Reading or Guessing? Visual Grounding Failures of Vision-Language Models for OCR in Ancient Greek Editions

Anonymous ACL submission

Abstract

Recent work has shown that vision-language models (VLMs) used for optical character recognition (OCR) can generate plausible but visually unsupported text, suggesting reliance on language priors. Comparing open-weight VLMs with traditional OCR baselines on low-resource Ancient Greek critical editions, we show that VLM errors often remain fluent even when wrong, producing plausible Greek substitutions where traditional engines produce local recognition noise. To analyze visual evidence during decoding, we introduce controlled image perturbations and token-level grounding measures based on conditional versus image-free decoding distributions. Under character-level perturbations, VLMs diverge sharply from the perturbed ground truth while traditional OCR remains comparatively faithful; however, token-level analysis shows that prior reliance is model-specific: in an OCR-specialist model, fluent lexical errors are produced with little reliance on the image, whereas general-purpose VLMs remain conditioned on the visual input even when wrong. Decode-time interventions fail to reliably restore grounding, while post-OCR language-model correction improves several systems only by repairing text after generation. Our results extend prior evidence of OCR language-prior reliance to low-resource historical documents and a broader set of models, showing that fluent output is not necessarily visually grounded.

1 Introduction

Vision-language models (VLMs) have displaced pipeline OCR systems as the dominant architecture for document understanding, achieving near-human performance on benchmarks (Bai et al., 2025; Wei et al., 2024; Liu et al., 2024). However, recent work has shown that these systems may rely on language priors: when linguistic context is perturbed, VLM accuracy degrades substantially, often producing semantically plausible but visually unsupported output (Shu et al., 2025; Liang et al., 2026; He et al., 2025; Gong et al., 2026). While prior-driven failures have primarily been studied in high-resource settings, much less is known about their behavior under weak language priors. We study this problem in Ancient Greek critical editions, a low-resource OCR setting with dense layouts, polytonic scripts, marginal annotations, and alphabetic numerals. These documents provide limited support for the lexical and formatting priors learned during VLM pretraining, making them a useful model-analysis probe of how visual evidence and language priors interact during generation.

We compare five open-weight VLMs (Bai et al., 2025; Taghadouini et al., 2026; Poznanski et al., 2025; Wei et al., 2026) with two traditional OCR baselines (Smith, 2007; Kiessling, 2025) on 90 scans from Ancient Greek critical editions (Anglerauad et al., 2026). We first analyze error distributions, then use controlled counterfactual perturbations that corrupt the on-page text before rendering, and finally measure token-level grounding with conditional versus image-free decoding distributions and evaluate decoding interventions designed to constrain prior-driven generation.

Our results show that VLMs diverge from the perturbed ground truth much more than traditional OCR systems, but that this prior reliance is model-specific: in the OCR-specialist olmOCR, fluent lexical errors are produced with little reliance on the image, while general-purpose VLMs remain conditioned on the visual input even when wrong. Perceptual errors stay visually grounded, and script-mixing errors are the most image-driven. Decode-time grounding constraints fail to reliably restore grounding. Together, these findings suggest that fluent OCR outputs can hide weak visual grounding, raising challenges for the interpretability and evaluation of multimodal language models in low-resource settings.

2 Related Work

Vision-language models have advanced OCR and document understanding, achieving strong performance on recent benchmarks (Bai et al., 2025; Wei et al., 2024; Liu et al., 2024; Poznanski et al., 2025; Wei et al., 2026; Taghadouini et al., 2026), but performance remains weaker for Unicode-heavy and low-resource scripts (Angleraud et al., 2026; Kargaran et al., 2026) such as Ancient Greek. Prior work has developed large-scale pipelines and specialized systems for polytonic Greek and mixed classical editions (Robertson and Boschetti, 2017; Romanello et al., 2021; Sichani et al., 2019; Kaddas et al., 2023; Perifanos and Goutsos, 2025), focusing mainly on recognition accuracy. Recent studies show that VLM-based OCR can rely on linguistic or semantic priors, producing plausible but visually unsupported text when lexical context or visual structure is disrupted (Shu et al., 2025; Liang et al., 2026; He et al., 2025; Gong et al., 2026). These findings raise several questions: do such prior-driven failures distinguish VLMs from traditional OCR systems, do they exist in low-resource historical settings, and can they be mitigated at inference time? Existing comparisons between widely used VLMs and off-the-shelf traditional OCR systems have so far mainly focused on videos (Fei et al., 2025; Shi et al., 2026), while recent historical-document OCR work has compared OCR-native transformers with VLMs only in English documents (Vesalainen et al., 2026). We study these questions comparing VLMs with traditional OCR baselines on Ancient Greek critical editions through error analyses, perturbation, image-gain attribution experiments, and decoding interventions.

3 RQ1: Do VLMs and traditional OCR systems fail differently?

We first ask whether VLM-based OCR systems make the same kinds of errors as traditional OCR engines. We evaluate five open-weight VLMs and two traditional OCR baselines on 90 scans from Ancient Greek critical editions (Angleraud et al., 2026) covering 30 authors, using Character Error Rate (CER) and Word Error Rate (WER). Model and inference details are provided in Appendix A. Table 1 shows that aggregate accuracy alone does not cleanly separate VLMs from traditional OCR systems: the best traditional engine (Kraken-CLLG) and the strongest VLM (Qwen3-VL-8B) are close in median CER. However, the relation between character- and word-level errors differs systematically: VLMs have a consistently higher median WER/CER ratio than the traditional baselines, suggesting that their errors often involve small character-level deviations that make many words incorrect. This motivates a closer inspection of the failure modes.

	CER (%)		WER (%)	
Model	Mean	Median	Mean	Median
Tesseract-grc	10.0	7.5	29.0	23.7
Kraken-CLLG	**5.2**	**4.1**	**12.5**	**11.7**
LightOn-1B	11.5	**4.1**	19.4	17.8
DeepSeek-OCR	7.8	6.7	30.6	26.2
Qwen3-VL-2B	10.3	8.0	41.2	34.0
OlmOCR-2-7B	9.2	6.4	32.1	26.3
Qwen3-VL-8B	5.6	5.1	22.9	22.3

Table 1: Character (CER) and Word (WER) Error Rates on the real-scan Ancient Greek critical-edition benchmark (n=90 pages, 30 editions). VLMs use greedy decoding (T=0); traditional OCR baselines are deterministic. Best per column in **bold**.

To define the categories, we manually inspected 20 pages spanning all systems to identify frequent error types, then classified each mismatch automatically. Each prediction is aligned to its ground truth with a word-level edit-distance backtrace, and every mismatch is mapped to one of eight mutually exclusive categories (Appendix B): *accent/diacritic*, *character confusion*, *cross-script* substitution, *word substitution*, *overgeneration*, *omission*, *page furniture*, and *punctuation*. The real-word/non-word split of substitutions is resolved against a 19,901-form lexicon.

The dominant failure mode is system-specific (Figure 3 in Appendix B). For the VLMs, most errors are perceptual mistakes, especially accent/diacritic errors and character confusions. The traditional systems show a different profile: Kraken is dominated by page-furniture errors, while Tesseract has a more mixed distribution across punctuation, page furniture, overgeneration, and local recognition errors. The key difference is qualitative: VLM substitutions tend to be fluent, in-vocabulary Greek words, whereas traditional OCR substitutions typically reflect local recognition noise.

