TD^2: Text-Driven Decoding for Mitigating Image-Induced Hallucination in Low-Resource Multimodal Translation

Yi Chen[1] Mieradilijiang Maimaiti[1†] Pei Yu[1] Jiaxin Wang[1]
Shuoshuo Hou[1] Zhexin Li[1] Zhengyi Guo[1] Wushouer Silamu[1]
[1]School of Computer Science and Technology, Xinjiang University, China
{chenyi, peiyu, wangjiaxin, yumoya, lizhexin, guozhengyi}@stu.xju.edu.cn
{miradeljan51,wushour}@xju.edu.cn

Abstract

Large Vision-Language Models (LVLMs) have shown strong potential for multimodal machine translation by leveraging visual context. However, visual inputs can also introduce image-induced priors, leading to visually plausible but textually unfaithful translations, especially for low-resource target languages. We propose Text-Driven Decoding (TD^2), a simple, training-free method that calibrates next-token logits by contrasting those obtained from the full multimodal input with their image-only counterparts. By suppressing source-agnostic image-prior logits while preserving the full source-image context, TD^2 improves source-faithful translation without model retraining. Experiments on VG and Multi30K with five LVLM backbones spanning three model families show improvements over direct decoding, while comparisons on two non-Qwen backbones further demonstrate gains over VCD and ICD. Entity-level analyses further indicate reduced source-unsupported hallucinations in multiple settings. The anonymized source code is available at [1].

1 Introduction

Recent advances in Large Vision-Language Models (LVLMs) have created new opportunities for multimodal machine translation (MMT) (Zhang et al., 2026; Pei et al., 2026). MMT uses images as auxiliary context for source-conditioned translation, helping resolve lexical ambiguity, ground textual mentions in the visual scene, and complement textual information in the source sentences (Elliott et al., 2016; Caglayan et al., 2019; Elliott, 2018; Li et al., 2021; Shen et al., 2024). With the instruction-following ability of recent LVLMs, multimodal translation can be performed directly from image-text inputs without training task-specific MMT architectures (Wang et al., 2024a; Bai et al., 2025; Chen et al., 2024; An et al., 2025). However, visual inputs can also introduce image-induced priors, leading to visually plausible but source-unsupported content, especially for low-resource target languages. This is problematic for translation, where outputs must remain faithful to the source sentence rather than merely plausible under the image.

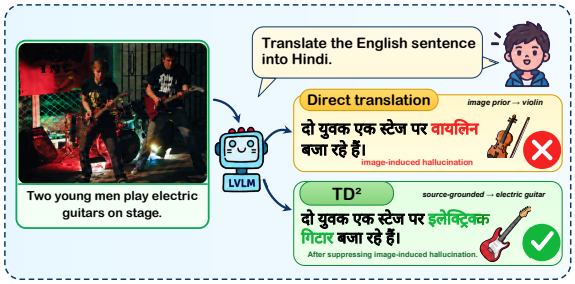


Figure 1: Comparison of Direct decoding and TD^2 for English-to-Hindi multimodal translation.

We frame source-unsupported generation in LVLM-based multimodal translation as *image-induced hallucination*. Unlike image captioning or visual question answering, MMT requires the source sentence to determine the primary semantics, with the image serving only as auxiliary evidence. Because direct multimodal decoding does not enforce this asymmetry, visual priors may override textual evidence and introduce unsupported entities.

We propose Text-Driven Decoding (TD^2) to promote source-faithful multimodal translation. As illustrated in Figure 1, TD^2 calibrates next-token logits by contrasting the full source-image distribution with an image-conditioned

†Corresponding author.

[1]https://github.com/chenyi240226-oss/Text-Driven-Decoding

prior computed without the source text. This contrast downweights visually favored but source-agnostic tokens while retaining the full multimodal distribution for decoding.

TD^2 is a training-free instance of inference-time logit calibration tailored to source-faithful MMT, requiring neither architectural modifications nor auxiliary verification models.

This work makes four main contributions.

- We identify image-induced priors as an important source of source-unsupported errors in LVLM-based multimodal translation.
- We introduce TD^2, a training-free method that mitigates source-agnostic visual priors through inference-time logit calibration.
- Across five backbones from three LVLM families, TD^2 achieves higher COMET than direct decoding on all five and higher BLEU on four.
- Comparisons with VCD and ICD demonstrate gains in both COMET and BLEU, while entity-level analyses show reduced source-unsupported hallucinations across multiple settings.

2 Related Work

Low-resource machine translation. Neural machine translation addresses data scarcity through transfer learning, pretrained lexical representations, and linguistically guided data augmentation (Maimaiti et al., 2019, 2020, 2021, 2022). More recently, large language models have been adapted to low-resource translation through prompting, parameter-efficient fine-tuning, and cross-lingual alignment (Zhang et al., 2023; Guo et al., 2024; Mao and Yu, 2024; Hou et al., 2026).

Multimodal machine translation. Multimodal machine translation uses images as auxiliary context for source-conditioned translation, especially when the source is ambiguous (Elliott et al., 2016; Caglayan et al., 2019; Shen et al., 2024). However, models do not always use visual evidence reliably, as their predictions may remain insensitive when images are perturbed or replaced with unrelated ones (Elliott, 2018; Li et al., 2021). This motivates a source-faithful setting where the source sentence defines the primary semantics and the image serves only as auxiliary evidence.

Algorithm 1 Text-Driven Decoding (TD^2)

Require: Source sentence x; image v; shared LVLM f_θ; contrast strength α; maximum length T

Ensure: Source-faithful translation y

Initialize the generated prefix $y_{<1} \leftarrow \varnothing$
for $t = 1$ to T **do**
  $\ell_t^{\text{full}} \leftarrow f_\theta(y_{<t}, x, v)$
  $\ell_t^{\text{image}} \leftarrow f_\theta(y_{<t}, v)$
  $\Delta_t \leftarrow \ell_t^{\text{full}} - \ell_t^{\text{image}}$ ▷ Text-driven contrast
  $\ell_t^{\text{TD}^2} \leftarrow \ell_t^{\text{full}} + \alpha\Delta_t$
  $p_t^{\text{TD}^2} \leftarrow \text{softmax}\left(\ell_t^{\text{TD}^2}\right)$
  $y_t \leftarrow \text{Decode}\left(p_t^{\text{TD}^2}\right)$
  if $y_t = [\text{EOS}]$ **then**
    break
  end if
end for
return $y = (y_1, \ldots, y_t)$

Hallucination in LVLMs. LVLMs may generate visually plausible but ungrounded content due to statistical bias, object co-occurrence, or unimodal priors (Li et al., 2023; Leng et al., 2024). In multimodal translation, such content becomes erroneous when it is not entailed by the source sentence, even if it is plausible under the image. This work therefore focuses on source-unsupported entity hallucination in LVLM-based translation, where the key criterion is source faithfulness rather than visual plausibility alone.

