WHEN MLLMS MISS WHAT HUMANS READ: DIAGNOSING AND LEARNING TO READ SQUINT TEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal LLMs can fail to read text embedded in an image's coarse structure—text a human recovers by squinting—and report that no text is present. We name this class *Squint Text* and study it as a recognition deficiency, motivated by cases from a production moderation pipeline. We introduce SQUINTBENCH, comprising 1076 production-derived moderation images, and SQUINTCTRL, a controlled synthetic suite for mechanism diagnosis. Frozen-weight analysis shows that hidden text leaves a coarse signal while glyph identity is weakly represented; identity becomes substantially more accessible when the correct spatial regions are supplied, implicating spatial selection as a major bottleneck. Suppressing high-frequency content also improves reading at a fixed visual-token budget, while visual token count may also contribute. These findings motivate SQUINT OPSD: a frozen teacher receives a downscaled view and the ground-truth transcription as privileged context, and its token-level distributions supervise a student along on-policy trajectories. At inference, the student reads the original image alone. SQUINT OPSD reaches 88.61 ANLS on SQUINTBENCH, versus 26.50 for the unadapted backbone, 24.44 for the best general-purpose model on the original view, and 68.60 at best for an off-the-shelf model under a downscaled view. These results substantially narrow the gap between available visual evidence and its generated read-out without auxiliary views at inference.

1 INTRODUCTION

Multimodal LLMs can miss text embedded in an image's coarse structure even when a human can recover it by squinting or viewing the image at reduced scale. We call this failure ***Squint Text***; we study why the visual evidence is difficult for current models to access and show that the corrective viewing condition can be internalized through training, enabling original-only inference.

Modern VLMs are strong scene-text readers, yet this capability breaks sharply on *Squint Text*. As Figure 1 shows, frontier commercial and open-source models (Singh et al., 2026; Bai et al., 2025) can report that no text is present even when the hidden string is readily recoverable by a person. The failure also persists with model scale (Table 3). Downscaling or low-pass filtering recovers part of the missing text, but leaves a substantial residual. *Squint Text* therefore exposes a specific recognition deficiency that stronger generic perception and inference-time preprocessing do not reliably remove.

This deficiency matters in deployment. Images are widely used to carry illicit promotion and other policy-violating content (Meta, 2026c;b; X Corp., 2026; TikTok, 2026), and the harmful semantics can reside in text drawn into the image rather than in the depicted scene (Guo et al., 2024a; Kiela et al., 2020; Wu et al., 2025b). In production moderation traffic, we observe domains, contact strings, and slogans embedded into coarse image structure so that human viewers can read them while the reviewing model reports no text. Such examples create a direct blind spot at the visual-review stage of an automated moderation pipeline (Meta, 2026a).

To study the failure, we introduce SQUINTBENCH, containing 1076 production moderation images, and SQUINTCTRL, a controlled synthetic suite for isolating what the model represents and where the failure arises. Frozen-weight diagnostics show that hidden text leaves a strong coarse signal, while glyph identity is only weakly represented; identity becomes substantially more accessible when the correct spatial regions are supplied. Controlled image interventions on SQUINTBENCH itself

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

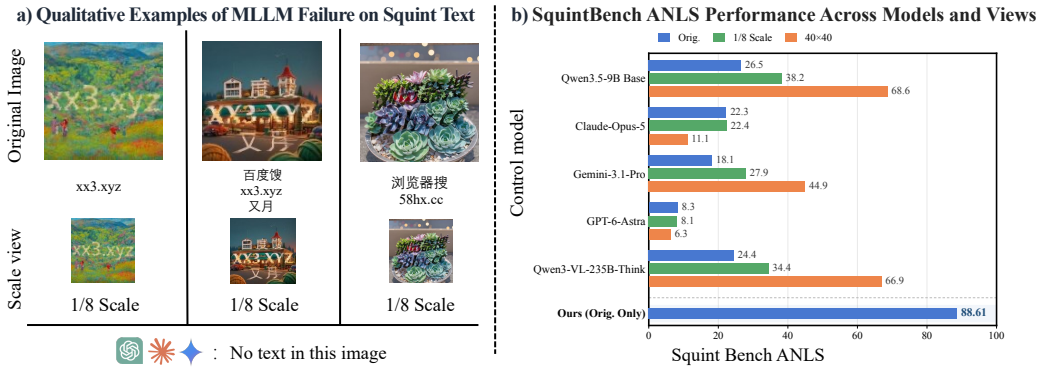


Figure 1: ***Squint Text*, and why the deployed defence does not remove it.** **(a)** Pipeline images whose embedded string a human reads—directly or by stepping back—while GPT, Gemini and Claude each answer *"No text in this image"*, on the original and the scaled view alike. **(b)** Downscaling before review recovers only part of the hidden text for every off-the-shelf model.