Takeaway. VLM errors are dominated by perceptual surface-orthography mistakes, with outputs that often remain fluent rather than noisy. This helps explain their higher WER/CER ratio: small character-level deviations can turn many otherwise

System	Word-level clean	Word-level scr.	Word-level Δ	Char-level clean	Char-level scr.	Char-level Δ
Tesseract-grc	.02	.02	−.00	.02	.08	+.06***
Kraken-CLLG	.01	.01	+.00	.01	.04	+.03***
LightOnOCR-1B	.04	.07	+.03	.04	.16	**+.12*** **
DeepSeek-OCR	.02	.04	+.02	.02	.39	**+.38*** **
Qwen3-VL-2B	.13	.18	+.05	.14	.82	**+.68*** **
OlmOCR-2-7B	.05	.06	+.01	.05	.33	**+.28*** **
Qwen3-VL-8B	.05	.06	+.01	.05	.26	**+.21*** **

Table 2: CER (no spaces/diacritics) on clean vs. scrambled (most-disrupted `random`) renderings, scored against the scrambled GT. Stars: per-image paired one-sided Wilcoxon (H_1: scrambled > clean); $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

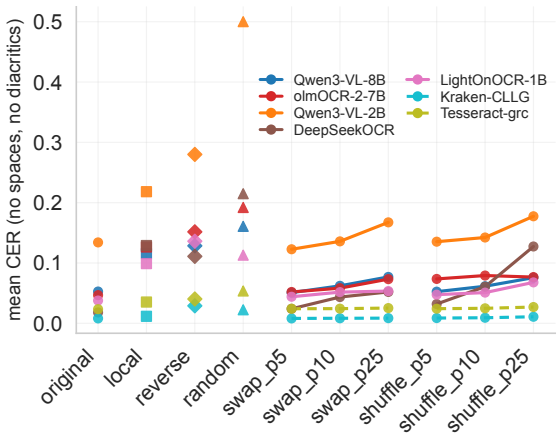

Figure 1: Perturbations averaged across test1 and test2, showing mean CER without diacritics by condition; only proportional swap/shuffle series are connected.

plausible Greek words into word-level errors. By contrast, the traditional engines evaluated here lack an open-ended autoregressive decoder, and their errors stay more tightly tied to visible glyphs, page layout, or segmentation. Whether fluent VLM substitutions reflect a language prior rather than visual confusion is the question we isolate in RQ2.

4 RQ2: Is there reliance on language priors?

To test whether the recognition failures reflect language-prior correction rather than visual confusion, we use two approaches: a behavioral counterfactual perturbation experiment and a token-level image-gain analysis. In the perturbation experiment, we modify the synthetic Ancient Greek corpus by Angleraud et al. (2026), so each image faithfully displays the perturbed string and the perturbed string itself is the ground truth. A visually grounded system should reproduce these glyphs even when they form implausible Greek; a system relying on lexical or orthographic priors should instead "repair" them toward more plausible forms, increasing error against the perturbed ground truth.

Across perturbation levels, VLM error rises sharply while the traditional baselines remain comparatively stable (Figure 1). This behavioral divergence is consistent with prior-driven repair, but could also reflect greater sensitivity to out-of-distribution glyph sequences. We therefore use the image-gain probe to separate these explanations.

We analyze token-level image reliance by comparing each token's image-conditioned decoding distribution with an image-free one (Appendix C.2, Table 7). The result is *model-specific*. The low-image-gain signature of prior-driven errors holds cleanly only for the OCR-specialist olmOCR: its fluent lexical substitutions remain probable even when the image is removed (median image gain 0.17, far below its perceptual confusions at 1.40). The Qwen-3VL models instead stay image-engaged even when wrong, with lexical gain at or above perceptual gain (1.97 vs. 1.30 for Qwen3-VL-8B; 2.04 vs. 1.88 for Qwen3-VL-2B). Across the three models for which image-gain is available, local perceptual confusions retain substantial image dependence, and script-mixing (Greek→Latin) substitutions are the *most* image-driven of all: visually triggered rather than image-ignoring.

Takeaway. Under controlled perturbation, traditional OCR remains comparatively faithful to the degraded glyphs, while every VLM diverges more strongly from the perturbed ground truth. But this is not a single prior-override mechanism: image gain shows clear prior override only for the OCR-specialist olmOCR, whereas the general-purpose VLMs stay image-engaged even when they substitute whole words.

5 RQ3: Can inference-time interventions restore grounding?

Finally, we ask whether certain failures can be mitigated at inference time. We evaluate five interventions spanning three families: (i) *Unicode-range script constraints*—token-level logit mask for Greek + Greek extended characters with length-based abstention; (ii) *contrastive, grounding-aware decoding*—Visual Contrastive Decoding (VCD) (Leng et al., 2024) and image-free con-

Model	script-mask	len-abstain	VCD	M3ID	LM-correct
	Δmed Δmean	Δmed Δmean	Δmed Δmean	Δmed Δmean	Δmed Δmean
Tesseract-grc	–	–	–	–	-0.90 -1.51^{***}
Kraken-CLLG	–	–	–	–	-0.34 -0.38^{***}
LightOnOCR-1B	$+64.5$ $+54.1^{***}$	$+0.32$ $+4.30$	–	–	-0.33 -2.37^{ns}
DeepSeek-OCR	$+12.0$ $+17.8^{***}$	$+0.42$ $+10.3$	–	–	-0.30 -0.68^{***}
Qwen3-VL-2B	$+74.7$ $+77.5^{***}$	$+0.52$ $+9.95$	$+0.25$ -4.47^{ns}	-1.06 -3.70^{***}	-0.28 -1.41^{***}
olmOCR-2-7B	$+57.5$ $+52.7^{***}$	$+0.29$ $+8.98$	$+1.31$ $+1.28^{\text{ns}}$	$+2.30$ $+0.28^{\text{ns}}$	$+0.13$ -1.21^{*}
Qwen3-VL-8B	$+62.1$ $+65.4^{***}$	$+0.32$ $+10.5$	$+0.13$ -0.88^{ns}	-1.73 -5.76^{***}	-0.02 $+1.00^{\text{ns}}$

Table 3: Change in CER under RQ3 interventions relative to each intervention's matched baseline. Negative values indicate improvement. We report changes in median and mean CER (Δmed, Δmean). Stars: per-page paired one-sided Wilcoxon(H_1: helping direction for each intervention); $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$, ns = not significant.

trastive decoding (M3ID) (Favero et al., 2024); and (iii) *post-hoc text correction*—few-shot post-OCR language-model correction. Method and hyperparameter details are given in Appendix D. Each family targets a different failure. Script-restricted decoding masks tokens outside the expected Greek-script and punctuation set, testing whether cross-script and markup errors are removed simply by restricting the output space. Length abstention suppresses outputs whose length departs from the visible text, targeting overgeneration. Contrastive decoding instead reweights tokens by their dependence on the image, penalizing continuations that remain likely without it—VCD contrasts the clean image against a noised one, M3ID against an image-free pass. Post-OCR correction instead targets the dominant surface-form errors observed in RQ1, accent and diacritic errors, by rewriting the recognized text after generation.

Table 3 shows that output-space constraints do not restore visual grounding. Script constraints are *catastrophic*: median CER rises roughly seven- to ten-fold (e.g., Qwen3-VL-8B $0.07 \rightarrow 0.69$), and only 1 of 90 pages improves. Length abstention at best *matches* the baseline on the pages it keeps ($\sim$12% abstention), trading coverage for no gain. Contrastive decoding is mixed: VCD leaves CER essentially unchanged (all p ns), while M3ID significantly improves both Qwen-VL models (Qwen3-VL-2B $\Delta\text{med} = -1.06$, Qwen3-VL-8B $\Delta\text{med} = -1.73$; both $p<.001$) but not the OCR-specialist olmOCR. Thus, contrastive decoding does not provide a model-agnostic grounding fix.[1] Post-OCR correction gives the broadest gains, with statistically significant CER improvements for five out of seven systems.

Takeaway. Prior-driven OCR failures are not solved by simple output constraints or by model-agnostic contrastive decoding. Script masking and length abstention fail to restore grounding, while contrastive decoding helps only some model families. Post-hoc text correction gives the broadest gains, but only by repairing outputs after generation rather than preventing the loss of grounding. This points to a mismatch between visual evidence and generative decoding, rather than a purely perceptual recognition problem.