Contrastive decoding. Contrastive decoding contrasts a desired distribution with a weaker or biased one to improve generation (Waldendorf et al., 2024). VCD contrasts clean and distorted visual inputs, whereas ICD contrasts standard and disturbance-instruction conditions to suppress image-grounding errors (Leng et al., 2024; Wang et al., 2024b). TD^2 follows this general inference-time calibration principle, but adapts it to the source-faithfulness constraint of multimodal translation. Specifically, it contrasts the full source-image condition with

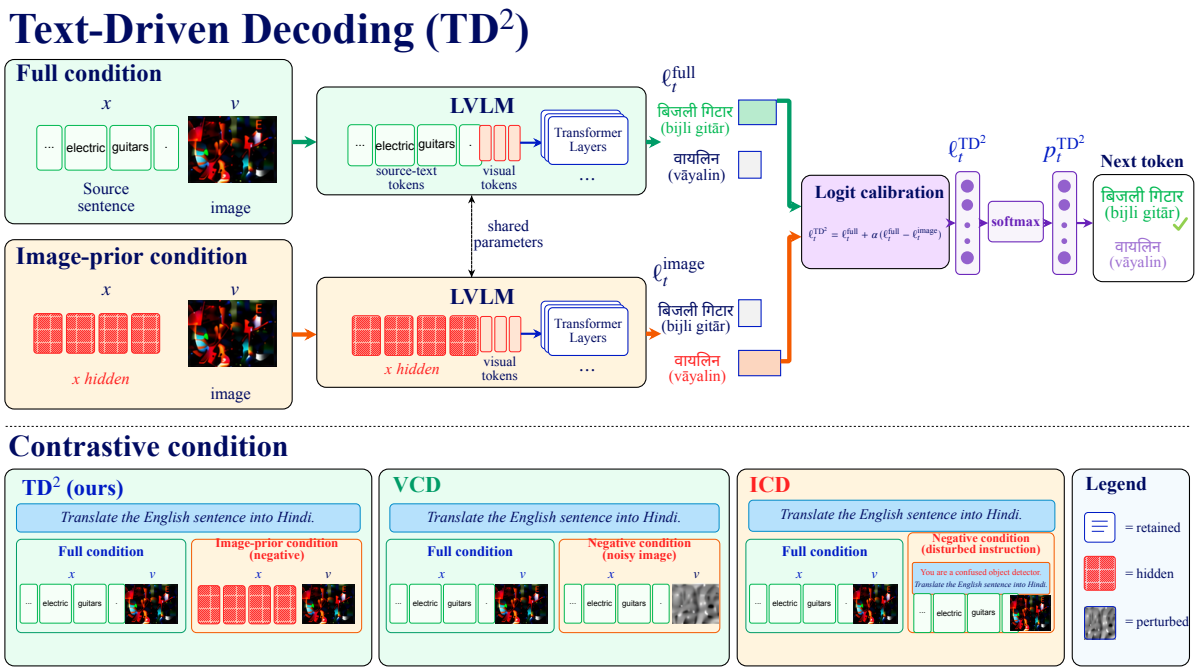


Figure 2: The architecture of Text-Driven Decoding (TD^2). At each decoding step, the full source–image condition and the source-hidden image-prior condition are processed by the same LVLM. Their next-token logits are calibrated to suppress source-agnostic visual priors. The bottom panel compares the contrastive conditions used by TD^2, VCD, and ICD.

a source-hidden image-prior condition within the same LVLM, suppressing source-agnostic image-prior logits while preserving the original source-conditioned multimodal decoding path.

3 Method

3.1 Overview

In LVLM-based multimodal translation, visual context may introduce source-agnostic priors that lead to source-unsupported token choices. TD^2 calibrates decoding by contrasting full source-image logits with source-hidden image-prior logits, thereby penalizing visually plausible but source-unfaithful predictions. The overall framework is illustrated in Figure 2.

3.2 Text-Driven Decoding

We first formalize standard multimodal decoding, which provides the base distribution.

At decoding step t, a standard multimodal LVLM predicts the next-token logits conditioned on the generated prefix $y_{<t}$, the source sentence x, and the paired image v:

$$\ell_t^{\text{full}} = f_\theta(y_{<t}, x, v), \tag{1}$$

where f_θ is the LVLM and ℓ_t^{full} denotes the full multimodal logits.

To estimate image-induced priors, we construct an image-prior condition by removing the source sentence:

$$\ell_t^{\text{image}} = f_\theta(y_{<t}, v). \tag{2}$$

This condition approximates source-agnostic token preferences induced by the visual context.

TD^2 calibrates the next-token logits by contrasting the full multimodal logits with the image-prior logits:

$$\ell_t^{\text{TD}^2} = \ell_t^{\text{full}} + \alpha \left(\ell_t^{\text{full}} - \ell_t^{\text{image}} \right), \tag{3}$$

where $\alpha \geq 0$ controls the contrast strength. When $\alpha = 0$, TD^2 reduces to standard full multimodal decoding. A larger α more strongly penalizes tokens that are disproportionately favored by the image-prior condition relative to the full source-image condition.

The calibrated next-token distribution is then computed as follows:

$$p_t^{\text{TD}^2} = \text{softmax}\left(\ell_t^{\text{TD}^2} \right). \tag{4}$$

The next token is generated from $p_t^{\text{TD}^2}$ using the same decoding strategy as the base LVLM.

