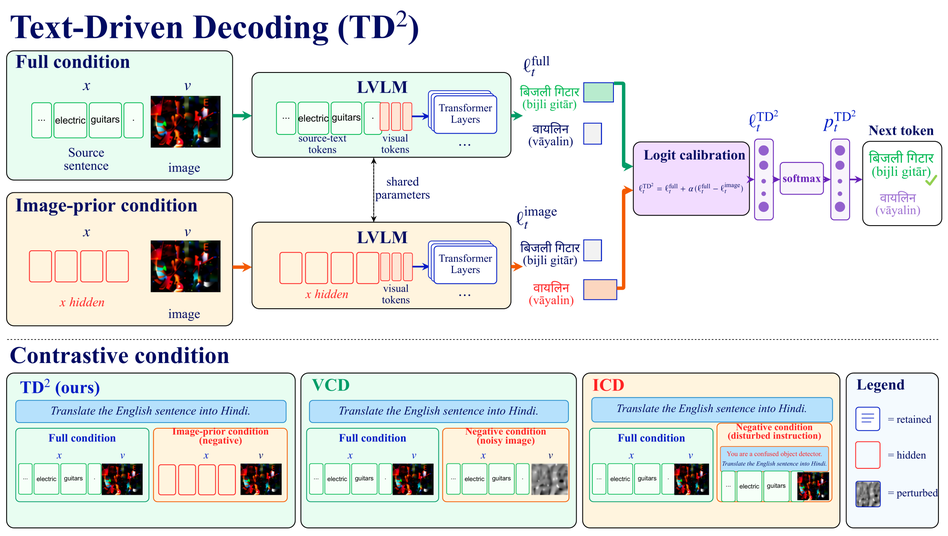


Fig. 2. Full and image-prior branches; comparison with VCD and ICD.

Two conditioning branches

- **Full context:** Condition on the source sentence and image.
- **Image prior:** Hide the source; retain the image and target language. Both branches share one LVLM and output prefix.

Text-driven logit calibration

- **Contrastive update:** Full + α × (Full − Image prior). Suppress source-agnostic visual preferences before greedy decoding.
- **Contrast strength:** α = 0 recovers Direct decoding.

Evaluation setup

- **Coverage:** 5 backbones, 3 families; VG (ha, ma, bn, hi) and Multi30K (hi, tr, lv).
- **Metrics:** BLEU and COMET (×100).

EXPERIMENTAL RESULTS & ANALYSIS

COMET improves on 5/5 backbones; BLEU improves on 4/5

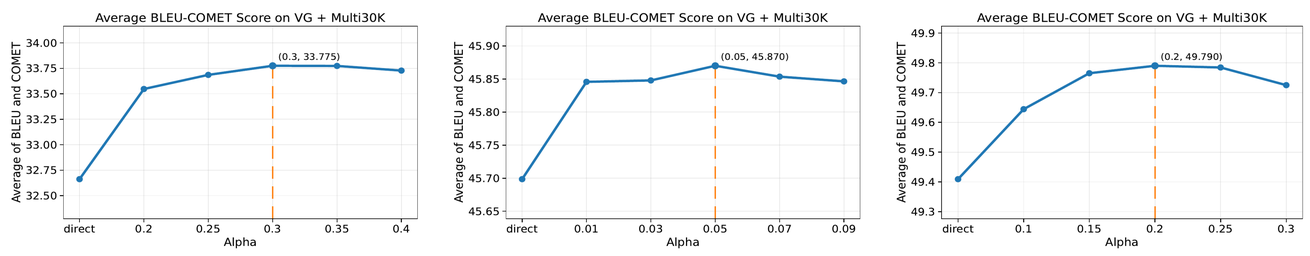
Backbone	BLEU Direct	BLEU TD²	Δ BLEU	COMET Direct	COMET TD²	Δ COMET
Qwen2-VL-7B	10.30	**10.94**	+0.64	55.02	**56.61**	+1.59
Qwen3-VL-8B	19.50	**19.76**	+0.26	71.93	**71.98**	+0.05
Qwen3-VL-32B	23.51	**23.89**	+0.38	75.30	**75.69**	+0.38
InternVL2-2B	5.00	**4.88**	-0.12	37.84	**38.84**	+1.00
LLaVA-OV-1.5-4B	15.51	**16.61**	+1.10	65.02	**65.25**	+0.23

Table 1. Seven-pair macro means; Qwen deltas use unrounded means.