6 Conclusion

We studied visual grounding failures in VLM-based OCR using Ancient Greek critical editions as a low-resource, structurally dense testbed. Across model families, perturbations, and interventions, VLMs fail differently from traditional OCR systems: their outputs can remain fluent even when visual grounding weakens. Token-level analyses show that this prior reliance is model-specific: olmOCR exhibits fluent lexical errors with little image dependence. This is consistent with concurrent evidence that end-to-end OCR-specialist VLMs systematically rely on linguistic priors over visual evidence (Liang et al., 2026), even in a low-resource language where such priors should be weaker than for English. General-purpose VLMs remain image-conditioned even when wrong. Inference-time interventions do not provide a general fix: output-space constraints fail, contrastive decoding helps only some models, and post-hoc text correction repairs outputs rather than preventing grounding failures. These findings motivate OCR evaluations that go beyond aggregate accuracy and test whether generated text is visually grounded, especially for low-resource and historically underrepresented scripts.

[1] The single-token repetition first observed under M3ID is an artifact of its custom decoding loop rather than of the image-free contrast; a no-contrast pass through the same loop collapses identically.

Limitations

Our benchmark targets Ancient Greek prose, a polytonic, multi-script setting that serves as a case study in low-resource scholarly OCR. This focus is a deliberate depth-over-breadth choice: adding more scripts would make it harder to carry out the controlled perturbation, error-taxonomy, token-level grounding, and intervention analyses that are central to this work. It also lets us compare VLMs with strong domain-specialized traditional OCR baselines. To our knowledge, comparable resources are not yet widely available for many other historical scripts: the analysis requires real-scan pages, reliable transcriptions, synthetic perturbation support, and strong script-specific OCR baselines. Future work should extend this analysis to other languages where such resources exist.

The RQ2 perturbation stimuli are rendered rather than scanned, which gives exact ground truth for the perturbed text and isolates the effect of language-prior repair. We vary fonts and backgrounds to increase visual diversity, while leaving fully realistic scan noise to future work. Our error taxonomy is automatically applied using fixed rules derived from manual inspection and checked against hand-labeled examples by the authors. Post-OCR correction is evaluated with one text-only corrector (Qwen3-VL-8B) and a fixed three-shot prompt, so its gains should be read as evidence that some errors are text-repairable, not as the most optimized post-correction pipeline. Finally, as with most open-weight VLM evaluations, pretraining exposure to related material cannot be exhaustively verified; our analysis therefore focuses on relative failure modes under controlled perturbations rather than claims about memorization.

Ethical Considerations

This work uses scholarly editions of Ancient Greek texts and their transcriptions for OCR evaluation; it raises no human-subjects concerns. The CLLG resources (Angleraud et al., 2026) are described as derived from openly licensed TEI/XML corpora; we cite the dataset paper and follow its stated research/evaluation use. Our findings caution against deploying VLMs on under-resourced historical and cultural-heritage material without human verification: the failures we document are *fluent*, producing plausible but fabricated text that could silently corrupt digitized corpora and downstream scholarship. We release stimuli, code, and per-error classifications to support auditing and reproduction; the released material derives from openly licensed editions. AI assistants were used for language editing and drafting support. All scientific claims, experiments, analyses, and final text were reviewed and verified by the authors. The total compute budget for the experiments reported in this paper is approximately 100 GPU-hours, measured with SLURM `sacct`, comprising $\sim$21 hrs for RQ1 (greedy + 5-seed stochastic sampling across the five VLMs), $\sim$23 hrs for RQ2 (seven perturbation conditions across seven systems), and $\sim$58 hrs for RQ3 (VCD, M3ID, post-OCR LM correction, script-restricted decoding, and length abstention). Experiments ran on a SLURM cluster with mixed NVIDIA accelerators (H100, H200, A100, RTX 8000, RTX 6000, RTX 2080 Ti, V100); main VLM inference used vLLM primarily on H100s, while VCD and M3ID used HuggingFace Transformers primarily on A100s for per-step logit access. Tesseract-grc and Kraken-CLLG ran CPU-side.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, and 1 others. 2024. Pixtral 12B. *arXiv preprint arXiv:2410.07073*.

Nicolas Angleraud, Antonia Karamolegkou, Benoît Sagot, and Thibault Clérice. 2026. Structure-aware text recognition for ancient greek critical editions. *Preprint*, arXiv:2603.02803.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, and 1 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Alix Chagué and Hugo Scheithauer. 2023. Gallicalbum.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14303–14312.

Yulin Fei, Yuhui Gao, Xingyuan Xian, Xiaojin Zhang, Tao Wu, and Wei Chen. 2025. Do current video LLMs have strong OCR abilities? a preliminary study. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9860–9876, Abu Dhabi, UAE. Association for Computational Linguistics.

Weile Gong, Xin He, Yiping Zuo, Weibei Fan, Shi Jin, Zijian Lu, and Lianyong Qi. 2026. From plausibility to verifiability: Risk-controlled generative OCR for vision-language models. *arXiv preprint arXiv:2603.19790*.

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. *arXiv preprint arXiv:2506.20168*.

Panagiotis Kaddas, Basilis Gatos, Konstantinos Palaiologos, Katerina Christopoulou, and Konstantinos Kritsis. 2023. Text line detection and recognition of greek polytonic documents. In *Machine Learning*, Communications in Computer and Information Science. Springer.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.

Amir Hossein Kargaran and 1 others. 2026. GlotOCR: A benchmark for multilingual OCR across 158 scripts. *arXiv preprint arXiv:2604.12978*. Verify author list and exact title from arXiv before submission.

Benjamin Kiessling. 2025. Version 5 of the Kraken ATR engine for the humanities. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 443–458. Springer.

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 (CVPR)*, pages 13872–13882.

Yunhao Liang, Ruixuan Ying, Bo Li, Hong Li, Kai Yan, Qingwen Li, Min Yang, Okamoto Satoshi, Zhe Cui, and Shiwen Ni. 2026. Visual merit or linguistic crutch? a close look at DeepSeek-OCR. *arXiv preprint arXiv:2601.03714*.

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

Konstantinos Perifanos and Dionysis Goutsos. 2025. Logios: An open source greek polytonic optical character recognition system. *Preprint*, arXiv:2506.21474.

Jake Poznanski, Aman Rangapur, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Aman Rangapur, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. 2025. olmocr: Unlocking trillions of tokens in pdfs with vision language models. *Preprint*, arXiv:2502.18443.

Bruce Robertson and Federico Boschetti. 2017. Large-scale optical character recognition of ancient greek. *Mouseion*, 14(3):341–359.

Matteo Romanello, Sven Najem-Meyer, and Bruce Robertson. 2021. Optical character recognition of 19th century classical commentaries: The current state of affairs. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, HIP '21, pages 1–6. Association for Computing Machinery.

Yang Shi, Huanqian Wang, Wulin Xie, Huanyao Zhang, Lijie Zhao, YiFan Zhang, Xinfeng Li, Chaoyou Fu, Zhuoer Wen, Wenting Liu, Zhuoran Zhang, Xinlong Chen, Bohan Zeng, Sihan Yang, Yushuo Guan, Zhang Zhang, Liang Wang, Haoxuan Li, Zhouchen Lin, and 4 others. 2026. MME-videoOCR: Evaluating OCR-based capabilities of multimodal LLMs in video scenarios. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yan Shu, Hangui Lin, Yexin Liu, Yan Zhang, Gangyan Zeng, Yan Li, Yu Zhou, Ser-Nam Lim, Harry Yang, and Nicu Sebe. 2025. When semantics mislead vision: Mitigating large multimodal models hallucinations in scene text spotting and understanding. *arXiv preprint arXiv:2506.05551*.

Anna-Maria Sichani, Panagiotis Kaddas, and Basilis Gatos. 2019. Ocr for greek polytonic (multi-accent) historical printed documents. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2019, pages 91–96. Association for Computing Machinery.

R. Smith. 2007. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.

Said Taghadouini, Adrien Cavaillès, and Baptiste Aubertin. 2026. Lightonocr: A 1b end-to-end multilingual vision-language model for state-of-the-art ocr. *Preprint*, arXiv:2601.14251.