Method	ha				ma				bn				hi			
	BLEU	COMET	BLEURT	chrF++	BLEU	COMET	BLEURT	chrF++	BLEU	COMET	BLEURT	chrF++	BLEU	COMET	BLEURT	chrF++
Qwen2-VL-7B-Instruct																
Direct	2.60	42.15	**24.12**	9.35	**4.89**	49.67	31.87	**11.96**	14.96	62.72	42.02	**25.80**	17.03	57.34	45.87	26.34
TD^2	**3.78****	**45.81****	23.89	**11.15****	4.64	**50.82****	**33.25****	10.44	**15.06**	**63.75**	**42.85**	25.00	**18.15****	**58.16**	**46.87**	**27.18**
Qwen3-VL-8B-Instruct																
Direct	4.67	52.41	26.28	**14.40**	16.62	69.44	54.51	32.50	25.49	**78.25**	**62.50**	**42.54**	30.49	75.13	67.09	47.94
TD^2	**4.81**	**52.64**	**26.66**	14.35	**17.00**	**69.71**	**54.80**	**32.74**	**25.57**	78.05	62.05	42.06	**30.96**	**75.17**	**67.22**	**48.30**
Qwen3-VL-32B-Instruct																
Direct	5.46	53.22	29.04	16.42	20.92	73.76	61.99	37.05	**30.75**	**82.08**	**68.16**	**47.34**	35.36	77.86	70.17	53.36
TD^2	**6.23****	**55.10****	**29.89**	**17.54****	**21.46**	**74.38**	**62.92***	**37.26**	29.79	81.84	67.99	46.28	**35.87**	**77.88**	**70.49**	**53.43**

Table 1: Results on VG for EN→XX translation using BLEU, COMET, BLEURT, and chrF++. Bold numbers indicate the best result. Stars indicate statistically significant improvements of TD^2 over direct decoding after Holm correction: $^{*}p < 0.05$ and $^{**}p < 0.01$.

Method	hi				tr				lv			
	BLEU	COMET	BLEURT	chrF++	BLEU	COMET	BLEURT	chrF++	BLEU	COMET	BLEURT	chrF++
Qwen2-VL-7B-Instruct												
Direct	12.04	51.43	46.06	30.79	13.76	76.08	**60.83**	43.52	6.84	45.76	33.49	22.60
TD^2	**12.83****	**52.63****	**47.85****	**32.75****	**14.26**	**76.23**	60.78	**44.25**	**7.86****	**48.87****	**35.93****	**25.37****
Qwen3-VL-8B-Instruct												
Direct	**21.94**	**69.18**	**61.49**	47.55	21.05	**84.97**	**72.60**	**54.42**	16.25	74.14	63.10	44.11
TD^2	21.90	69.05	61.40	**47.67**	**21.41**	84.87	72.47	54.41	**16.66**	**74.38**	**63.15**	**44.64***
Qwen3-VL-32B-Instruct												
Direct	25.64	71.78	64.11	51.85	24.64	87.59	**76.78**	58.32	21.83	80.84	70.96	51.80
TD^2	**25.97**	**72.00**	**64.29**	**52.16**	**25.44***	**87.60**	76.70	**59.04****	**22.48***	**81.02**	**71.14**	**52.51****

Table 2: Results on Multi30K for EN→XX translation using BLEU, COMET, BLEURT, and chrF++. Bold numbers indicate the better result. Stars indicate statistically significant improvements of TD^2 over direct decoding after Holm correction: $^{*}p < 0.05$ and $^{**}p < 0.01$.

Since TD^2 only modifies logits at inference time, it requires no additional training, external tools, auxiliary models, or architectural changes. It can therefore be applied to different LVLM backbones and translation directions without model-specific training.

4 Experiments

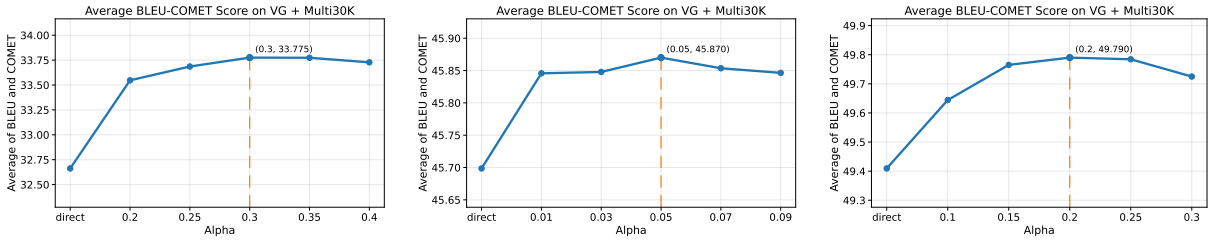
4.1 Setup

We evaluate TD^2 on EN→XX multimodal translation, where XX denotes the target language, using VG (Krishna et al., 2017) and Multi30K (Guo et al., 2022). VG provides object-rich visual contexts for analyzing source-unsupported entity generation and includes Hausa (ha), Malayalam (ma), Bengali (bn), and Hindi (hi), with 1,595 examples per language. Multi30K is a standard multimodal translation benchmark covering Hindi (hi), Turkish (tr), and Latvian (lv), with 1,000 examples per language. Each example includes an image, an English source sentence, and a reference translation in the target language.

Our main evaluation uses Qwen2-VL-7B-Instruct (Wang et al., 2024a), Qwen3-VL-8B-Instruct, and Qwen3-VL-32B-Instruct (Bai et al., 2025). We further test InternVL2-2B (Chen et al., 2024) and LLaVA-OneVision-1.5-4B-Instruct (An et al., 2025), covering five backbones from three LVLM families. For the two non-Qwen backbones, we additionally compare against Visual Contrastive Decoding (VCD) (Leng et al., 2024) and Instruction Contrastive Decoding (ICD) (Wang et al., 2024b). We use BLEU and COMET as the main metrics (Papineni et al., 2002; Rei et al., 2022).

4.2 Main Results

We present the main EN→XX results on VG and Multi30K in Tables 1 and 2. Compared with Direct decoding, TD^2 yields higher macro-averaged BLEU and COMET across the three Qwen-VL backbones. The largest aggregate differences occur on Qwen2-VL-7B-Instruct, the backbone that also has the highest entity-hallucination rate under Direct decoding. The aggregate differences are smaller on the two Qwen3-VL backbones and vary by



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

Figure 3: Sensitivity of TD^2 to the contrast coefficient α. We report the averaged BLEU–COMET score on VG and Multi30K across three LVLM backbones. Dashed vertical lines indicate the model-specific α selected for the main experiments.

Backbone	Method	BLEU	COMET
InternVL2-2B	Direct	**5.00**	37.84
	VCD	4.16	37.61
	ICD	3.94	37.46
	TD^2	4.88	**38.84**
LLaVA-OV-1.5-4B	Direct	15.51	65.02
	VCD	14.08	62.27
	ICD	13.04	60.77
	TD^2	**16.61**	**65.25**

Table 3: Cross-backbone comparison, macro-averaged over all seven dataset–language pairs. Detailed results appear in Appendix B.1.

language pair. Thus, the results support the effectiveness of TD^2 as a training-free, inference-time calibration method across the evaluated datasets, languages, and backbones.