further show that suppressing high-frequency content can substantially improve reading at matched visual-token budgets, implicating high-frequency interference as a major contributor while leaving room for a secondary token-count effect. These findings motivate SQUINT OPSD: during training, a frozen teacher receives a downscaled view and the ground-truth transcription as privileged context, and its token-level distribution supervises a student that sees only the original image. At inference, the student receives the original image alone.

SQUINT OPSD reaches 88.61 ANLS on SQUINTBENCH, compared with 26.50 for the base model and 58.96 when the corrective view is supplied only at inference. The capability transfers to external hidden-text benchmarks while retaining substantial general vision-language capability with measurable, task-dependent trade-offs. Analysis after training shows that glyph representations become more discriminative, reading depends causally on glyph-bearing regions under our visual-token embedding intervention, and the residual candidate distribution becomes more visually structured. Code and the evaluation suite are released at `https://anonymous.4open.science/r/Squint_Text-3456`.

Contributions.

- **A deployment-grounded failure mode and two complementary datasets.** We define *Squint Text* and introduce SQUINTBENCH, comprising 1076 production moderation images, for realistic evaluation, together with SQUINTCTRL for controlled diagnosis of hidden-text representation, localization, and recoverability.
- **A diagnosis of why current MLLMs miss *Squint Text*.** Hidden text leaves a strong coarse signal but weak glyph identity; the identity signal becomes more accessible at glyph-bearing locations, and high-frequency interference is a major contributor, with token count a possible secondary factor.
- **A diagnosis-driven training method.** SQUINT OPSD uses a downscaled privileged view together with a ground-truth hint to provide token-level distributional supervision for original-only inference, reaching 88.61 ANLS on SQUINTBENCH.

2 PRELIMINARIES AND DATA

2.1 PROBLEM SETTING

Given an image x and an instruction to transcribe its text, the model must report every string a human reader would, including text carried in the image's coarse structure. We use *read-out* to refer to the mapping from encoded visual representations to the generated transcription. Let $v = g(x)$ denote a bicubically downscaled view of x, at $1/8$ linear scale unless stated otherwise. We compare a single-image input x with a two-view input $\langle x, v \rangle$ to test whether changing the viewing condition makes the same underlying visual evidence easier for a frozen model to use. Section 3 performs this comparison with identical off-the-shelf weights; the training method is introduced separately in Section 4.

On SQUINTBENCH we report ANLS (Biten et al., 2019), with predicted and reference regions matched as a set before scoring so that a correct transcription written in a different order is not penalised.

2.2 DATASETS

Showing that the phenomenon occurs in deployment requires real model failures; attributing the failure to a particular mechanism requires controlled examples. We therefore use two complementary datasets.

SQUINTBENCH comprises 1076 squint-text images collected from a production moderation pipeline. All 1076 examples are production-derived. These images are typically profile photos, in which a human reader recognizes the embedded text almost effortlessly. Human annotators transcribe all visible and hidden text. Thus, SQUINTBENCH is used as the realistic end-to-end evaluation set.

SQUINTCTRL supplies the controlled counterpart. A glyph string is rendered as a binary conditioning image and embedded into a scene using QR Code Monster v2 (Monster Labs, 2023), a ControlNet (Zhang et al., 2023) over Stable Diffusion 1.5 (Rombach et al., 2022), with scene prompts sampled from JourneyDB (Sun et al., 2023). For each controlled comparison, non-target generation variables are held fixed; because the conditioning image is known, each sample also has an exact token-level glyph mask without detector or human annotation.