Ari Vesalainen, Eetu Mäkelä, Laura Ruotsalainen, and Mikko Tolonen. 2026. Error patterns in historical ocr: A comparative analysis of trocr and a vision-language model. *Preprint*, arXiv:2602.14524.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, and 1 others. 2024. General OCR theory: Towards OCR-2.0 via a unified end-to-end model. In *arXiv preprint arXiv:2409.01704*. Also referred to as GOT-OCR. Verify final venue.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2026. Deepseek-ocr 2: Visual causal flow. *arXiv preprint arXiv:2601.20552*.

A Models and Inference

A.1 System selection.

We compare systems representing two OCR paradigms: traditional recognition-based engines and end-to-end vision-language OCR models.

Traditional OCR systems. The traditional group consists of TESSERACT v5 (Smith, 2007) with its released Ancient Greek model (denoted TESSERACT-GRC in our tables) and KRAKEN-CLLG, a KRAKEN v5 (Kiessling, 2025) model trained on the CLLG corpus of Ancient Greek critical editions (Angleraud et al., 2026). Both systems perform image-conditioned sequence recognition rather than open-ended language generation: their recurrent recognizers can encode contextual dependencies within a line, but they do not include an autoregressive language-model decoder capable of freely continuing text from a prompt. The VLMs, by contrast, transcribe entire page images end-to-end autoregressively, with no explicit segmentation step. TESSERACT is a widely used open-source OCR engine. Since version 4 its default recognizer has been an LSTM that reads a whole line as a sequence, replacing the legacy character-classifier engine; we run the current v5 release, with layout analysis and line segmentation also performed by Tesseract. The released `grc` model is trained on synthetic rendered text in polytonic Greek fonts (the standard `tessdata` pipeline) rather than on scanned critical editions, so it is not in-domain for our test set. KRAKEN is widely used for historical and manuscript OCR in digital-humanities workflows. Kraken v5 uses a hybrid CNN–LSTM recognizer by default. We pair it with the in-domain recognizer released by Angleraud et al. (2026), trained on the CLLG corpus of Ancient Greek critical editions, so Kraken-CLLG is in-domain for our test set. For the real-scan benchmark, line polygons are taken from the gold eScriptorium ALTO export accompanying the test set, so the reported scores primarily reflect recognition quality rather than segmentation errors. For the RQ2 perturbation stimuli no gold layout is available, so we run Kraken-CLLG end-to-end: line baselines are detected automatically with Kraken's BLLA segmenter (CLLG-trained), each detected line is then recognized with the same CLLG recognizer (with bidirectional reordering enabled), and per-line predictions are concatenated. RQ2 Kraken-CLLG numbers therefore reflect both segmentation and recognition.

Vision-language models. For VLMs, we evaluate five open-weight models that fall into two categories. The first, *OCR-specialist VLMs*, comprises models post-trained on document-transcription data: LIGHTONOCR-1B (Taghadouini et al., 2026), a compact multilingual end-to-end OCR model built on the Pixtral architecture (Agrawal et al., 2024); DEEPSEEK-OCR (Wei et al., 2026), a 3B-parameter mixture-of-experts model designed around aggressive vision-token compression for efficient document inference; and OLMOCR-2-7B (Poznanski et al., 2025), a 7B OCR-specialist post-trained from a Qwen2.5-VL initialization. The second category consists of *general-purpose VLMs*, QWEN3-VL-2B and QWEN3-VL-8B (Bai et al., 2025), two scales of the Qwen3-VL family, which are not OCR-specialized but are widely deployed as document-recognition systems in practice. The lineage relationship between OlmOCR-2 and Qwen2.5-VL lets us isolate the effect of OCR-specific post-training while holding the underlying architecture nearly fixed. We choose these models because they are open-weight, recent, and frequently included in OCR and document-understanding evaluations such as OCR-Bench, OlmOCR-Bench, and recent multilingual or historical-script OCR studies (Liu et al., 2024; Angleraud et al., 2026; Kargaran et al., 2026). This selection lets us ask whether visually ungrounded behaviour is specific to large general-purpose VLMs, or whether it also appears in OCR-specialist VLMs that are increasingly used as practical document-recognition systems.

Decoding Setup.

Decoding setup. All main-text experiments use deterministic greedy decoding via vLLM (`temperature`=0, no repetition or frequency penalty, `max_tokens`=2048, image budget `max_pixels`=1,003,520 $\approx$ 1 MP, pages above that budget resized with LANCZOS). This setting reduces generation variance and ensures reproducible error analysis across models and perturbation conditions. To evaluate decoding sensitivity, we additionally compare greedy against stochastic sampling for all five VLMs, using the Qwen3-VL model-card defaults (`temperature`=0.7, `top_p`=0.8, `top_k`=20) over five independent seeds, with all other settings held fixed. Sampling does not systematically improve CER or WER: in the median-CER view (Table 4), greedy lies within the seed-to-seed spread for every

model; in the mean view, sampling additionally inflates the tail for some models, most notably DeepSeek-OCR (sampling mean CER $12.9 \pm 6.5\%$ vs. greedy 7.8%). Figure 2 visualizes per-seed CER.

Per-model prompts follow each model card. QWEN3-VL-2B and QWEN3-VL-8B use "*You are an OCR system for ancient Greek and Latin printed scholarly texts. Transcribe the page exactly as it appears without the running header and page numbering. Keep all characters, accents, ligatures, punctuation, and spacing.*". OLMOCR-2-7B uses the `olmocr` toolkit's `build_no_anchoring_v4_yaml_prompt`; the YAML preamble is stripped from the prediction. DEEPSEEK-OCR uses the card-default "`Free OCR.`" prompt and LIGHTONOCR-1B runs image-only with no text prompt. TESSERACT-GRC runs in v5 with `--psm 6 -l grc`; KRAKEN-CLLG uses gold ALTO line polygons on the real benchmark and the CLLG-trained BLLA segmenter with bidirectional reordering on the RQ2 stimuli.

A.2 Analysis coverage.

Not every system enters every analysis, and Table 5 summarizes the coverage. The aggregate error metrics (CER/WER and the word-level taxonomy) and the perturbation probe cover all seven systems: they require only a predicted transcription and a page image. The token-level image-gain analysis (Appendix C.2) and the contrastive decoding interventions (VCD, M3ID) instead require a HuggingFace teacher-forcing path that exposes well-defined per-step image-conditioned logits, available only for the three Qwen-VL–architecture models (Qwen3-VL-2B, Qwen3-VL-8B, and OlmOCR-2-7B); LightOnOCR's inference path did not provide a valid image-conditioned versus image-free logit comparison for this analysis, DeepSeek-OCR's architecture is unsupported by this analysis, and the traditional baselines have no autoregressive decoder logits to query. The decode-time output-space constraints (script-restricted decoding, length abstention) apply to the five VLMs but not to the traditional baselines, which produce a single deterministic transcription per page with no per-step logits to constrain. Post-OCR LM correction, by contrast, is text-only — it rewrites an existing transcription without access to the page image — and so applies to all seven systems, including both traditional baselines. A practical consequence is that the token-level grounding signature of RQ2 is established on a single OCR-specialist model, OlmOCR-2-7B.

B Error Taxonomy

For RQ1, we define a word-level error taxonomy and use it to classify each mismatch between prediction and ground truth. The pipeline combines word-level alignment, fixed normalisation, rule-based category assignment, and a 19,901-form lexicon for distinguishing real-word from non-word substitutions. It is fully automated and reproducible from the released analysis notebook.

Alignment. For each page we align the prediction to its ground truth at the word level with a Levenshtein edit-distance backtrace, yielding a sequence of operations: *match*, *substitution* (one GT word read as a different word), *deletion* (a GT word with no counterpart), and *insertion* (a predicted word with no counterpart). Every non-matching operation is assigned to exactly one category, so the categories are mutually exclusive and the per-system counts partition that system's errors.