Generalization beyond Qwen-VL. Results for InternVL2-2B and LLaVA-OneVision-1.5-4B-Instruct are summarized in Table 3. TD^2 achieves the highest macro-averaged COMET on both backbones and the highest BLEU on LLaVA-OneVision. Compared with direct decoding, it increases COMET by 0.99 despite a marginal BLEU decrease of 0.12 on InternVL2, while improving BLEU and COMET by 1.10 and 0.22, respectively, on LLaVA-OneVision. It also outperforms VCD and ICD in both aggregate metrics on both backbones. After Holm correction, paired tests indicate statistically significant improvements in 9 of 14 comparisons on InternVL2 and 7 of 14 on LLaVA-OneVision. Language-level results and significance markers are provided in Appendix B.1.

4.3 Analysis

The effect of α. We examine sensitivity to α using averaged BLEU–COMET scores across VG and Multi30K, as shown in Figure 3. Because TD^2 operates only at inference time, the LVLM parameters remain frozen, and larger values of α impose stronger penalties on tokens favored by the source-hidden image-prior condition. We select a single α for each backbone and keep it fixed across datasets and languages. Specifically, we use $\alpha = 0.30$ for Qwen2-VL-7B, $\alpha = 0.05$ for Qwen3-VL-8B, and $\alpha = 0.20$ for Qwen3-VL-32B. For each candidate, we average BLEU and COMET across the seven dataset–language pairs and then average the two metrics, both of which are reported on a 0–100 scale. The results indicate that preferred contrast strength varies across backbones; we provide complete sensitivity curves and language-level results in Appendix B.2.

The effect of visual-prior signals. We compare TD^2 with Text-only and Image-prior decoding to examine whether the gains come from simply removing visual input.

Model	Direct	Text-only	Img-prior	TD^2
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**

Table 4: Ablation on visual-prior suppression. Scores are averaged BLEU-COMET values over VG and Multi30K. Text removes the image, Img-prior hides the source sentence, and TD^2 preserves the full source-image context while suppressing image-prior logits.

Text-only decoding removes the image to assess the contribution of visual context. Image-prior decoding omits the source sentence while

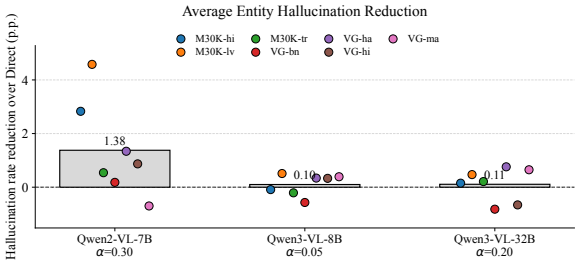


Figure 4: Entity hallucination reduction over Direct decoding. Bars show average reduction across dataset-language pairs, and dots indicate individual results.

retaining the image and target-language instruction, thereby estimating generation preferences that arise without source-text conditioning. As summarized in Table 4, Image-prior decoding performs far worse than source-conditioned decoding, indicating that the source sentence remains essential for translation. Meanwhile, Text-only remains competitive with Direct decoding, showing that visual input is not always beneficial when used without calibration. TD^2 achieves the highest averaged BLEU–COMET score across all three backbones. This is consistent with suppressing harmful image-prior signals while retaining the full source–image condition, although the ablation does not isolate which individual visual cues are beneficial. Table 7 provides the aggregate and language-level results.

Hallucination mitigation. We further evaluate whether TD^2 reduces source-unsupported entity hallucination. We measure entity hallucination reduction as the hallucination rate of Direct decoding minus that of TD^2, using GPT-5.5 as the main evaluator. As illustrated in Figure 4, TD^2 reduces entity hallucination across several dataset–language pairs, with the largest mean reduction on Qwen2-VL-7B. The reductions are smaller on the two Qwen3-VL backbones, so the evidence supports mitigation in the reported settings rather than a uniform reduction across models and languages. To reduce reliance on a single LLM-based evaluator, we also run a complementary Stanza-based, reference-grounded entity-matching analysis on the supported Multi30K languages. Details are provided in Appendix C.

Inference time. Figure 5 reports the macro-averaged wall-clock time over the language-level measurements. Because latency depends on the generated sequence length and implementation details, interpret these measurements as an empirical cost profile rather than an architecture-independent complexity comparison.

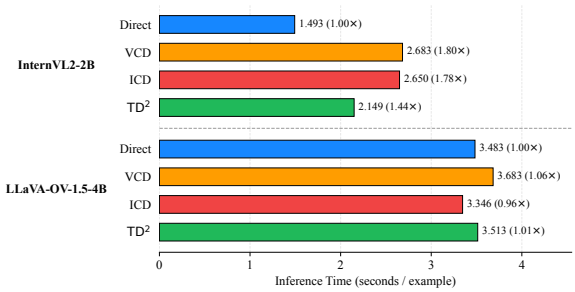


Figure 5: Macro-averaged inference time per example across languages. Relative cost is normalized to direct decoding within each backbone.

5 Conclusion

We propose TD^2, a training-free decoding method that downweights tokens favored by source-agnostic visual priors for source-faithful LVLM-based multimodal translation. Across VG and Multi30K, TD^2 yields higher macro-averaged COMET than Direct decoding on all five evaluated backbones and higher BLEU on four of five. It also exceeds VCD and ICD in both metrics on the two non-Qwen backbones. Entity-level analyses indicate reduced source-unsupported hallucination in multiple, but not all, settings. Within the evaluated scope, these results support inference-time visual-prior calibration as a training-free approach to improving source faithfulness in multimodal translation.

6 Limitations

Fixed contrast strength. The contrast coefficient α is selected separately for each model and then held fixed across datasets and target languages. Although this design simplifies deployment, the optimal contrast strength may vary across language pairs, domains, and individual inputs. Adaptive or instance-level calibration remains a direction for future work.

Approximate visual-prior estimation. TD^2 uses the source-hidden condition as a

proxy for source-agnostic visual preferences. This approximation cannot perfectly separate harmful image-induced priors from visual evidence that helps resolve source ambiguity. The contrastive adjustment may therefore occasionally downweight useful visual information, particularly when the source text is underspecified.

Benchmark and model coverage. Our evaluation is limited to VG and Multi30K, which primarily contain short, image-grounded text, while entity-level analysis covers only the three Qwen-VL backbones. We also observe smaller gains on larger backbones, although differences in model architecture and training preclude attributing this trend solely to parameter count. Generalization to longer texts, broader domains, more diverse languages, and larger LVLMs requires further investigation.

