CCMT 2026

Kunming, China · 13–15 November 2026

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

Training-free calibration for source-faithful 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

† Corresponding author: miradeljan51@xju.edu.cn

Outline

01

Background and Motivation

Image-induced hallucination in translation

02

Methodology

Method overview and source-faithful decoding

03

Experimental Setup

Datasets, model backbones and evaluation metrics

04

Main Results

Translation quality and visual-prior ablation

05

Analysis and Conclusion

Contrast strength, hallucination and inference cost

Background and Motivation

Source faithfulness

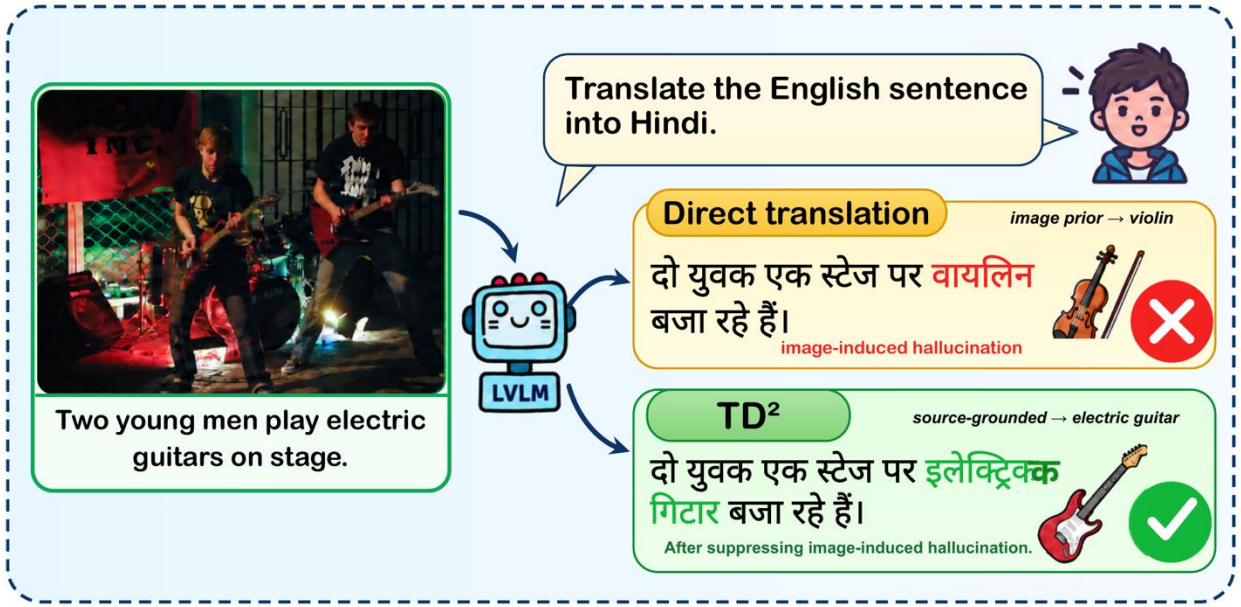
The source sentence determines the meaning.
The image provides auxiliary context.

The failure mode

Visual priors can override textual evidence and introduce unsupported entities.

