

Explicit State Gating and Implicit Refusal for Unreadable Scene Text

Dayeon Jang^{1,‡}, Yuseok Bae², Kimin Yun^{2,3†}

¹Sungkyul University, Anyang, South Korea

²Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea

³University of Science and Technology (UST), Daejeon, South Korea

[†]Corresponding author. [‡]Work done during an internship at ETRI.

Emails: clavexn@gmail.com, {baeys, kimin.yun}@etri.re.kr

Abstract

Scene text recognition systems must decide what to do when an image contains no reliably readable text. We compare two abstention contracts for a vision-language scene-text recognizer. Under *implicit refusal*, the model signals unreadability inside its own transcription output. Under *explicit state gating*, a separate five-way state classifier decides whether a transcription is emitted at all. Across four training variants sharing the same backbone, direct OCR yields similar accuracy on readable text; the two contracts differ mainly in how they handle inputs with no readable text. Gating reduces false transcription (emitting text where none is readable) from 45.9% to 1.7%, at the cost of falsely abstaining on 3.8% of readable inputs. Which contract is preferable, however, cannot be determined without first specifying how much a correct abstention is worth. When rejected inputs are excluded from scoring, all gated configurations show similar risk; when a correct abstention is credited as a useful outcome, gating lowers risk on the diagnostic set. We present this preliminary analysis not as a context-free model ranking, but as a framework for choosing an abstention contract at deployment time. These results motivate future systems that use the predicted state to select a recovery action before abstaining.

1 Introduction

Modern vision-language models (VLMs) can transcribe scene text well enough to be benchmarked as general-purpose readers (Liu et al., 2024; Fu et al., 2025), yet they often produce plausible strings when a crop is mirrored, collapsed, missing its target, or too degraded to read (He et al., 2025; Guan et al., 2026). A deployed system therefore needs a contract for saying “I cannot read this.” Selective prediction frames this as a risk-coverage trade-off (Chow, 1970; Geifman and El-Yaniv, 2017), and hallucination work asks whether a model can avoid

unsupported output (Li et al., 2023). This paper asks whether abstention should be represented inside the transcription or by an explicit input-state prediction before transcription.

An *in-band* contract expresses refusal in the transcription channel, for example through an [UNREADABLE] token or a verbal declaration (Guan et al., 2026). This keeps refusal available whenever transcription is called, but uses the same output channel for recognized content and control signals. An *out-of-band* contract instead predicts an input state before transcription. We implement this as a five-way gate (*normal* / *mirrored* / *target absent* / *collapsed-furled* / *degraded-other*) that invokes OCR only for crops classified as *normal*. This separates state prediction from transcription, but readable inputs misclassified by the gate are discarded.

Using matched Qwen3.5-4B (Qwen Team, 2026a) experiments and 546 human-verified crops, we find that (i) readable-text accuracy remains similar across four training variants; (ii) gating reduces false transcription from 45.9% to 1.7%, while rejecting 3.8% of readable inputs; and (iii) scoring conventions can reverse the comparison between direct and gated inference for the same fixed predictions.

2 Related Work

Reject-option classification formalizes the trade-off between prediction error and the cost of declining to answer (Chow, 1970; El-Yaniv and Wiener, 2010), and selective classification evaluates confidence-based rejection through risk-coverage curves (Geifman and El-Yaniv, 2017). Our setting also allows abstention to be represented as a semantic input state. Such a state can specify a possible cause, such as mirroring, which may support state-specific recovery in future work. Empty-reference scene-text crops provide a direct halluci-