Evaluation scope and inference cost. BLEU, COMET, and entity-level hallucination measures may not capture all forms of semantic distortion or source–translation inconsistency. Moreover, TD^2 computes both the full multimodal distribution and the source-hidden image-prior distribution at each decoding step, increasing inference latency and computational cost relative to direct decoding.

7 Ethical considerations

This work uses public multimodal translation benchmarks and publicly available LVLMs. No private, user-generated, or personally identifiable data is introduced.

Although TD^2 aims to reduce source-unsupported hallucination, model outputs may still contain mistranslations or hallucinated content, especially for low-resource languages. Such outputs should not be used in high-stakes settings without human verification.

Acknowledgements

We are grateful to our collaborators and anonymous reviewers for their valuable contributions. This work was supported by the National Natural Science Foundation of China (Grant No.62406316), and the Xinjiang 'Tianchi Talent' Recruitment and Introduction Program (awarded to Mieradilijiang Maimaiti).

References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, and 4 others. 2025. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Hongcheng Guo, Jiaheng Liu, Haoyang Huang, Jian Yang, Zhoujun Li, Dongdong Zhang, and Zheng Cui. 2022. LVP-M3: Language-aware visual prompt for multilingual multimodal machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural*

Language Processing, pages 2862–2872, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and Heyan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697, Torino, Italia. ELRA and ICCL.

Shuoshuo Hou, Mieradilijiang Maimaiti, Zhexin Li, Jiaxin Wang, Ahmad Hassan, Kaishaer Jiapaer, Nilufar Abdurakhmonova, Roza Urinbayeva, Madina Mansurova, Shormakova Assem, Gulnar Murat, Wu Le, and Wushour Silamu. 2026. Vision-to-text: Benchmarking multimodal LLMs on extremely low-resource languages. In *2026 International Joint Conference on Neural Networks (IJCNN)*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision matters when it should: Sanity checking multimodal machine translation models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, Zegao Pan, and Maosong Sun. 2021. Improving data augmentation for low-resource NMT guided by pos-tagging and paraphrase embedding. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20(6):107:1–107:21.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2019. Multi-round transfer learning for low-resource NMT using multiple high-resource languages. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 18(4):38:1–38:26.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2022. Data augmentation for low-resource languages NMT guided by constrained sampling. *Int. J. Intell. Syst.*, 37(1):30–51.

Mieradilijiang Maimaiti, Yang Liu, Maosong Sun, and Huanbo Luan. 2020. Enriching the transfer learning with pre-trained lexicon embedding for low-resource neural machine translation. *Tsinghua Science and Technology*, page 1.

Zhuoyuan Mao and Yen Yu. 2024. Tuning LLMs with contrastive alignment instructions for machine translation in unseen, low-resource languages. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 1–25, Bangkok, Thailand. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yu Pei, Mieradilijiang Maimaiti, Yi Chen, Wu Le, Zhuofei Xie, and Jiawei Chen. 2026. Dmg-rag: Dynamic multi-grained retrieval augmented generation for multi-hop qa. In *2026 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 578–585, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Huangjun Shen, Liangying Shao, Wenbo Li, Zhibin Lan, Zhanyu Liu, and Jinsong Su. 2024. A survey on multi-modal machine translation:

Tasks, methods and challenges. *arXiv preprint arXiv:2405.12669*.

Jonas Waldendorf, Barry Haddow, and Alexandra Birch. 2024. Contrastive decoding reduces hallucinations in large multilingual machine translation models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2539, St. Julian's, Malta. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15840–15853, Bangkok, Thailand. Association for Computational Linguistics.

Shibo Zhang, Mieradilijiang Maimaiti, Zhengyi Guo, Dezhi Wang, Wu Le, Zhuofei Xie, Jiawei Chen, and Wushouer Silamu. 2026. Retrieve, refine, and translate: Llm-based translation for low-resource languages. In *2026 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

Appendix

A Implementation Details

Input fields. For VG, we use `src_en` as the English source and `tgt` as the reference. For Multi30K, we use `source` and `target`, respectively. References are used only for evaluation and are never provided during generation.

Prompt templates. We use the prompts in Table 5. The full multimodal and image-prior conditions share the same system prompt, differing only in whether the source sentence is visible.

Decoding implementation. Direct decoding uses greedy `model.generate` with `do_sample=False`, `use_cache=True`, and `max_new_tokens=64`. TD^2 uses manual greedy decoding over calibrated logits. We verify that the manual decoder exactly reproduces Direct decoding when $\alpha = 0$, ensuring that differences come from logit calibration rather than implementation details. We set `HEAD_BETA=0` and `MIN_NEW_TOKENS=0`. Images are processed with the processor configuration released with each LVLM checkpoint, without additional task-specific image preprocessing.

For the cross-family evaluation, VCD and ICD use their released contrastive formulations with the same checkpoints, prompts, and maximum generation length as Direct. We use a single TD^2 coefficient for each non-Qwen backbone, $\alpha = 0.5$ for InternVL2-2B and $\alpha = 0.2$ for LLaVA-OneVision-1.5-4B-Instruct, and keep it fixed across both datasets and all languages.

Evaluation implementation. We evaluate each generated translation against its target-language reference. For COMET, the English source sentence is used as the source input, and the reference translation is used as the target reference. We use SacreBLEU to compute sentence-level BLEU scores and report their arithmetic mean. COMET is computed with `wmt22-comet-da` and multiplied by 100 for readability. All reported decoding results are from a single deterministic greedy-decoding run. We do not report error bars over repeated sampling runs, and we explicitly indicate when scores are averaged across dataset–language pairs.

Statistical testing. For the InternVL2 and LLaVA-OneVision comparisons, we align Direct and TD^2 outputs by sample identifier and verify that their source, reference, and image fields match. The paired analysis contains 900 Multi30K examples and 1,595 VG examples per language. For BLEU, we test the difference in mean sentence-level scores using a two-sided paired bootstrap with 10,000 resamples and random seed 42. We apply the plus-one correction when estimating p-values and report 95% percentile confidence intervals. For COMET, we use a two-sided paired Wilcoxon signed-rank test with zero differences removed. We apply Holm's step-down correction separately to the seven dataset–language comparisons within each model and metric. Stars mark only positive differences with adjusted $p < 0.05$.

Averaged BLEU-COMET score. For the reported α sensitivity analysis, we average BLEU and COMET over all evaluated dataset–language pairs, and then take the mean of the two metrics. Both metrics are on a 0–100 scale.