Our objective

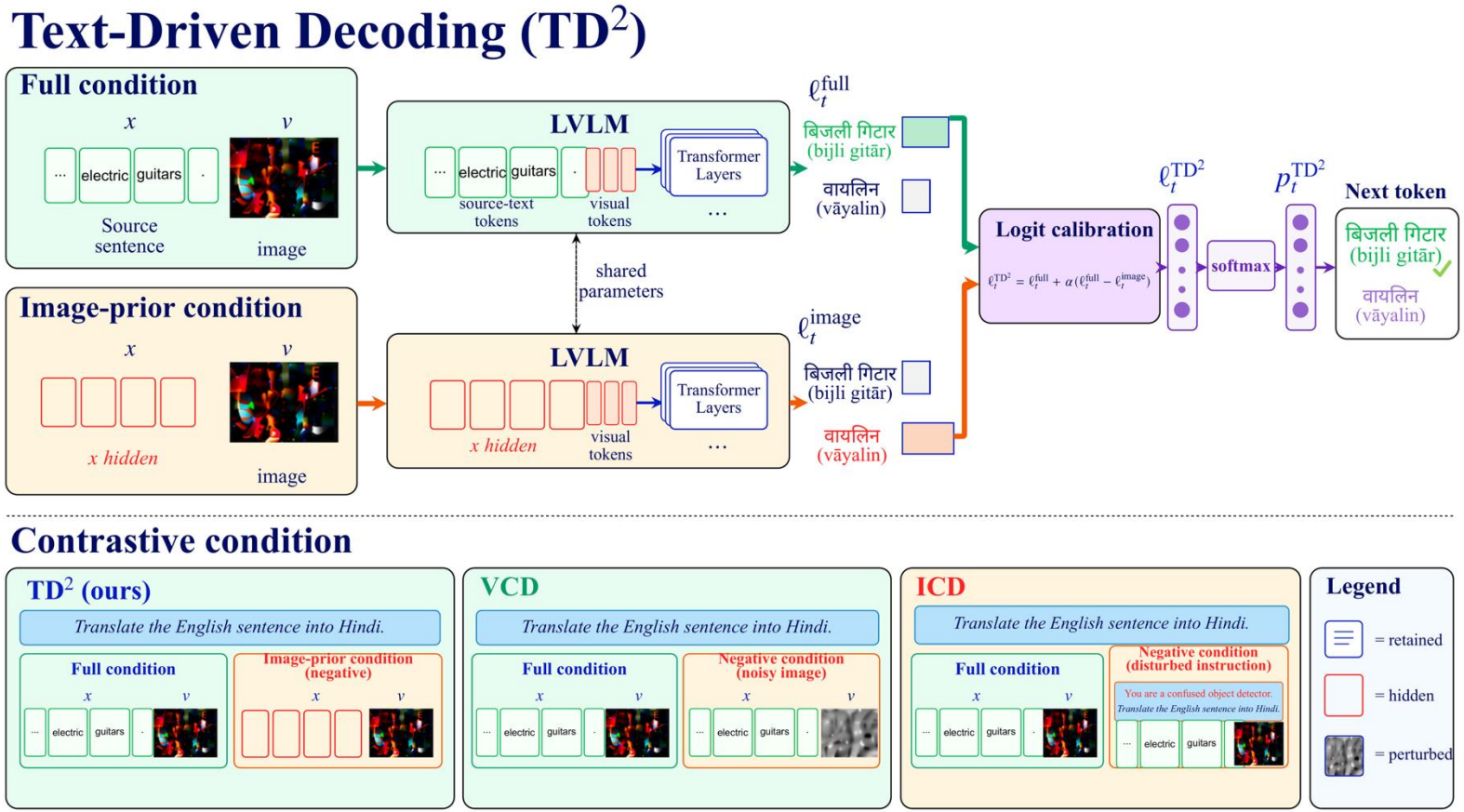
Keep useful visual context while suppressing source-agnostic image priors.



Direct: “violin”
TD²: “electric guitars”

The example illustrates source-unsupported entity substitution.

Method



One shared LVLM

1 Full condition
Source + image + prefix

2 Image-prior condition
Hide the source sentence

3 Logit calibration
Suppress prior-favored token preferences

Both branches use the same generated prefix.

The LVLM parameters are shared and frozen.

Datasets and Evaluation

English-to-target multimodal translation

Benchmark	Target languages	Examples / language
VG	Hausa, Marathi, Bengali, Hindi	1,595
Multi30K	Hindi, Turkish, Latvian	1,000