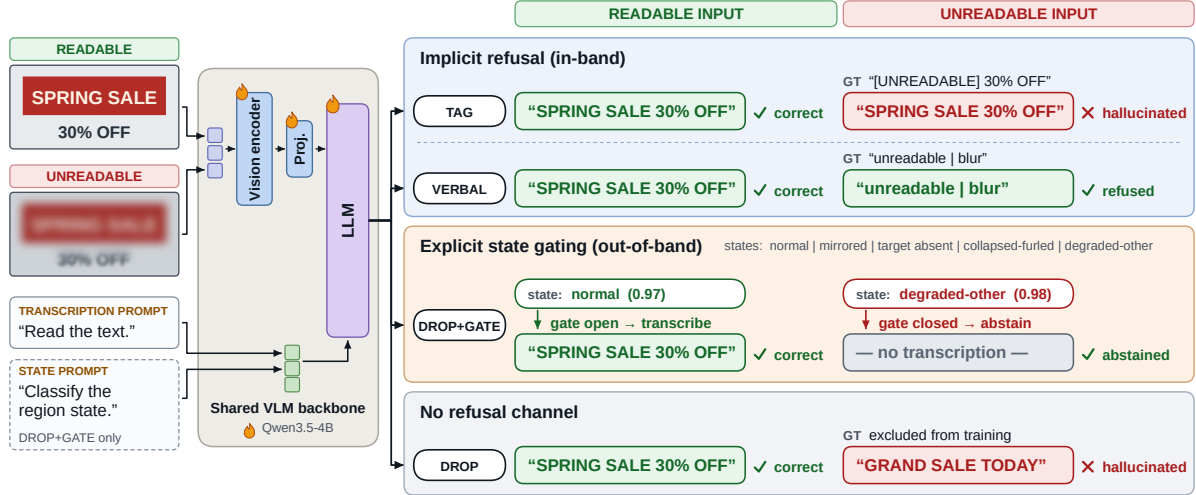


Figure 1: Schematic comparison of four training variants on a shared VLM backbone. TAG and VERBAL represent unreadability in the transcription channel. DROP+GATE predicts a five-way state before transcription, while DROP contains no refusal mechanism. At inference, the same state gate can be placed before any variant’s OCR task.

nation test because any nonempty transcription is unsupported (Li et al., 2023), while readable crops still require character-level metrics such as CER and chrF (Popović, 2015). We compare in-band and out-of-band abstention under shared evaluation conditions and report recognition accuracy, false transcription, coverage, and action-level risk separately.

Abstention in scene text and OCR. OCR-oriented benchmarks for large multimodal models measure how much text a model reads, so unreadable input usually appears as recognition error rather than as a state the system may declare (Liu et al., 2024; Fu et al., 2025). Related work represents uncertainty through in-band tags (Guan et al., 2026), legibility supervision (He et al., 2025), or post-hoc probes on latent representations (Yao et al., 2025). We instead compare in-band and out-of-band signal placement on a shared backbone and source dataset. The measured outcomes of both contracts depend on the scoring assigned to abstention and on the frequency of unreadable inputs. Our study is closest to uncertainty-aware decision making: it compares where an unreadable state is represented, how that representation controls selective prediction, and how the resulting action is valued.

3 Abstention Contracts

3.1 Training variants

We study four training variants that differ in target construction and state supervision. All variants

use the same Qwen3.5-4B backbone (Qwen Team, 2026a) and share the same OCR task format.

TAG uses targets constructed by a separate Qwen3.6-27B model (Qwen Team, 2026b), which assigns an inline [UNREADABLE] marker to unreadable inputs. VERBAL uses targets rewritten by the same Qwen3.6-27B model; these targets separate the transcription from a reason field and express unreadability verbally within the transcription channel. DROP excludes human-verified unreadable crops from OCR training and uses only the original readable-text targets. DROP+GATE uses the same filtered OCR data as DROP and additionally trains the shared backbone on a five-way input-state classification task. Thus, DROP serves as the plain OCR baseline for measuring the effect of co-training the state gate.

At inference, each OCR variant is evaluated under both direct and gated protocols. The gated protocol uses the state classifier trained by DROP+GATE and places it before each variant’s OCR task. We reserve the term *contract* to denote where the refusal signal is exposed to a caller: TAG and VERBAL provide an in-band contract through the transcription channel, whereas the state gate provides an out-of-band contract before transcription. DROP alone provides no refusal mechanism.

Because both TAG and VERBAL rely on targets constructed by Qwen3.6-27B, comparisons among TAG, VERBAL, and DROP change target construction as well as refusal representation. They should therefore be interpreted as comparisons among

complete training variants, not as estimates of the isolated causal effect of an abstention contract. Human review is used to identify unreadable crops for filtering, not to correct their transcriptions.

3.2 Inference protocols

We evaluate each training variant under two inference protocols. Under **Direct-OCR**, the transcription task is invoked for every crop, regardless of its predicted input state. Under **Gated** inference, the five-way state classifier is applied first; transcription is invoked only for crops classified as *normal*, while all other predictions result in a typed abstention. The gate is trained jointly with DROP+GATE, but it can be placed before the OCR head of another variant without additional training. This shared-gate evaluation separates the observed effect of invoking the gate at inference from the target-construction differences among OCR variants. Gate-based rejection affects system actions only when deployment follows the gated inference path.

3.3 Representation, decision, and evaluation

Our comparison separates three choices that are often conflated in uncertainty-aware recognition: representation, decision, and evaluation. In-band contracts represent unreadability in the transcription channel, whereas the state gate represents an observed input condition outside that channel. Neither representation alone determines the system action. Under direct OCR, transcription is always invoked; under gated inference, the predicted state controls whether transcription is attempted.