Hardware. All experiments are conducted in offline mode using local model checkpoints on NVIDIA H100 GPUs. We use mixed-precision inference for LVLM decoding.

Artifact licenses and terms. We use publicly available datasets, models, and evaluation

Prompt Templates

Component	Prompt

Shared system prompt	`/no think`
	`You are a professional translator.`
	`Your job is to translate the given English source sentence into {target_language}.`
	`STRICT RULES:`
	`1) Faithful: Do NOT add, remove, or change objects, attributes, actions, or relations.`
	`2) Use the image only when it helps resolve ambiguity.`
	`3) Do not turn the task into image captioning.`
	`4) Output ONLY the translation in {target_language}. No explanations.`
	`5) NEVER output any <think>...</think>content.`

Full multimodal condition	`[Image]`
	`English SOURCE:`
	`{source_text}`
	`Translate into {target_language}.`

Image-prior condition	`[Image]`
	`English SOURCE:`
	`[SOURCE SENTENCE HIDDEN]`
	`Translate into {target_language}.`

Table 5: Prompt templates used in Direct decoding and TD^2.

tools following their original licenses and terms of use. VG and Multi30K are used solely for research evaluation. Qwen-VL backbones are used through publicly released model checkpoints under their corresponding model licenses. SacreBLEU, COMET, and Stanza are used as evaluation and analysis tools under their respective open-source licenses. We do not redistribute the original datasets, model checkpoints, or third-party evaluation tools.

B Additional Translation Results

B.1 Cross-Family Evaluation

Table 6 reports language-level results for InternVL2-2B and LLaVA-OneVision-1.5-4B-Instruct. We use one model-level coefficient across all datasets and languages: $\alpha = 0.5$ for InternVL2-2B and $\alpha = 0.2$ for LLaVA-OneVision. VCD and ICD follow their released inference-time decoding formulations under the same model checkpoints, inputs, prompts, and evaluation pipeline as Direct and TD^2.

Statistically significant improvements over direct decoding are marked with * for Holm-adjusted $p < 0.05$ and ** for Holm-adjusted $p < 0.01$. Full testing details are provided in Appendix A. On InternVL2-2B, 9 of the 14 BLEU/COMET comparisons are significant improvements, while three are significant degradations. On LLaVA-OneVision, 7 of the 14 comparisons show significant improvements, while two COMET comparisons show significant degradations. These results show gains in a majority of comparisons, but also confirm that the effect is not uniform across languages and metrics.

B.2 Hyperparameter Sensitivity

This section reports the language-level results under the candidate contrast coefficients used for analyzing the effect of α. To keep TD^2 simple, we use a single backbone-level coefficient for each LVLM rather than tuning α separately for each dataset, language, or example. The selected coefficient is then fixed across all dataset–language pairs for the corresponding backbone.

We provide the language-level results for Qwen2-VL-7B-Instruct, Qwen3-VL-8B-Instruct, and Qwen3-VL-32B-Instruct in Tables 11, 12, and 13, respectively. For each backbone, we also report the average over all dataset–language pairs in the last row and highlight the best average value in bold. Based on this model-level analysis, we use $\alpha = 0.30$ for Qwen2-VL-7B-Instruct, $\alpha = 0.05$ for Qwen3-VL-8B-Instruct, and $\alpha = 0.20$ for Qwen3-VL-32B-Instruct. Because the coefficients are selected from the reported aggregate sensitivity results, this setting should be interpreted as a coarse backbone-level analysis rather than an independently validated hyperparameter choice.

B.3 Language-Level Visual-Prior Ablation

Aggregate ablation results are reported in Table 4 in the main text. Table 7 provides the corresponding language-level breakdown for text-only and image-prior decoding.

C Entity Hallucination Evaluation

This section reports the entity-level statistics used in our hallucination analysis. We use GPT-5.5 as the main evaluator to apply one evaluation protocol across datasets, languages, and LVLM backbones. Given the source sentence, reference translation, and model output, GPT-5.5 identifies entity mentions in the output that are supported by neither the source sentence nor the reference translation. The hallucination rate is computed as the number of hallucinated entity instances divided by the number of reference entities.

To reduce reliance on a single evaluator, we also report a Stanza-based reference-grounded entity matcher on the supported Multi30K languages. This matcher serves as a complementary consistency check rather than the main metric, since it has limited language coverage and does not cover all cross-lingual entity variants.

GPT-5.5 evaluation prompt. Table 8 shows the prompt template used for GPT-5.5-based entity hallucination evaluation. The model is instructed to focus only on source- or reference-unsupported entity mentions, while ignoring attributes, harmless wording differences, and grammatical variations.

Dataset	Lang.	Direct		VCD		ICD		TD^2	
		BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
InternVL2-2B									
Multi30K	lv	4.34	29.98	4.37	29.79	4.48	30.15	**4.58***	**30.98****
	tr	6.05	37.82	5.97	37.22	5.98	36.85	**6.90****	**40.64****
	hi	**4.47**	29.88	3.91	29.13	3.61	28.61	4.28	**30.38****
VG	ha	1.52	35.15	1.52	38.13	1.57	**39.21**	**3.24****	**39.21****
	ma	**3.16**	**44.95**	2.06	43.97	1.50	43.11	1.23	41.76
	bn	**5.89**	46.63	3.36	45.85	3.04	45.66	4.53	**47.35****
	hi	**9.58**	40.51	7.95	39.17	7.37	38.64	9.41	**41.52****
LLaVA-OneVision-1.5-4B-Instruct									
Multi30K	lv	9.75	56.99	8.73	52.02	7.92	49.52	**10.18**	**58.68****
	tr	17.61	**80.28**	16.26	77.42	16.27	75.89	**17.75**	79.60
	hi	19.09	64.37	16.59	60.67	15.37	58.85	**20.64****	**64.91***
VG	ha	1.86	47.53	1.98	50.08	1.67	**50.76**	**2.67****	46.99
	ma	11.86	62.28	9.76	57.90	8.08	55.65	**12.61**	**62.62**
	bn	21.62	**73.16**	19.61	70.02	17.58	68.26	**23.42****	**73.16**
	hi	26.81	70.56	25.60	67.81	24.41	66.48	**29.02****	**70.77***

Table 6: Language-level cross-backbones results. Bold indicates the best result. Stars indicate statistically significant improvements of TD^2 over direct decoding after Holm correction: $^{*}p < 0.05$ and $^{**}p < 0.01$.