Five backbones, three families

Qwen2-VL-7B
Qwen3-VL-8B and Qwen3-VL-32B
InternVL2-2B
LLaVA-OneVision-1.5-4B

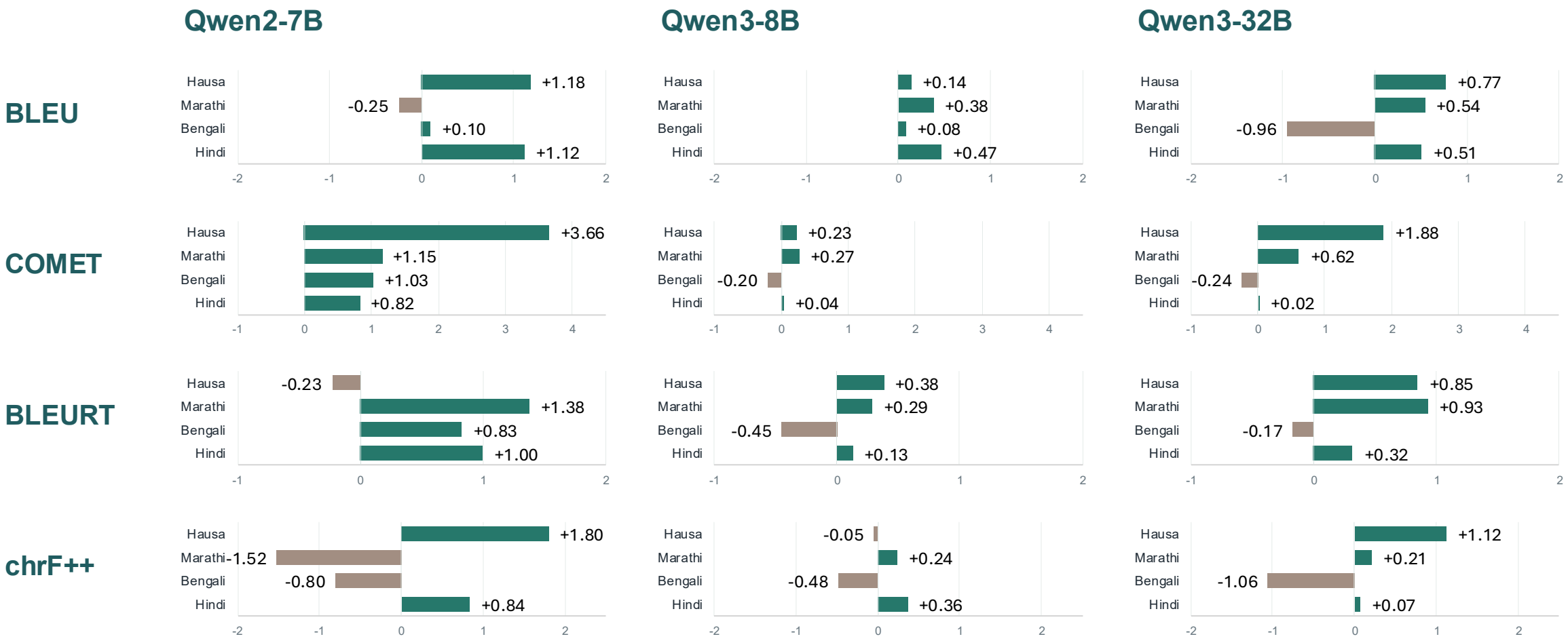
Metrics and decoding

BLEU: mean sentence-level SacreBLEU
COMET: wmt22-comet-da, scaled by 100
BLEURT and chrF++ for main Qwen results
Single deterministic greedy run

VG: 1,595 examples/language; Multi30K: 1,000 examples/language. Five backbones, three model families.