Normalisation. Before alignment we apply a fixed normalisation to both strings: NFC Unicode normalisation; removal of GT-only annotation markup (⟨ref⟩, ⟨note⟩, section markers) that is not visible on the page; rejoining of line-break and stem hyphenation; unification of elision apostrophes; isolation of editorial brackets and dashes as separate tokens; and splitting of digit–letter junctions. This prevents annotation conventions and typography from being scored as model errors.

Categories. We use ten fine categories folded into eight headline categories (Table 6), grouped by the edit operation that produces them:

- **Substitutions:** *accent/diacritic* (same letters; differs only in accent, breathing, case, or final sigma); *character confusion* (same-script misread of one–two letters, ≤ 2 edits on the bare-letter form); *cross-script* (Greek letters replaced by Latin look-alikes, or vice versa); *word substitution* (a different whole word, > 2 edits, same script); and *punctuation* (difference confined to punctuation, quotation, or elision marks).

- **Deletion:** *omission* (a GT content word produced by no prediction).

Model	Variant	CER (%)		WER (%)	
		Greedy	Sampling	Greedy	Sampling
LightOnOCR-1B	raw	4.1	4.5 ±0.3	17.8	18.9 ±0.3
	no-diac.	2.8	3.1 ±0.3	11.4	12.5 ±0.4
DeepSeek-OCR	raw	6.7	7.1 ±0.2	26.2	27.0 ±1.0
	no-diac.	4.4	4.6 ±0.1	15.7	16.6 ±0.7
Qwen3-VL-2B	raw	8.0	8.4 ±0.2	34.0	35.4 ±0.9
	no-diac.	4.0	4.1 ±0.1	17.6	18.0 ±0.7
OlmOCR-2-7B	raw	6.4	6.6 ±0.0	26.3	26.7 ±0.2
	no-diac.	3.4	3.6 ±0.1	16.2	16.5 ±0.5
Qwen3-VL-8B	raw	5.1	5.2 ±0.1	22.3	22.2 ±0.3
	no-diac.	2.9	3.1 ±0.1	13.9	14.2 ±0.3

Table 4: Decoding strategy for the five VLMs on the real-scan benchmark: greedy (T=0) vs. stochastic sampling (5 seeds). Values are median per-page error rates; sampling reports mean±std across the five seed-level medians. We report raw text and a diacritic-insensitive variant (no-diac.) for both CER and WER.

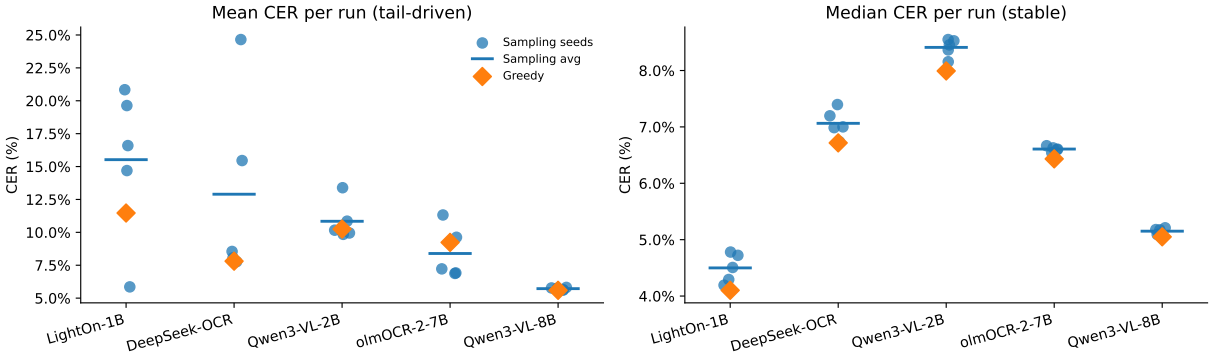

Figure 2: Per-seed CER for the five VLMs (dots = the five sampling seeds, bar = sampling mean, diamond = greedy). Greedy sits within the seed cluster for every model.

System	RQ1		RQ2		RQ3		
	C/W	Tax	Prt	Gain	Dec	LMc	Con
Tesseract-grc	•	•	•	–	–	•	–
Kraken-CLLG	•	•	•	–	–	•	–
LightOnOCR-1B	•	•	•	–	•	•	–
DeepSeek-OCR	•	•	•	–	•	•	–
Qwen3-VL-2B	•	•	•	•	•	•	•
olmOCR-2-7B	•	•	•	•	•	•	•
Qwen3-VL-8B	•	•	•	•	•	•	•

Table 5: Which system enters which analysis (• included, – not applicable). **C/W**: CER/WER; **Tax**: error taxonomy; **Prt**: perturbation; **Gain**: token-level image gain; **Dec**: script masking and length abstention; **LMc**: post-OCR LM correction; **Con**: contrastive decoding. Image gain and contrastive decoding require per-step image-conditioned logits, available only for the three Qwen-VL–architecture models.

- **Insertions:** *page furniture & ref.* (running heads, page / line / section numbers, and bibliographic Latin that the transcription omits) and *overgeneration* (other spurious tokens, including decoding-loop repetition).

Word substitution merges three fine cases—real-word (in-lexicon) substitution, non-word (out-of-lexicon) substitution, and segmentation (word splitting/merging). A run of ≥ 5 identical predicted tokens is routed to *overgeneration* as decoding collapse. Inserted or deleted numerals, Latin words, and all-caps Greek are routed to *page furniture* regardless of operation.

Real-word lexicon. To split large within-script substitutions into fluent real-word swaps versus non-word strings, we use a lexicon of 19,901 base word-forms (accents, breathings, case, and final sigma stripped) collected from the ground-truth transcriptions of the full train/dev/test corpus by Angleraud et al. (2026). A substitution is labelled *real-word* if its bare-letter form is attested in this lexicon and *non-word* otherwise. The lexicon is shipped with the release so the split is reproducible without the corpus.

Per system patterns. The dominant failure mode is system-specific (Figure 3). Perceptual errors—accent/diacritic and character-confusion

Category	Definition	Examples
Accent & diacritic	Same letters; differs only in accent, breathing, case, or final sigma (surface orthography).	αὐτοῖς→αὐτοῖς; ἐπειδὴ→ἐπειδή; Ἄρ'→Ἄρ'
Character confusion	Same-script visual misread of one–two letters ($\leq$ 2 edits on the bare-letter form).	καὶ→χαὶ; ὅπως→δπως; προσβολὴν→προσδολὴν
Cross-script	A Greek word rendered with Latin look-alike letters (or vice versa).	Παῦλος→ΠάULO; ὅρκου→ϑρsche
Word substitution	A different whole word replaces the target ($>$ 2 edits, same script). In VLMs this is typically a fluent, in-vocabulary real word; in LM-free systems a non-word string.	μισϑός→ἐκούσιος; εἴκοσιν→ἐκείνων; *(non-word)* φησί→σπαο; κόλασις→οττς
Overgeneration	Spurious extra tokens not on the page, including repetition / decoding-loop collapse.	+γὰρ; +αὐτοῦ; loop: ὅ ὅ ὅ…; X X X…
Omission	A ground-truth word produced by no prediction (skipped or truncated text).	−ζητεῖτε; −ἡνεωγμένον; −ἀπῆλϑεν
Page furniture & ref.	Running heads, page / line / section numbers, and bibliographic Latin that the GT transcription omits. Mostly *faithful reads* of page elements excluded by the transcription standard, not transcription mistakes.	+ΕΠΙΣΤΟΛΑΙ; +[ΑΙΣΧΙΝΟΥ]; +141; +PORPHYR
Punctuation	Difference confined to punctuation / quotation / elision marks.	καταφανεῖς·→καταφανεῖς"; αὐτοῦ;→αὐτοῦ

Table 6: Word-level OCR error taxonomy. We align each prediction to its ground truth with a word-level edit-distance backtrace and assign every non-matching operation (substitution, insertion, deletion) to exactly one mutually-exclusive category, so per-model counts partition the errors. Before alignment we normalise Unicode, unify elision apostrophes, rejoin line-break hyphenation, and drop GT-only annotation markup (⟨ref⟩, section markers) that the model cannot see. The *Page furniture* category is a known confound: these tokens inflate the raw error totals of furniture-reading systems (notably the LM-free pipelines) even though most are correct reads of material the GT excludes. Per-model counts and the finer ten-way breakdown (real-word vs. non-word substitution, word splitting/merging) are in the released per-model CSVs.