GPT-5.5-based entity hallucination statistics. We provide detailed GPT-5.5-based entity hallucination statistics for all three Qwen-VL backbones in Table 9. For each dataset-language pair, we record the number of hallucinated entity instances, the number of reference entities, the hallucination rate, and the hallucination-rate change from Direct decoding to TD^2. Downward arrows indicate reduced hallucination rates under TD^2, while upward arrows indicate increased rates.

Because the evaluation follows a strict source- and reference-grounded entity criterion, the absolute hallucination rates should be interpreted as evaluator-specific entity mismatch rates rather than human-calibrated error rates. This is especially relevant for low-resource target languages, where entity translation is more prone to lexical variation, untranslated forms, and source-unsupported substitutions. Therefore, our main analysis focuses on the relative change between Direct decoding and TD^2 under the same evaluator.

Complementary Stanza-based entity matching. We further conduct a complementary Stanza-based reference-grounded entity matching analysis on Multi30K. (Qi et al., 2020) The automatic matcher extracts noun and proper-noun spans from the reference and model output. An output entity is marked as hallucinated if it cannot be matched to the reference through exact surface-form, lemma-based, phrase-level, fuzzy, or alias-equivalence matching. This matching protocol is intentionally strict: an output entity is counted as hallucinated unless it can be grounded in the reference through one of the supported equivalence checks. The hallucination rate is normalized by the number of reference entities. We use this analysis as an automatic consistency check for the GPT-5.5-based evaluation and provide the results for the supported languages, Hindi, Latvian, and Turkish, in Table 10.

C.1 Case Study

Figure 6 presents a representative Hindi example. Direct decoding introduces a visually plausible but source-unsupported entity, while TD^2 preserves the source-grounded entity.

Dataset	Lang.	Method	Qwen2-7B		Qwen3-8B		Qwen3-32B	
			BLEU	COMET	BLEU	COMET	BLEU	COMET
VG	ha	Text-only	3.02	43.15	4.58	52.62	5.53	53.72
		Img-prior	0.55	34.94	1.51	42.53	0.78	41.64
	bn	Text-only	14.99	63.64	25.42	77.83	30.08	82.18
		Img-prior	0.66	37.78	1.07	47.13	2.61	51.75
	ma	Text-only	5.45	51.43	16.88	69.81	20.64	73.52
		Img-prior	0.00	31.98	2.14	47.35	3.07	48.89
	hi	Text-only	18.47	58.65	29.95	75.06	33.82	77.76
		Img-prior	0.88	33.34	1.31	36.96	2.77	42.51
Multi30K	hi	Text-only	13.03	52.46	22.11	69.24	25.57	71.92
		Img-prior	0.95	25.45	1.90	33.73	5.27	48.01
	lv	Text-only	7.21	46.99	16.18	73.65	21.82	80.23
		Img-prior	2.35	35.19	2.57	42.37	2.96	57.10
	tr	Text-only	14.86	76.86	21.66	84.53	25.08	84.66
		Img-prior	1.61	37.00	2.70	46.40	3.12	62.24

Table 7: Detailed language-level ablation results for Text-only and Image-prior decoding. Text-only removes the image, while Img-prior hides the source sentence and keeps the image with the target-language instruction.

GPT-5.5 Entity Hallucination Evaluation Prompt

```
You are an evaluator for entity hallucination in
multimodal machine translation.
Given an English source sentence, a reference
translation, and a model translation, identify
entity-level hallucinations in the model translation.

An entity hallucination is an entity mention in the
model translation that is supported by neither the
source sentence nor the reference translation.

Count hallucinations only for entity mentions,
including objects, people, animals, places, named
entities, and numbers.
Do not count entity attributes, harmless paraphrases,
grammatical differences, word-order differences, or
strictly equivalent synonyms as hallucinations.

Return the result in JSON format with the following
fields:
{
  "reference_entities": […],
  "hallucinated_entities": […],
  "hallucinated_entity_count": integer,
  "reason": "brief explanation"
}

English source: {source_text}
Reference translation: {reference_translation}
Model translation: {model_translation}
```

Table 8: Prompt template for GPT-5.5-based entity hallucination evaluation.

Case Study: Source-Unsupported Entity Hallucination



Source.
Two young men play electric guitars on stage.
Reference.
दो युवक मंच पर इलेक्ट्रिक गिटार बजाते हैं।
do yuvak manch par ilekṭrik giṭār bajāte haiṃ. (Two young men play electric guitars on stage.)
Direct decoding.
दो युवक एक स्टेज पर वायलिन बजा रहे हैं।
do yuvak ek sṭej par vāyalin bajā rahe haiṃ. (Two young men are playing the violin on a stage.)
TD^2.
दो युवक एक स्टेज पर इलेक्ट्रिक गिटार बजा रहे हैं।
do yuvak ek sṭej par ilekṭrik giṭār bajā rahe haiṃ. (Two young men are playing electric guitars on a stage.)
Observation.
Direct decoding incorrectly translates the source entity "electric guitars" as "violin", introducing a source-unsupported entity hallucination. By suppressing image-only visual priors, TD^2 restores the source-supported entity "electric guitars" and produces a more faithful translation.

Figure 6: Hindi case study: Direct hallucinates "violin", whereas TD^2 preserves "electric guitars".

Model	Dataset	Lang.	Direct H./Total	Direct HR	TD^2 H./Total	TD^2 HR	ΔHR
Qwen2-7B	Multi30K	hi	3249/4622	70.29%	3118/4622	67.46%	↓ **2.83**
		lv	3880/4907	79.07%	3655/4907	74.49%	↓ **4.58**
		tr	2083/4695	44.37%	2058/4695	43.83%	↓ **0.54**
	VG	bn	1893/2800	67.61%	1888/2800	67.43%	↓ **0.18**
		ha	2817/2925	96.31%	2778/2925	94.97%	↓ **1.34**
		hi	2275/3348	67.95%	2246/3348	67.08%	↓ **0.87**
		ma	3712/4022	92.29%	3740/4022	92.99%	↑ 0.70
Qwen3-8B	Multi30K	hi	1871/4622	40.46%	1874/4622	40.55%	↑ 0.06
		lv	1945/4907	39.64%	1920/4907	39.13%	↓ **0.51**
		tr	1529/4695	32.57%	1539/4695	32.78%	↑ 0.21
	VG	bn	1466/2800	52.19%	1482/2800	52.76%	↑ 0.57
		ha	2835/2925	96.92%	2825/2925	96.58%	↓ **0.34**
		hi	601/3348	17.95%	590/3348	17.62%	↓ **0.33**
		ma	2538/4022	63.10%	2522/4022	62.71%	↓ **0.39**
Qwen3-32B	Multi30K	hi	1741/4622	37.67%	1734/4622	37.52%	↓ **0.15**
		lv	2058/4907	41.94%	2035/4907	41.47%	↓ **0.47**
		tr	1201/4695	25.58%	1191/4695	25.37%	↓ **0.21**
	VG	bn	1249/2800	44.61%	1272/2800	45.43%	↑ 0.82
		ha	2683/2925	91.73%	2661/2925	90.97%	↓ **0.76**
		hi	953/3348	28.46%	975/3348	29.12%	↑ 0.66
		ma	2375/4022	59.05%	2349/4022	58.40%	↓ **0.65**