VG: changes across four evaluation metrics

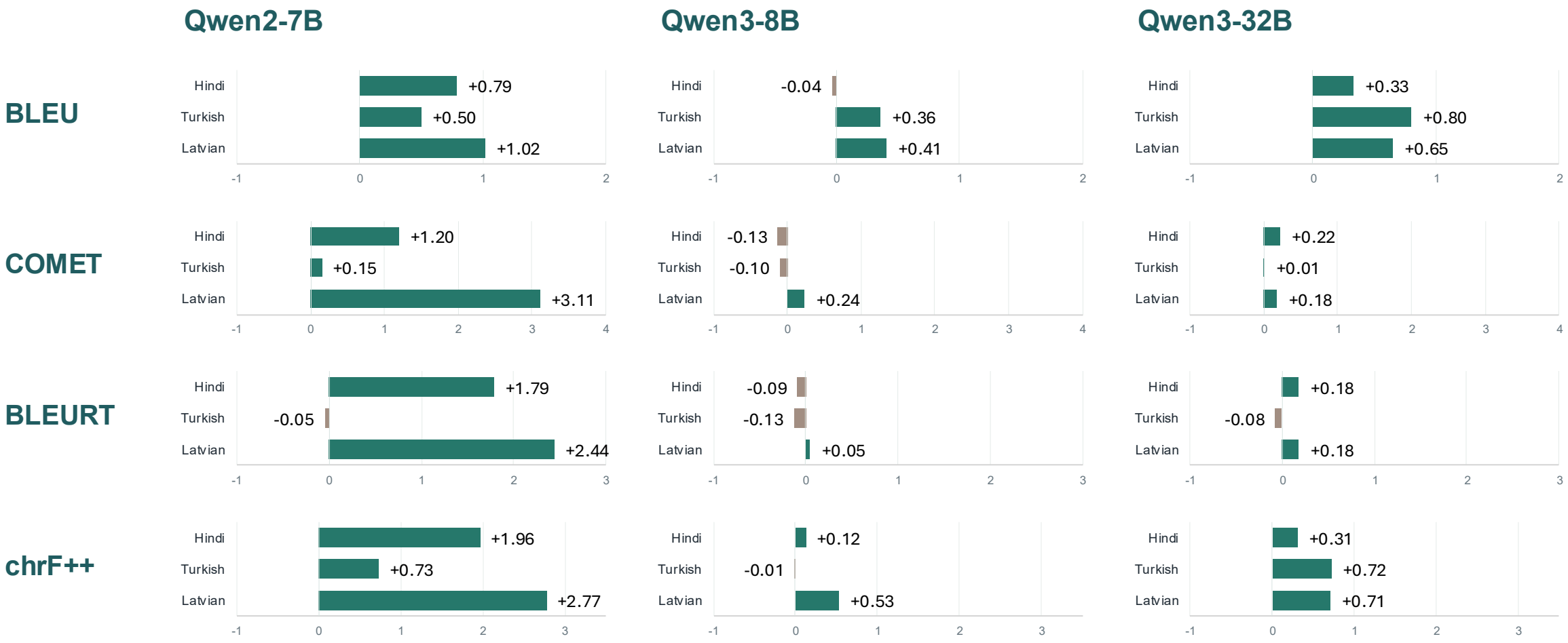
TD² − Direct Teal: improvement Warm gray: decline



Differences use reported scores. Each metric uses a common axis across models; scales vary between metrics.

Multi30K: changes across four evaluation metrics

TD² − Direct Teal: improvement Warm gray: decline



Differences use reported scores. Each metric uses a common axis across models; scales vary between metrics.

Cross-Family Results and Visual-Prior Ablation

Cross-family comparison

Backbone	Method	BLEU	COMET
InternVL2-2B	Direct	5.00	37.84
InternVL2-2B	VCD	4.16	37.61
InternVL2-2B	ICD	3.94	37.46
InternVL2-2B	**TD²**	**4.88**	**38.84**
LLaVA-OV-4B	Direct	15.51	65.02
LLaVA-OV-4B	VCD	14.08	62.27
LLaVA-OV-4B	ICD	13.04	60.77
LLaVA-OV-4B	**TD²**	**16.61**	**65.25**

TD² exceeds VCD and ICD in both metrics.
InternVL2 BLEU remains below Direct.

Visual-prior ablation

Backbone	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**

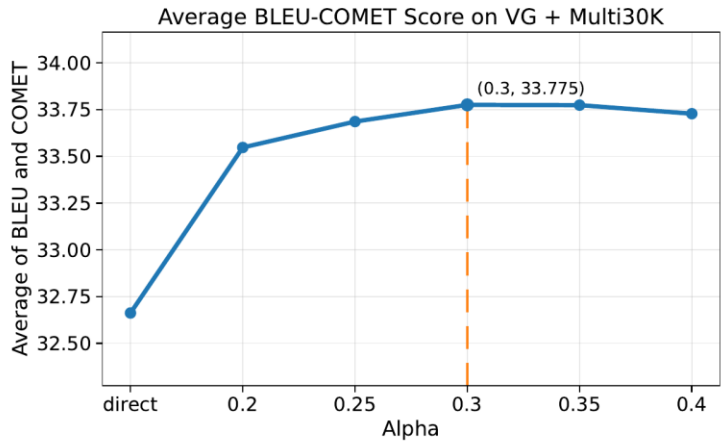
Mean of BLEU and COMET over seven pairs.