substitutions— account for a large share of VLM errors, especially for Qwen3-VL-2B (accent/diacritic 57%), DeepSeek-OCR (53%), Qwen3-VL-8B (42%), and the OCR-specialist models OlmOCR-2-7B and LightOnOCR-1B (both 37%). Traditional systems show a different profile: Kraken is dominated by page-furniture and reference insertions (32%), while Tesseract distributes errors across punctuation, page furniture, overgeneration, and local recognition errors. Cross-script substitutions are rare across systems, but remain diagnostically useful because RQ2 shows that they are strongly image-driven. Whole-word substitutions also do not dominate any system. Instead, the relevant distinction is qualitative: VLM word substitutions are more often fluent Greek forms, whereas traditional substitutions more often reflect local recognition noise or segmentation artifacts. This motivates the RQ2 tests of whether fluent substitutions reflect prior-driven correction rather than visual confusion.

Traditional-OCR errors remain image-tied. The traditional baselines accumulate substantial absolute error, but their insertions are mostly tied to visible page content or local recognition noise. Many inserted tokens correspond to page furniture—running heads, numerals, apparatus, or bibliographic material that is visible in the scan but excluded from the transcription standard. Other insertions arise from stray glyphs, over-segmentation, hyphenation fragments, or short recognition artifacts rather than fluent continuation. Thus, although the traditional recognizers evaluated here can encode local sequential regularities, they lack the open-ended autoregressive decoder that allows VLMs to continue generating plausible text beyond the visual evidence. This distinction helps explain why traditional OCR remains comparatively faithful under the perturbations in RQ2, whereas VLMs more often drift toward plausible Greek.

C RQ2 Perturbation and Image-Gain Details

We probe reliance on language priors with two complementary methods. The *behavioral* counterfactual-perturbation experiment (Ap-

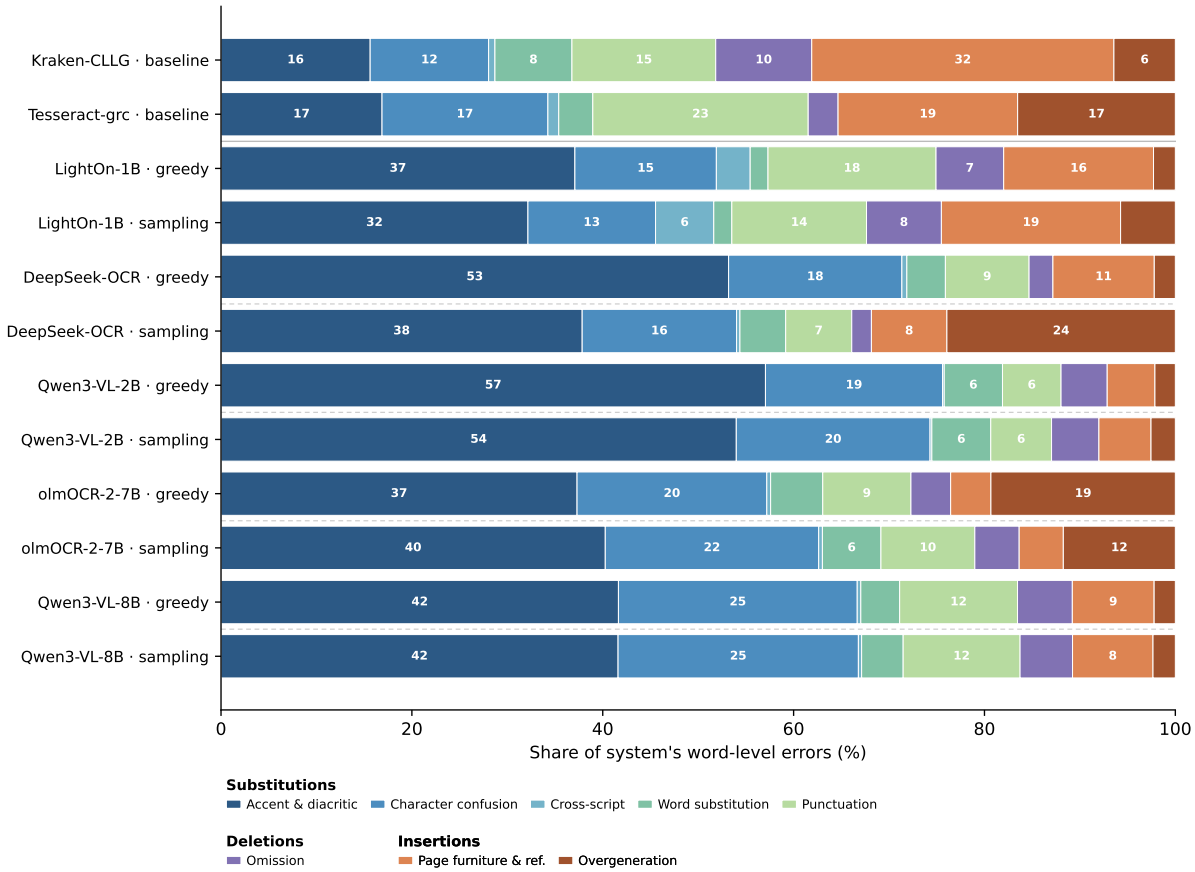

Figure 3: Composition of word-level errors per model on the real Greek test set, as the share of each model's errors falling in each category of Table 6; the right-hand column gives the absolute rate (errors per 1,000 GT words).

pendix C.1) asks whether a system transcribes what is on the page or silently "corrects" it toward fluent Greek. The *mechanistic* token-level image-gain analysis (Appendix C.2) measures, token by token, how much each emitted character actually depends on the image.

C.1 Counterfactual perturbation stimuli

Perturbation logic. A visually grounded reader should transcribe whatever glyphs are on the page, even when they spell nonsense; a reader that leans on a Greek language prior will "repair" implausible strings toward real words. We therefore perturb Ancient Greek text *before rendering it*, so the rendered image faithfully depicts the perturbed string and the ground truth *is* the perturbed string. Fidelity to the page then shows up as low CER against the perturbed ground truth, whereas prior-driven correction drives CER *up* as the rendered text becomes less word-like.

Source and rendering. Stimuli are built from 22 Ancient Greek prose documents which consists the CLLG synthetic test set. Each text is rendered to A5 pages with `lualatex` in the GFS Didot polytonic font (two visual styles per document, up to six pages per rendering) following Angleraud et al. (2026). The background-colored document images are from (Chagué and Scheithauer, 2023). Per-page ground truth is extracted directly from the compiled PDF with PyMuPDF, so there is no pagination or line-alignment uncertainty between image and reference.

Perturbation axes. We separate two prior types along orthogonal axes that share a single variant taxonomy:

- **Test 1 (word-level)** disrupts *word order* while leaving every word intact (a syntactic/positional prior).
- **Test 2 (character-level)** disrupts *characters within words* while leaving word boundaries intact (a lexical/orthographic prior).

Each axis applies the same operations (acting on words for Test 1, on characters within words for Test 2):

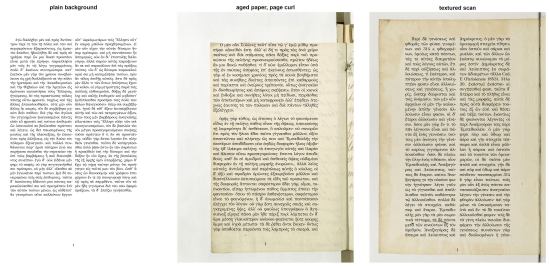

Figure 4: Rendered examples from the CLLG synthetic test set used in RQ2. Ancient Greek prose is typeset with `lualatex` in the GFS Didot polytonic font and composited onto document-style backgrounds at A5 size, producing controlled page images with varied visual styles.