The four subsets serve distinct diagnostics. SQUINTCTRL-PAIR contains 500 matched text-present/text-free pairs for testing whether hidden-text presence is encoded. The text-free member is regenerated through the same ControlNet branch using an all-black conditioning image, avoiding a synthesis pipeline shortcut. SQUINTCTRL-GLYPH contains two 500-image closed vocabularies over ten Chinese glyphs (加女可度搜月百看网馊) and ten Latin letters for glyph-identity and localization analyses. SQUINTCTRL-STRENGTH varies 200 scenes across 5 conditioning strengths to study the recoverability boundary. SQUINTCTRL-READ contains 200 open-vocabulary images for end-to-end read-out and post-training analysis. Detailed generation and control protocols are provided in Appendix F.6, and representative examples of each subset are shown in Figure 8.

3 DIAGNOSING SQUINT TEXT

Unless stated otherwise, all diagnostic experiments use Qwen3.5-9B (Qwen Team, 2026) with frozen weights. We report set-matched ANLS for transcription and out-of-fold AUC for probing. The failure is not confined to this checkpoint: four commercial and open-weight MLLMs obtain only 8.3–24.4 ANLS on the original SQUINTBENCH images (Table 3).

3.1 HIDDEN-TEXT REPRESENTATION

To test what survives in the vision encoder, we fit logistic-regression probes to frozen, pre-merger patch features at blocks L6/12/18/24/27. Each probe uses mean-pooled features and is evaluated out of fold against shuffled-label controls; fitting it does not update the MLLM. On SQUINTCTRL-PAIR, whose 500 pairs differ only in whether a string was embedded, text presence reaches 0.956–0.981 AUC. Label shuffling returns chance-level performance (0.47–0.52 AUC), while matched generation paths control for synthesis-pipeline artefacts (Appendix A.3). Going beyond mere presence, we further test how well each individual glyph can be read out on the closed-vocabulary SQUINTCTRL-GLYPH set. Glyph identity is weaker than presence but remains above chance: Chinese macro-AUC peaks at 0.718 at L18 and averages 0.676 across the five blocks, with Latin at 0.612 and 0.581, respectively. Thus, the encoder detects hidden text, although glyph identity is only weakly represented.

3.2 SPATIAL LOCALIZATION

We extract visual features before the patch merger and, using the known SQUINTCTRL-GLYPH mask, compare three poolings of the same features: ORACLE averages vectors at glyph-bearing locations, GLOBAL averages all vectors, and RANDOM averages a random subset of the *same size as the* ORACLE *set*. RANDOM controls for sample count: its mean estimates the global mean and empirically differs from it by at most 0.004 AUC. Averaged over five blocks, ORACLE/GLOBAL/RANDOM reach 0.799/0.676/0.675 AUC on Chinese and 0.664/0.581/0.580 on Latin, respectively (Figure 2a),

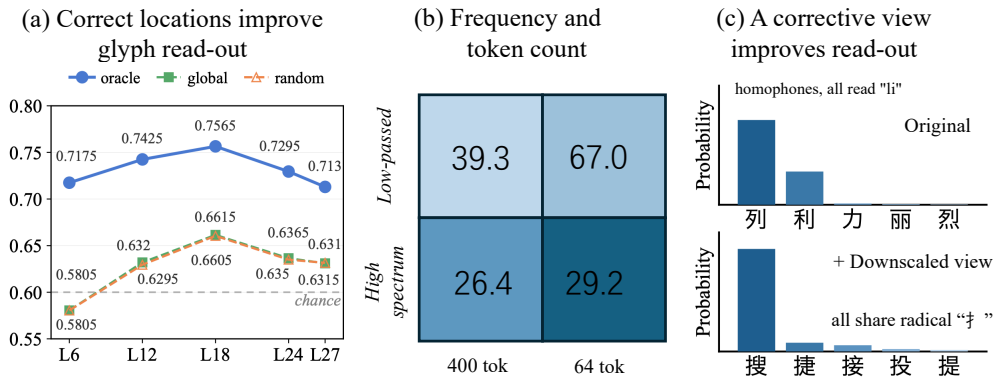


Figure 2: **Three diagnostics on frozen off-the-shelf weights. (a)** Restricting the read-out to glyph-bearing tokens improves glyph-identity prediction, supporting spatial selection as a major bottleneck (per-layer expansion in Appendix A). **(b)** At a fixed 64-token budget, suppressing high-frequency content raises squint-text ANLS from 29.2 to 67.0 (per-σ results in Appendix B). **(c)** For the same target glyph 搜, adding a $1/8$-scale view of the same image shifts the base model's phonetic predictions toward visually matching glyphs, indicating stronger visual grounding.

so ORACLE improves by about 0.10 AUC over both controls. Glyph identity is therefore substantially more accessible at the correct locations, identifying spatial selection as a major bottleneck (Wu & Xie, 2024; Zhang et al., 2025). Passing these features through the patch merger changes separability by at most 0.018 AUC (Appendix A.2).