The gate should not be interpreted as a general estimator of epistemic uncertainty. Its five labels describe operational input states supervised in the training data, and its abstention behavior is evaluated only on the diagnostic distribution considered here. The scoring convention then assigns utility to the resulting action: covered-only risk evaluates retained answers, whereas unified risk credits a correct abstention and penalizes an incorrect one. This decomposition makes explicit that uncertainty representation, selective action, and evaluation are distinct design choices.

3.4 Two scoring conventions

We define *risk* as the average penalty for the system’s final actions: transcription errors, false transcriptions on empty-reference crops, and incorrect abstentions. How abstention is counted depends

on deployment. **Covered-only** (COVERED) evaluates only answered samples, which is appropriate when abstentions are handled downstream. **Unified** (UNIFIED) evaluates all samples, treating a correct abstention as loss 0 and abstention on readable text as loss 1.

Let $y_i = 1$ indicate readable text, $a_i = 1$ an answered sample, and ℓ_i the answering loss: normalized CER in $[0, 1]$ for readable crops and the false-transcription indicator for empty-reference crops. We compute

$$R_{\text{cov}} = \frac{\sum_i a_i \ell_i}{\sum_i a_i}, \quad (1)$$

$$R_{\text{uni}} = \frac{1}{N} \sum_i [a_i \ell_i + (1 - a_i) y_i]. \quad (2)$$

Thus, R_{cov} measures the quality of retained answers, whereas R_{uni} measures end-to-end action quality over all N inputs. Under direct OCR, every input is answered ($a_i = 1$ for all i), so the two risks coincide. Neither convention is universally preferable; the choice between them makes the assumed utility of a correct abstention explicit.

4 Experiments

4.1 Setup

We evaluate 546 Korean banner crops verified by human review. Of these, 365 contain readable text and 181 have empty references because of mirroring, missing targets, collapsed banners, or severe degradation. Empty-reference crops account for 33.2% of the diagnostic set. We report macro CER and chrF on readable crops, false-transcription rate (FT) on empty-reference crops, coverage, and the risks defined in Section 3.4. All variants use the same gate predictions.

4.2 Recognition and gate behavior

Under direct OCR, readable-text performance varies within a narrow range across the four training variants: macro CER ranges from 28.5 to 31.2, and chrF ranges from 66.4 to 68.0 (Table 1). No variant shows a clear advantage on the readable subset of this diagnostic set.

The fixed state gate blocks 177 of 181 empty-reference crops and 14 of 365 readable crops (Table 2). For DROP+GATE, applying this gate reduces the false-transcription rate from 45.9% to 1.7% and coverage from 100% to 65%. Of the four empty-reference crops passed to OCR, three produce nonempty transcriptions. The blocked

Table 1: Integrated results on 546 human-verified crops under two scoring conventions (§3.4). CER and chrF use all 365 readable crops for direct OCR and the 351 readable crops passed by the gate for gated inference. FT uses all 181 empty-reference crops. For direct OCR, covered-only and unified risk coincide because all inputs are answered.

Variant	Stratum-level			Coverage	Risk ↓	
	Macro CER ↓	ChrF ↑	FT ↓		covered-only	unified
<i>Direct OCR (readable $n = 365$; no abstention)</i>						
TAG	29.8	67.7	39.2	100%	32.81	32.81
VERBAL	31.2	66.4	17.1	100%	26.43	26.43
DROP	30.2	67.0	38.1	100%	32.77	32.77
DROP+GATE	28.5	68.0	45.9	100%	34.15	34.15
<i>Gated inference (answered readable $n = 351$; overall coverage 355/546)</i>						
TAG	28.3	69.5	1.7	65%	28.79	21.28
VERBAL	29.7	67.9	0.0	65%	29.37	21.66
DROP	28.0	69.2	1.7	65%	28.50	21.10
DROP+GATE	26.9	69.7	1.7	65%	27.47	20.42

Table 2: Decisions of the fixed state gate on the 546-crop diagnostic set. Blocking readable text is a false abstention; passing an empty-reference crop is a gate miss.

Reference state	Passed to OCR	Blocked
Readable ($n = 365$)	351 (96.2%)	14 (3.8%)
Empty reference ($n = 181$)	4 (2.2%)	177 (97.8%)

readable crops are also more difficult to transcribe: when OCR is applied to these 14 crops, their macro CER is 66.4, compared with 26.93 for the 351 readable crops passed by the gate. Under gated inference, however, all 14 blocked readable crops are counted as false abstentions.