- `swap_p{5,10,25}`: with probability p, swap adjacent units (adjacent words; or two adjacent characters in $p\%$ of words).
- `shuffle_p{5,10,25}`: relocate $p\%$ of words; or fully scramble the characters of $p\%$ of words.
- `local`: shuffle within non-overlapping windows of three units (words; or characters in every word).
- `reverse`: reverse the order of all units (whole-paragraph word order; or characters within each word).
- `random`: maximal disruption—per sentence, fully shuffle all words (Test 1) or permute all letters across word slots keeping per-word lengths (Test 2). The `local`, `reverse`, and `random` conditions are adapted from impossible-language perturbations, which use local shuffles, reversals, and random shuffles to remove natural word-order structure (Kallini et al., 2024). The graded `swap/shuffle` series follows Liang et al. (2026), who perturb OCR inputs to test whether VLMs rely on language priors rather than visual evidence.

Figure 4 shows example renderings and Figure 5 illustrates every condition on a single passage.

Scoring. Every system is scored against the *perturbed* ground truth. We report CER after stripping spaces and diacritics, isolating letter-level grounding from spacing and accent noise. The headline comparison (Table 2) restricts to single-column pages: two-column renderings introduce a reading-order confound (Kraken's segmenter and some VLMs mis-order columns) unrelated to grounding. We summarize each system's susceptibility by the clean→`random` CER increase (and the slope of CER across the disruption gradient), which we read as "prior-correction strength."

Word-level reordering leaves *all* systems near their clean CER ($\Delta \leq 0.05$ for every VLM): when each word is still a real word, there is nothing to "correct." Character-level corruption instead inflates every LM-equipped VLM (clean→`random` Δ up to $+0.68$ for Qwen3-VL-2B; $+0.21$–0.38 for the others) while the traditional baselines stay near-faithful (Tesseract $+0.06$, Kraken $+0.03$); all VLM increases are significant (paired one-sided Wilcoxon, $p < 0.001$). The asymmetry localizes the prior that overrides the image: it is *lexical/orthographic* (real-word) expectation, not sentence-level word order.

C.2 Token-level image-gain analysis

Image gain. For each model we teacher-force its *own* baseline transcription and, at every generated token t, compute

$$\text{gain}(t) = \log p(t \mid \text{image}, \text{prompt}) - \log p(t \mid \text{prompt}),$$

the log-probability the image adds to the emitted token (an attribution/faithfulness measure). The image-free pass uses the identical prompt with the image removed. We also record the top-1 probability and entropy of the image-conditioned distribution (calibration) and, via a logit lens, the share of probability mass on Greek vs. non-Greek vocabulary at intermediate layers.

Token labeling and the fair comparison. Each predicted character is aligned to the ground truth (`difflib`) and labeled `correct`, `substitution`, or `overgeneration`; a token inherits its characters' dominant label. Comparing all errors to all correct tokens is confounded—"correct" is dominated by easy, text-predictable tokens that are naturally low-gain, and the prior-driven failures concentrate on rare tokens. We therefore (i) compare *within Greek* ground-truth positions and (ii) split substitutions by the ground truth they replace: `perceptual` (same base letter, differing only in diacritic/case/final-sigma—a glyph confusion), `cross-script` (Greek→Latin look-alike—the script-mixing failure), and `lexical` (same script, different word—a fluent lexical swap). This pits error subtypes *against each other*, all at hard token positions. The analysis covers the three

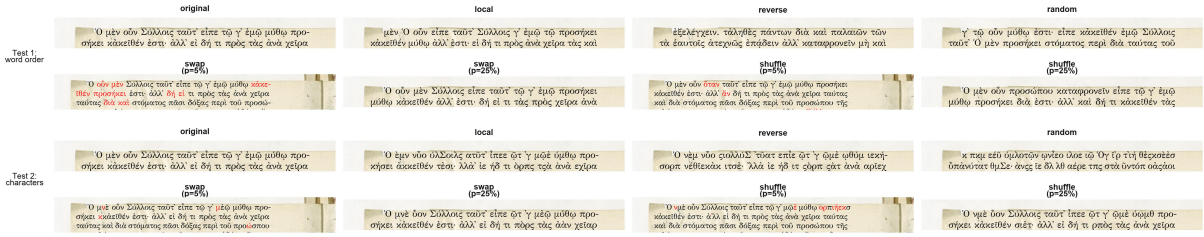

Figure 5: Perturbation conditions illustrated on one passage. **Test 1** is word-level: it changes word order while keeping each word intact. **Test 2** is character-level: it applies the same operations inside words, preserving word boundaries but turning words into non-words. `swap`/`shuffle` perturb a proportion p of units; `local`, `random`, and `reverse` apply stronger sequence disruptions.

	median image gain (within-Greek)			
Model	correct	percept.	cross-scr.	lexical
Qwen3-VL-2B	1.12	1.88	4.10	2.04
OlmOCR-2-7B	0.68	1.40	5.67	**0.17**
Qwen3-VL-8B	0.45	1.30	5.14	1.97

Table 7: Median image gain by token class for the three Qwen-VL–architecture models, restricted to Greek ground-truth positions. **percept.** denotes perceptual substitutions (same base letter, differing only in diacritic/case/final-sigma), **cross-scr.** Greek→Latin look-alikes, and **lexical** same-script substitutions to a different word. Higher values indicate stronger image dependence; the lowest lexical value is bolded.

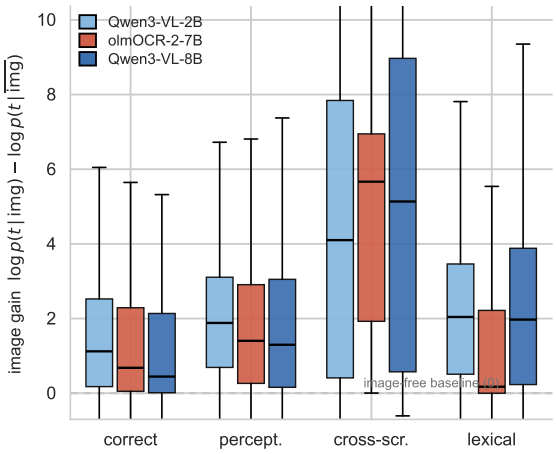

Figure 6: Distribution of token-level image gain by token class, for the three Qwen-VL–architecture models. Image gain is the log-probability an emitted token receives with the page image included minus without; positive values (above the dashed zero line) indicate image-supported predictions. Box = IQR, line = median, whiskers = 1.5×IQR; outliers omitted. Token classes are defined as in Table 7.

Qwen-VL–architecture models whose HuggingFace teacher-forcing path exposes well-defined image-conditioned logits (Qwen3-VL-8B/2B and OlmOCR-2-7B, bf16).

Table 7 reports within-Greek median image gain for the three Qwen-VL–architecture models with well-defined image-conditioned logits, and Figure 6 shows the corresponding per-token distributions. The low-image-gain signature of prior override appears clearly only for the OCR-specialist olmOCR: its lexical errors have median gain 0.17, far below its perceptual errors (1.40) and correct Greek tokens (0.68). The Qwen models show the opposite pattern: lexical errors remain image-engaged, with gain at or above perceptual errors (Qwen3-VL-8B: 1.97 vs. 1.30; Qwen3-VL-2B: 2.04 vs. 1.88). Across all three models, cross-script errors have the highest image gain (median 4–6), indicating that script mixing is visually triggered rather than a language-prior override.

Relation to the main-text claim. This token-level result refines the simplified statement in §4: the prior-override mechanism—a fluent error produced largely without the image—is *real but model-specific*, appearing as a low-image-gain signature only in the OCR-specialist olmOCR, while general-purpose VLMs remain grounded even when wrong and script mixing is visually triggered in all models.

D RQ3 Intervention Details

All interventions are evaluated on the same benchmark by (Angleraud et al., 2026) and scored with the same normalization as the main text (NFKC, structural-tag stripping). We report median CER as the headline metric because the prior-driven failures are heavy-tailed (Section 3).