Main results

- **Largest Qwen gain:** Qwen2-VL gains 1.59 COMET; Qwen3 gains are more modest.
- **Metric trade-off:** InternVL2 gains COMET but loses 0.12 BLEU.
- **Baseline comparison:** TD² exceeds VCD and ICD on both non-Qwen backbones.

Parameter sensitivity (Fig. 3)

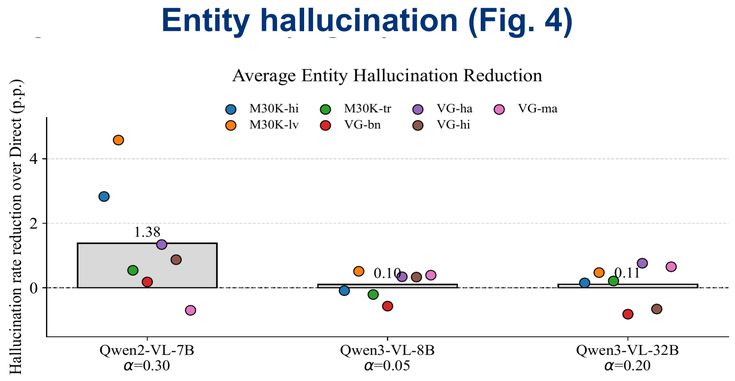


(a) Qwen2-VL-7B (b) Qwen3-VL-8B (c) Qwen3-VL-32B

Backbone-specific α

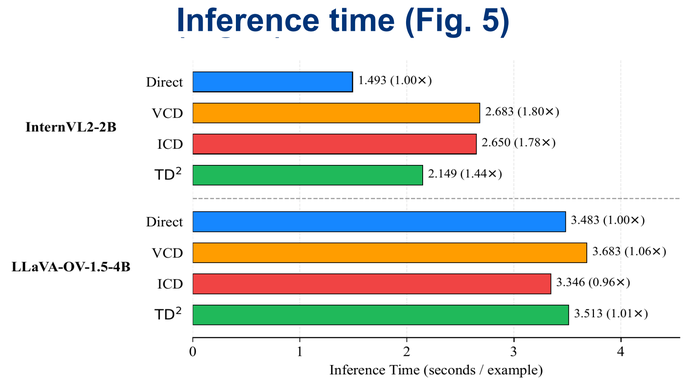
- **Reported peaks:** α = 0.30 / 0.05 / 0.20 for Qwen2-7B / Qwen3-8B / Qwen3-32B.
- **Calibration strength:** Larger α does not always help. Peaks reflect aggregate sensitivity, not independent validation.

Entity hallucination (Fig. 4)



GPT-5.5 entity-mismatch evaluation; rates are evaluator-specific.

Inference time (Fig. 5)



Hallucination reduction

- **Mean reduction:** 1.38 p.p. on Qwen2-VL versus 0.10–0.11 on Qwen3; some language pairs worsen.

Runtime trade-off

- **Measured overhead:** 1.44× / 1.01× Direct time on InternVL2 / LLaVA-OV; costs vary by backbone and output length.

Ablation: conditioning inputs (Table 2)

Model	Direct	Text-only	Img-prior	TD²
Qwen2-7B	32.66	33.57	17.33	33.78
Qwen3-8B	45.72	45.68	22.12	45.87
Qwen3-32B	49.41	49.04	26.62	49.79

Mean BLEU–COMET across seven pairs.

Ablation analysis

- **Source text:** Image-prior-only decoding scores far below text-conditioned decoding.
- **Combined context:** TD² gives the highest mean score on all three Qwen backbones.

CONCLUSION

- TD² suppresses source-agnostic visual priors to improve source faithfulness without retraining.
- On VG and Multi30K, COMET improves on all five backbones and BLEU on four of five.
- TD² exceeds VCD and ICD in both metrics on the two non-Qwen backbones.
- Entity analyses show reduced source-unsupported hallucination in multiple, but not all, settings.

Acknowledgement: Supported by NSFC (Grant No. 62406316) and the Xinjiang "Tianchi Talent" Recruitment and Introduction Program.

† Corresponding author: Mieradilijiang Maimaiti (miradeljan51@xju.edu.cn)