This reduction in false transcription depends on invoking the gate at inference. Under direct OCR, DROP+GATE retains a false-transcription rate of 45.9%, compared with 38.1% for DROP. Co-training the state task alone does not suppress unsupported transcription. By contrast, VERBAL can express an in-band refusal when OCR is invoked without the gate.

4.3 Risk depends on how abstention is scored

The effect of gating depends on how rejected inputs are included in the evaluation. For VERBAL, gated inference increases covered-only risk from 26.43 to 29.37, but decreases unified risk from 26.43 to 21.66. Thus, the same gate appears unfavorable when evaluation is restricted to answered inputs and favorable when correct abstentions are credited.

This difference arises from the 191 inputs

blocked by the gate. Covered-only risk excludes these inputs and measures the quality of the remaining transcriptions. Unified risk instead assigns zero loss to the 177 correctly blocked empty-reference crops and unit loss to the 14 blocked readable crops. The two metrics measure different outcomes: transcription quality among answered inputs and action quality over the full diagnostic set.

The same pattern holds across the four gated variants. Covered-only risk ranges from 27.47 to 29.37, whereas unified risk ranges from 20.42 to 21.66. These results do not provide a single ranking of the training variants; the comparison depends on the value assigned to abstention. The VERBAL result should also be interpreted with caution because its targets were constructed by an external language model.

5 Conclusion

On this diagnostic set and under the evaluated protocols, changing the training variant does not yield a clear readable-text advantage. In contrast, placing the state gate at inference cuts the false-transcription rate by a factor of 27 (45.9% $\rightarrow$ 1.7%) while discarding 3.8% of readable inputs. Whether this trade-off is favorable depends on how abstention is scored and whether deployment consistently invokes the gate. These dependencies argue for reporting coverage, stratum-level errors, and an explicit abstention cost rather than a single deployment-independent ranking. Future work should evaluate a layered system that combines in-band refusal with state-specific recovery before final abstention.

Limitations

All evaluated recognizers share a single Qwen3.5-4B backbone, and the diagnostic set contains 546 crops from a single Korean banner domain. The TAG and VERBAL targets were constructed by larger external models, which introduces an additional source of variation across training variants. The diagnostic set deliberately oversamples empty-reference inputs, so the reported aggregate risks are specific to this diagnostic composition. The gate still permits three false transcriptions and its predicted states have not been tested in a downstream recovery pipeline. The reported risks are therefore conditional on this dataset and scoring design; larger and more diverse models, domains, and prospective deployment data remain to be evaluated.

Ethical Considerations

The gate is intended to reduce unsupported transcription, but incorrect abstention may suppress readable text. Deployment should preserve the input image, provide a recovery path, and monitor false abstention. Any release of scene-text images must also follow the source dataset’s privacy and licensing requirements.

Acknowledgments

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-II220124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities) and Electronics and Telecommunications Research Institute(ETRI) grant funded by the Korean government (26CB1200, Development and Application of Science-Specialized Multimodal Foundation Models)

References

- C. K. Chow. 1970. [On optimum recognition error and reject tradeoff](#). *IEEE Transactions on Information Theory*, 16(1):41–46.
- Ran El-Yaniv and Yair Wiener. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641.
- Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan,

Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, and 5 others. 2025. [OCR-Bench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning](#). *Computing Research Repository*, arXiv:2501.00321.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems* 30.

Shuhao Guan, Moule Lin, Cheng Xu, Jinman Zhao, and Derek Greene. 2026. [Teaching VLMs to admit uncertainty in OCR from lossy visual inputs](#). In *Proceedings of the 14th International Conference on Learning Representations (ICLR)*.

Zhentao He, Can Zhang, Ziheng Wu, Zhenghao Chen, Yufei Zhan, Yifan Li, Zhao Zhang, Xian Wang, and Minghui Qiu. 2025. [Seeing is believing? mitigating OCR hallucinations in multimodal large language models](#). In *Advances in Neural Information Processing Systems* 38.

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*.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024. [OCR-Bench: On the hidden mystery of OCR in large multimodal models](#). *Science China Information Sciences*, 67(12):220102.

Maja Popović. 2015. [chrF: Character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Qwen Team. 2026a. [Qwen3.5: Towards native multimodal agents](#).

Qwen Team. 2026b. [Qwen3.6-27B: Flagship-level coding in a 27B dense model](#).

Jihan Yao, Achin Kulshrestha, Nathalie Rauschmayr, Reed Roberts, Banghua Zhu, Yulia Tsvetkov, and Federico Tombari. 2025. [Reading between the lines: Abstaining from VLM-generated OCR errors via latent representation probes](#). *Computing Research Repository*, arXiv:2511.19806.