Table 9: GPT-5.5-based entity hallucination statistics for the three Qwen-VL backbones. H./Total denotes hallucinated entity instances over reference entities; ΔHR is Direct HR minus TD^2 HR.

Model	**Lang.**	**Direct H. Ent./Total**	**Direct HR**	TD^2 **H. Ent./Total**	TD^2 **HR**	**ΔHR**
Qwen3-VL-8B $\alpha = 0.05$	hi	1414/4330	32.66%	1399/4330	32.31%	↓ **0.35**
	lv	1772/3983	44.49%	1759/3983	44.16%	↓ **0.33**
	tr	1179/3642	32.37%	1216/3642	33.39%	↑ 1.02
Qwen3-VL-32B $\alpha = 0.20$	hi	1266/4330	29.24%	1247/4330	28.80%	↓ **0.44**
	lv	1507/3983	37.84%	1425/3983	35.78%	↓ **2.06**
	tr	1171/3642	32.15%	1157/3642	31.77%	↓ **0.38**
Qwen2-VL-7B $\alpha = 0.30$	hi	2529/4330	58.41%	2487/4330	57.40%	↓ **1.01**
	lv	3329/3983	83.58%	3050/3983	76.58%	↓ **7.00**
	tr	1923/3642	52.80%	1803/3642	49.51%	↓ **3.29**

Table 10: Complementary Stanza-based entity matching statistics on Multi30K. *H. Ent./Total* denotes hallucinated entity instances over total reference entities. ΔHR is computed as the hallucination rate of Direct decoding minus that of TD^2; downward arrows indicate reduced hallucination rates, while upward arrows indicate increased rates.

Dataset	**Lang.**	$\alpha = 0.20$		$\alpha = 0.25$		$\alpha = 0.30$		$\alpha = 0.35$		$\alpha = 0.40$	
		BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
VG	ha	3.41	44.87	3.66	45.39	3.78	45.81	3.92	46.25	4.12	46.50
	bn	15.33	63.78	15.21	63.81	15.06	63.75	14.94	63.78	14.90	63.64
	ma	4.69	50.35	4.72	50.57	4.64	50.82	4.68	50.88	4.56	50.87
	hi	17.90	58.24	18.10	58.15	18.15	58.16	18.03	57.86	17.63	57.69
Multi30K	hi	12.70	52.69	12.92	52.60	12.83	52.63	12.83	52.45	12.79	52.46
	lv	7.34	47.89	7.50	48.50	7.86	48.87	7.85	49.11	7.76	49.10
	tr	14.06	76.32	14.06	76.30	14.26	76.23	14.32	75.90	14.40	75.77
Average		10.776	56.306	10.881	56.474	**10.940**	**56.610**	10.939	56.604	10.880	56.576

Table 11: Language-level hyperparameter results for Qwen2-VL-7B-Instruct. For each candidate contrast coefficient α, we report BLEU and COMET on every dataset-language pair. The last row reports the average over all seven dataset-language pairs, with the highest average value for each metric highlighted in bold.

Dataset	Lang.	$\alpha = 0.01$		$\alpha = 0.03$		$\alpha = 0.05$		$\alpha = 0.07$		$\alpha = 0.09$	
		BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
VG	ha	4.67	52.41	4.68	52.41	4.81	52.64	4.80	52.76	4.86	52.88
	bn	25.64	78.27	25.63	78.18	25.57	78.05	25.51	78.01	25.65	78.02
	ma	17.03	69.56	16.95	69.61	17.00	69.71	16.97	69.62	17.03	69.63
	hi	30.94	75.18	31.01	75.22	30.96	75.17	30.93	75.07	30.87	75.03
Multi30K	hi	21.97	69.07	21.97	69.04	21.90	69.05	21.86	69.18	21.81	69.22
	lv	16.54	74.47	16.57	74.42	16.66	74.38	16.62	74.19	16.36	73.97
	tr	21.09	85.00	21.22	84.96	21.41	84.87	21.43	85.00	21.58	84.94
Average		19.697	**71.994**	19.719	71.977	**19.759**	71.981	19.731	71.976	19.737	71.956

Table 12: Language-level hyperparameter results for Qwen3-VL-8B-Instruct. For each candidate contrast coefficient α, we report BLEU and COMET on every dataset-language pair. The last row reports the average over all seven dataset-language pairs, with the highest average value for each metric highlighted in bold.

Dataset	Lang.	$\alpha = 0.10$		$\alpha = 0.15$		$\alpha = 0.20$		$\alpha = 0.25$		$\alpha = 0.30$	
		BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
VG	ha	5.78	54.11	6.07	54.70	6.23	55.10	6.32	55.31	6.49	55.50
	bn	30.09	81.94	30.11	81.90	29.79	81.84	29.64	81.62	29.09	81.41
	ma	21.62	74.55	21.57	74.49	21.46	74.38	21.20	74.50	21.13	74.42
	hi	35.68	77.91	35.78	77.91	35.87	77.88	35.96	77.86	35.83	77.74
Multi30K	hi	25.75	71.84	25.82	71.93	25.97	72.00	26.02	71.93	26.07	72.03
	lv	22.33	80.89	22.46	81.02	22.48	81.02	22.47	80.92	22.44	80.78
	tr	24.95	87.58	25.26	87.01	25.44	87.60	25.61	87.61	25.75	87.47
Average		23.743	75.546	23.867	75.566	**23.891**	**75.689**	23.889	75.679	23.829	75.621

Table 13: Language-level hyperparameter results for Qwen3-VL-32B-Instruct. For each candidate contrast coefficient α, we report BLEU and COMET on every dataset-language pair. The last row reports the average over all seven dataset-language pairs, with the highest average value for each metric highlighted in bold.