Text-only: remove the image.
Image-prior: hide the source sentence.
TD²: retain full input and suppress image priors.

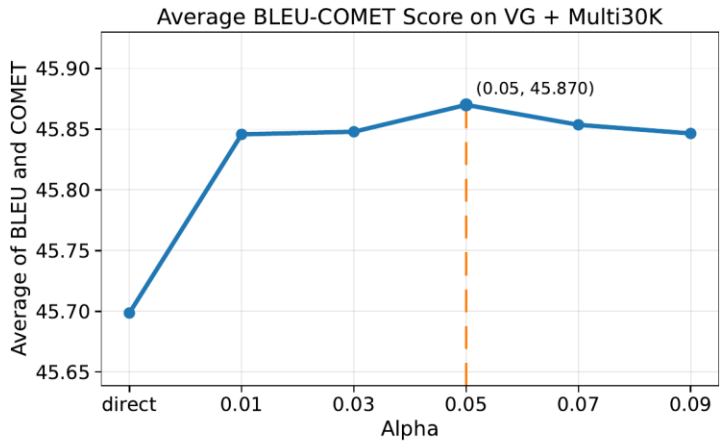
TD² has the highest mean score in all three rows.
Text-only remains competitive.

Source: main-text Tables 3–4. Table 3 reports macro BLEU / COMET; Table 4 reports their mean. LLaVA-OV-4B = LLaVA-OneVision-1.5-4B.

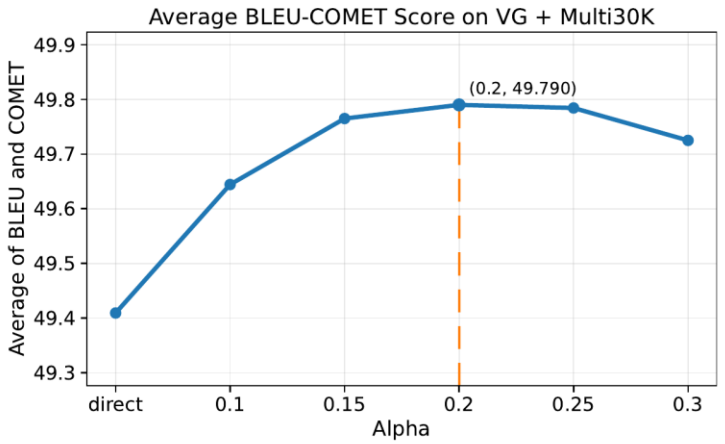
Sensitivity to Contrast Strength



(a) Qwen2-VL-7B



(b) Qwen3-VL-8B



(c) Qwen3-VL-32B

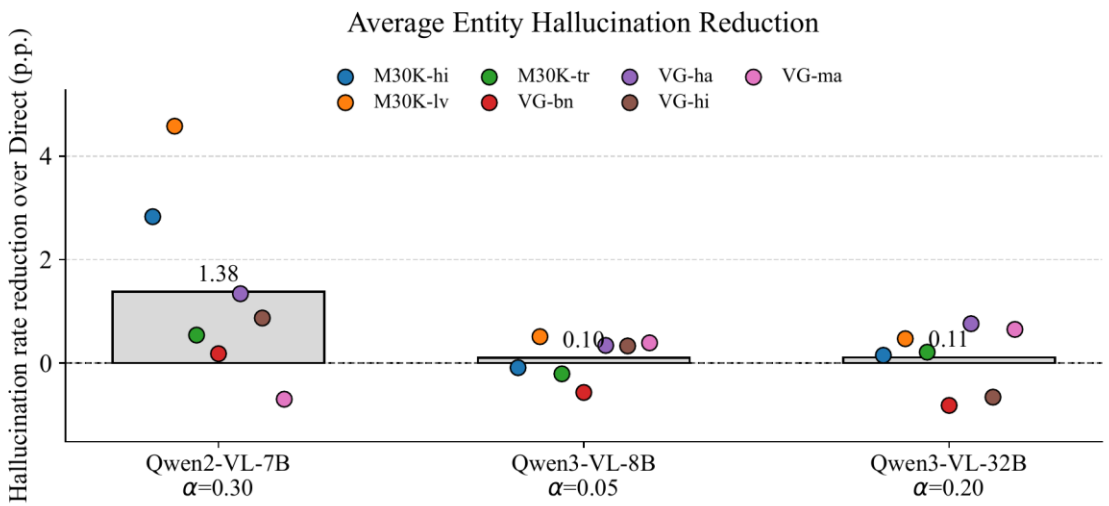
Qwen2-7B: α = 0.30 **Qwen3-8B: α = 0.05** **Qwen3-32B: α = 0.20**