3.3 FREQUENCY AND TOKEN COUNT

That downscaling or low-pass filtering makes such text legible is already established—HC-Bench evaluates after downscaling (Li et al., 2025) and Illusory VQA after low-pass filtering (Rostamkhani et al., 2025). What is not established is *which* property of those transforms is responsible, and the answer determines what one should build: (**H1**) high-frequency texture masks the glyph structure; (**H2**) too many tokens dilute attention; (**H3**) it is a contrast problem.

We test these explanations on SQUINTBENCH images (Figure 2b). At the original token budget (median 400), Gaussian low-pass filtering raises ANLS from 26.4 to 39.3, showing that suppressing high frequencies helps without shortening the visual sequence. Restricting the visual input to 64 tokens yields 29.2; applying a stronger low-pass transform at the same token budget raises ANLS to 67.0. This matched-budget comparison independently supports H1. However, constructing the 64-token input also alters its frequency content, so H2 cannot be fully disentangled from H1. We additionally test H3 with contrast-only transforms, which yield substantially smaller gains than frequency suppression (Appendix B), making contrast an unlikely primary explanation. Overall, these results identify high-frequency interference as a major contributor, while visual token count may also contribute, although its effect is not fully disentangled from frequency.

3.4 CORRECTIVE VIEWING

We next give the same frozen model the two-view input $\langle x, v \rangle$, with no change to its weights or prompt. On SQUINTBENCH, adding the $1/8$-scale view raises ANLS from 26.50 to 58.96. Figure 2c shows the corresponding shift from phonetic alternatives toward candidates that share the target's visual structure. Because $v = g(x)$ is deterministic and lossy, the gain is consistent with the same frozen weights exploiting coarse visual evidence more effectively under a different viewing condition. The result motivates learning a similar read-out behaviour while retaining original-only input.

Design requirements. The diagnosis yields three requirements. **(D1) Original-only inference:** the solution should receive x alone, because the needed evidence already survives in its representation. **(D2) Localization-bearing supervision:** a transcript says what the glyph is but not where its evidence lies, whereas predictions elicited by the downscaled view reflect evidence that has become easier to select. The method should transfer this implicit spatial information; whether it does so is

Table 1: Candidate solutions under the requirements of §3.4: **D1** uses original-only inference, **D2** uses localization-bearing supervision, and **D3** uses a distributional target. A dash denotes a method with no training target. Only SQUINT OPSD satisfies all three.

	D1	D2	D3
ground-truth SFT	✓	✗	✗
sparse-reward RL	✓	✗	✗
downscaled view at inference	✗	—	—
ground-truth-conditioned teacher	✓	✗	✓
SQUINT OPSD (view + hint teacher)	✓	✓	✓

tested in §6. **(D3) Distributional targets:** at each step the teacher assigns probability mass over possible next tokens, not only the ground-truth one; the relative probabilities of alternatives retain which glyphs remain visually plausible — information a one-hot label cannot express.

4 METHOD

Table 1 suggests supervising the student under a viewing condition in which Squint Text is easier to read. During training, the teacher receives privileged context consisting of the downscaled view and the ground-truth output, while the student receives only the original image. Inference remains original-only. Figure 3 illustrates the training framework.

4.1 SQUINT OPSD

Let x be the original image, $v = g(x)$ its downscaled view, q the task instruction, and h the ground-truth output provided only to the teacher. The student is $p_\theta(y \mid x, q)$ and the teacher is a frozen copy of the initial backbone conditioned on (x, v, q, h). At each update, we first sample a trajectory $\hat{y}$ from the current student policy and treat the sampled prefixes as stop-gradient states:

$$\hat{y} \sim \mathrm{sg}[p_\theta(\cdot \mid x, q)]\,. \tag{1}$$

We then minimize token-level Jensen–Shannon divergence between teacher and student distributions along that student-generated trajectory:

$$\mathcal{L}_{\mathrm{OPSD}}(\theta) = \mathbb{E}_x\, \mathbb{E}_{\hat{y}\sim \mathrm{sg}[p_\theta(\cdot|x,q)]} \sum_{t=1}^{T(\hat{y})} \mathrm{JSD}\big(\widetilde{p}_T(\cdot \mid x, v, q, h, \hat{y}_{<t}) \,\big\|\, \widetilde{p}_\theta(\cdot \mid x, q, \hat{y}_{<t})\big)\,, \tag{2}$$

where gradients flow only through the student logits. For each step, the teacher's top-K tokens define the explicit support; probability mass outside this support is aggregated into a single tail bucket for both teacher and student before computing JSD. We use $K = 100$. Rollouts terminate at EOS or the maximum response length. Full rollout and optimization details are given in Appendix F.3.