Choice of interventions. The interventions are chosen to localize *where* grounding could be restored, not to build the strongest OCR pipeline. We

Model	Variant	CER (%) test1	CER (%) test2	WER (%) test1	WER (%) test2
Kraken-CLLG	raw	5.6	5.6	65.3	65.3
	no-diac.	5.1	5.1	64.1	64.1
Tesseract-grc	raw	6.0	6.0	62.2	62.2
	no-diac.	5.5	5.5	55.9	55.9
LightOnOCR-1B	raw	7.9	7.9	67.4	67.5
	no-diac.	6.3	6.3	60.9	61.0
DeepSeek-OCR	raw	8.0	8.0	65.9	65.5
	no-diac.	6.0	5.9	59.7	59.4
Qwen3-VL-2B	raw	19.4	18.7	91.8	91.1
	no-diac.	10.3	10.2	69.1	69.1
olmOCR-2-7B	raw	12.0	12.0	76.1	76.1
	no-diac.	8.6	8.7	65.1	65.1
Qwen3-VL-8B	raw	9.8	9.7	72.0	71.6
	no-diac.	6.6	6.6	63.0	63.0

Table 8: RQ2 synthetic benchmark, clean-control (`original`) condition only. Values are median per-page error rates with jiwer-based scoring, reported for raw text and a diacritic-insensitive variant. test1 and test2 are the word-level and character-level perturbation axes; under `original`, no perturbation is applied, so differences reflect page-subset differences.

use them as model-analysis probes: output constraints test whether errors are simple vocabulary leakage, contrastive decoding tests whether visual dependence can be up-weighted at decode time, and post-hoc correction tests how much can be repaired from text alone.

Engines. The output-space constraints, length abstention, and post-OCR correction run in the same inference engine as our main baseline. VCD and M3ID require per-step access to the token-level logit distribution for a second forward pass, which our serving engine does not expose; we therefore run them, *and their greedy baseline*, in HuggingFace Transformers and compare contrastive decoding only against this matched greedy baseline (Table 9) to avoid cross-engine confounds.

Script-restricted decoding. At each step we mask all tokens whose decoded string contains characters outside the Greek and Greek Extended ranges (`U+0370-03FF`, `U+1F00-1FFF`), Greek-relevant punctuation, and whitespace; special tokens (EOS, role markers) are always allowed so generation can terminate. Masking operates on the *decoded* token string, so byte-level BPE pieces that realize Greek characters are correctly admitted. Digits are excluded to prevent the model from copying printed page numbers and from entering punctuation/length loops. The eval retains numeric content (e.g., `<ref>7</ref>`→`7`), so the constraint is mildly penalized for forbidding it; however, the degradation is roughly an order of magnitude (10–17× in median CER across the four strong VLMs; DeepSeek-OCR is milder at ∼2.8×), far larger than the few digits per page could explain, indicating that these models structurally reallocate probability to non-Greek and markup continuations when their preferred tokens are masked.

Length abstention. We abstain on a page when the prediction-to-reference length ratio exceeds a fixed threshold (targeting overgeneration). Table 3 reports the kept-set median CER (the dagger), with an abstention rate of ∼12%. For every model, abstention improves 0 pages and merely removes the longest tails; the kept-set CER therefore equals the baseline by construction rather than reflecting better grounding.

VCD. Visual Contrastive Decoding contrasts the logits from the clean image against those from a diffusion-noised image, $\ell_{\text{cd}} = (1{+}\alpha)\,\ell_{\text{clean}} - \alpha\,\ell_{\text{noisy}}$, with the adaptive plausibility constraint of Leng et al. (2024) (β). We use $\alpha{=}1.0$, $\beta{=}0.1$, and a noise step of $500/1000$.

M3ID. M3ID replaces VCD's noised-image branch with an *image-free* (text-only) contrast, $\ell_{\text{cd}} = (1{+}w_t)\,\ell_{\text{image}} - w_t\,\ell_{\text{text}}$, with a strength that grows over the sequence, $w_t = \min(\alpha,\, e^{\gamma t}{-}1)$ (zero at $t{=}0$, ramping to a ceiling α), following Favero et al. (2024) to avoid the first-token collapse that a constant contrast induces. We use $\alpha{=}0.5$ and $\gamma{=}0.02$ for all three models with a no-repeat-3-gram constraint; for the olmOCR checkpoint we additionally apply a mild repetition penalty (1.15). We compare M3ID against its own $\alpha{=}0$ (no-contrast) pass rather than the engine's native greedy decode, because M3ID requires a custom decoding loop and we traced the degenerate single-token repetition (`6666. . .`) we initially observed to that loop, not to the contrast: a plain $\alpha{=}0$ pass through the same loop loops identically (median CER >1 on a majority of pages), whereas the engine's native greedy decode does not, and the no-repeat constraint removes the loop in both cases. Even with the loop controlled, the $\alpha{=}0$ code path sits well above the native greedy baseline (e.g. 0.245 vs. 0.073 on Qwen3-VL-8B), so M3ID must be read against its own $\alpha{=}0$ column rather than the greedy baseline (Table 9). Read this way, M3ID is essentially flat for the general VLMs and degrades CER for the OCR-specialist olmOCR, con-

	VCD		M3ID	
Model	greedy	VCD	α=0	M3ID
Qwen3-VL-8B	.073	.074	.245	.228
Qwen3-VL-2B	.104	.106	.252	.241
olmOCR-2-7B	.070	.083	.346	.369

Table 9: Contrastive grounding-aware decoding, reported as median CER on the 90-page benchmark for the three Qwen-VL–architecture models. VCD = Visual Contrastive Decoding; M3ID = image-free contrastive decoding. VCD is compared to its matched greedy baseline, and M3ID to its own α=0 no-contrast baseline.

sistent with the image-free contrast being out-of-distribution for a model trained to always condition on a page.

Post-OCR LM correction. Each baseline transcription is rewritten by Qwen3-VL-8B in text-only mode (no image) with a fixed 3-shot prompt of Greek printed-text exemplars. Table 10 gives the per-page breakdown: correction is significant for five of seven systems but is neutral on the two strongest VLM baselines (LightOnOCR-1B p=0.06; Qwen3-VL-8B p=0.60), and the mean gain exceeds the median gain throughout—i.e., it repairs heavy-tail pages rather than restoring visual grounding.

	pages (n=90)			ΔCER		
Model	help	tie	hurt	med	mean	p
Traditional baselines						
Tesseract-grc	68	11	11	−0.009	−0.015	<0.001
Kraken-CLLG	59	11	20	−0.003	−0.004	<0.001
Vision-language models						
LightOnOCR-1B	35	24	31	−0.003	−0.024	0.058
DeepSeek-OCR	56	16	18	−0.003	−0.007	<0.001
Qwen3-VL-2B	61	3	26	−0.003	−0.014	<0.001
olmOCR-2-7B	42	22	26	+0.001	−0.012	0.037
Qwen3-VL-8B	31	26	33	−0.000	+0.010	0.603

Table 10: Per-page effect of post-OCR LM correction (3-shot Qwen3-VL-8B text-only) vs. the unconstrained baseline on the 90-page benchmark. *help*/*tie*/*hurt* count pages whose CER decreased / was unchanged / increased; ΔCER is LM-correct minus baseline (negative = improvement); p is a one-sided Wilcoxon signed-rank test (H_1: LM-correct < baseline). Correction significantly reduces CER for five of seven systems (both traditional baselines, DeepSeek-OCR, Qwen3-VL-2B at p<0.001; olmOCR-2-7B at p=0.04); for the two strongest VLM baselines it is not significant (LightOnOCR-1B p=0.06; Qwen3-VL-8B p=0.60, essentially neutral—31 helped vs. 33 hurt). Where it helps, Δmean exceeds Δmed (e.g. Tesseract-grc −0.015 vs. −0.009): the correction disproportionately repairs heavy-tail pages rather than restoring grounding.