Preferred contrast strength varies across backbones.
One α is fixed across datasets and languages for each backbone.

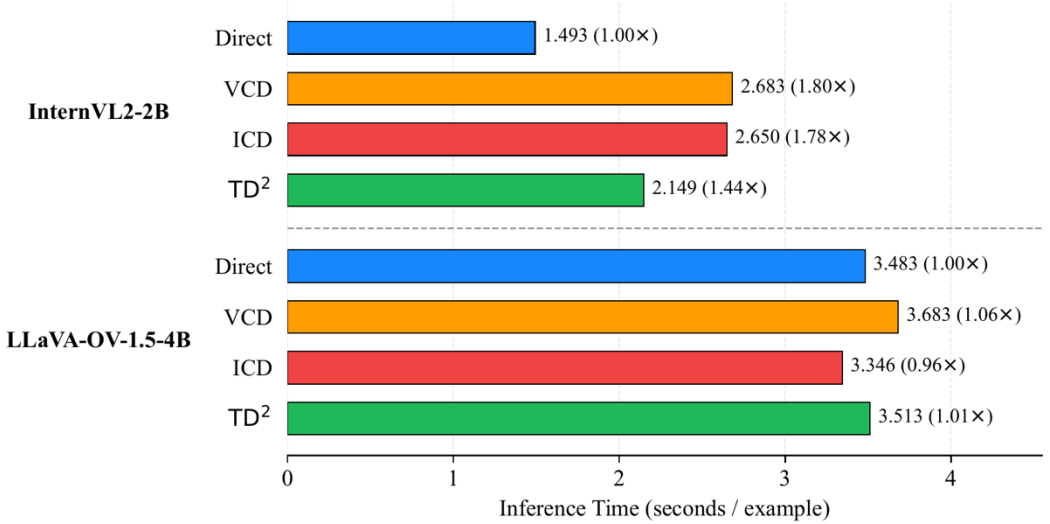
Vertical scales differ; scores average BLEU and COMET over seven dataset–language pairs.

Entity Hallucination and Inference Time

Entity hallucination reduction



Inference time per example



Mean reduction: 1.38 / 0.10 / 0.11 p.p. for Qwen2-7B / Qwen3-8B / Qwen3-32B.

Bars show means; dots show language-level results.
Some language pairs still show increased hallucination.

TD² / Direct: 1.44× on InternVL2-2B and 1.01× on LLaVA-OV-1.5-4B.

Empirical latency depends on output length and implementation.

GPT-5.5 entity evaluation; reduction in percentage points. Runtime in seconds/example.

Conclusion and Discussion

Takeaways

TD² calibrates image-prior logits without retraining.
It improves aggregate COMET on all five backbones.
Entity-level evidence supports mitigation in several settings.

Limits and next step

Fixed α, approximate prior estimation and extra inference cost.
Future work: adaptive calibration and broader domains.

Code

Text-Driven-Decoding

Corresponding author:
miradeljan51@xju.edu.cn

Thank you! Questions are welcome.

Supported by NSFC (Grant No. 62406316) and the Xinjiang “Tianchi Talent” Recruitment and Introduction Program.