The privileged view and the ground-truth hint play different roles. The view changes the visual condition under which the teacher reads the image, whereas the hint supplies privileged label information. Their contributions are separated empirically in Table 4 and Appendix C.

4.2 PRIVILEGED VIEW

Since $v = g(x)$ is a deterministic, lossy transform of the original image, it does not introduce new Shannon information conditional on x. Its value is therefore computational (Xu et al., 2020): the same visual content can become easier for a fixed model to exploit after high-frequency suppression and scale change. The ground-truth hint h, by contrast, does provide privileged label information. Accordingly, we do not assume that the complete teacher distribution $p_T(\cdot \mid x, v, q, h)$ is exactly realizable from (x, q) alone. Our objective is empirical: transfer the components of the teacher's reading behavior that the student can acquire while retaining original-only inference.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets and metrics. We train on 5905 prompts, including 3489 synthetic hidden-text images, 1213 OCR samples, and 1203 VQA samples. Our primary evaluation is set-matched ANLS on all

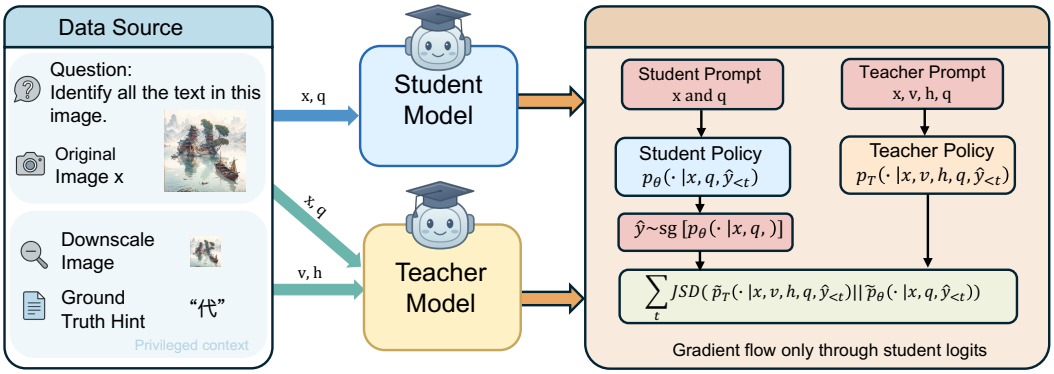


Figure 3: **Overview of SQUINT OPSD.** The teacher receives the original image, the downscaled view $v = g(x)$, the task instruction, and the ground-truth output as privileged context. Its token distributions supervise the student along stop-gradient, student-generated trajectories. Gradients flow only through the student logits, and inference uses the original image alone.

Table 2: SQUINTBENCH performance and general vision-language capability.

	SQUINTBENCH	general		
		AI2D	DocVQA	OCRBench
BASE	26.50	86.66	95.93	879
BASE $+1/8$ view (two-view)	58.96	52.36	95.37	834
SFT (full, single)	72.47	85.52	64.73	722
DAPO	73.55	82.12	92.31	728
SQUINT OPSD	**88.61**	85.33	92.98	833

1076 deployment-derived SQUINTBENCH images. We further evaluate transfer on Illusory VQA (Rostamkhani et al., 2025), AdvOCR (Xu et al., 2026), and HC-Bench (Li et al., 2025), and general vision-language capability on AI2D (Kembhavi et al., 2016), DocVQA (Mathew et al., 2021), and OCRBench (Liu et al., 2024). We reserve a 1,045-example held-out validation split from the same source mixture as training.

Baselines. We compare against the pretrained backbone (BASE), appending a downscaled view at inference (BASE $+1/8$ view, two-view), single-view SFT, and DAPO (Yu et al., 2025). For trained baselines, we focus on methods that use the original image at inference, matching the deployment constraint of SQUINT OPSD. We also report four off-the-shelf MLLMs as external references: claude-opus-5, gemini-3.1-pro, gpt-6-astra, and Qwen3-VL-235B-Think.

Model and training. All trained methods use Qwen3.5-9B. Unless stated otherwise, all parameters, including the vision tower, are updated. In the current experiments, the teacher is a frozen copy of the initial backbone and receives the privileged downscaled view and ground-truth hint, while the student receives only the original image. SQUINT OPSD, DAPO, and SFT are all trained for three epochs. Full optimization and rollout details are in Appendix F.3; the held-out mixture is described in Appendix F.2.

Inference. SQUINT OPSD receives only the original image at inference. Additional scaled views are provided only to baselines explicitly marked with them.

5.2 MAIN RESULTS

Target performance. SQUINT OPSD reaches 88.61 ANLS on SQUINTBENCH, compared with 26.50 for BASE and 58.96 when the corrective view is appended only at inference (Table 2). It also exceeds single-view SFT (72.47) and DAPO (73.55). SFT improves the target metric only partway and forgets heavily: its DocVQA falls from 95.93 to 64.73 and OCRBench from 879 to 722. SFT warm starts likewise underperform direct OPSD, suggesting that first fitting a hard one-hot target

Table 3: Comparison with the base model and off-the-shelf MLLMs on SQUINTBENCH and three external hidden-text benchmarks. WER is lower better. The "only" conditions replace the original image with the reduced view, applied before the model preprocessor; further protocol details are in Appendix B.1.

benchmark	cond.	SQUINT OPSD	Qwen3.5-9B base	claude-opus-5	gemini-3.1-pro	gpt-6-astra	qwen3-vl-235b-think
SQUINTBENCH	orig.	**88.61**	26.50	22.26	18.15	8.33	24.44
	$1/8$ only	**49.64**	38.22	22.41	27.90	8.13	34.37
	40×40 only	**76.12**	68.60	11.07	44.91	6.26	66.88
IllVQA *Illusion*	WER↓	**67.00**	178.10	166.50	190.13	119.60	177.97
IllVQA *Filtered*	WER↓	**64.87**	76.80	83.53	70.00	65.27	91.33
AdvOCR[†]	orig.	**66.7**	3.7	3.7	3.7	11.1	3.7
	$1/8$ only	**70.4**	44.4	33.3	33.3	0.0	51.9
	40×40 only	**70.4**	51.9	22.2	44.4	44.4	55.6
HC-Bench	orig.	**44.6**	1.8	5.4	5.4	1.79	1.8
	$1/8$ only	**46.4**	26.8	19.6	33.9	12.5	28.6
	40×40 only	19.6	19.6	8.9	19.6	12.5	**25.0**

Table 4: Ablations on SQUINTBENCH. We vary teacher input, target form, and trainable modules. [†]The mismatched-view row is paired with a matched-view run under the same setting (84.60), giving a 6.0-point drop.

configuration		SQUINTBENCH
teacher input		
x	(neither)	27.96
$x + h$	(hint only)	41.79
$x + v$	(privileged view only)	54.68
$x + v + h$	(default)	**88.61**
$x + v_{\text{mis}} + h$	(mismatched view)[†]	78.60
$x + v + h_{\text{wrong}}$	(wrong hint)	64.00
target signal		
ground truth, one-hot	(SFT)	72.47
teacher, one-hot	(SFT-distil)	68.31
teacher, top-K distribution	(SQUINT OPSD)	**88.61**
*trainable modules (*SQUINT OPSD*)*		
LM only, vision tower frozen		29.25
vision tower only, LM frozen		71.12
all parameters		**88.61**

is less effective than preserving distributional supervision from the outset. We also ran a series of teacher-view ablations that vary only v in the fixed teacher input (x, v, q, h); none of the alternative views or their aggregations improves on the default (Appendix C).

External transfer. The capability transfers to external hidden-text benchmarks (Table 3). On original inputs, SQUINT OPSD reaches 66.7 on AdvOCR and 44.6 on HC-Bench, compared with 3.7 and 1.8 for BASE. On Illusory VQA, it reduces WER from 178.10 to 67.00, even outperforming BASE given the filtered view (76.80). Increasing model scale alone also does not remove the failure: Qwen3-VL-235B (Bai et al., 2025) reaches only 24.44 on the original SQUINTBENCH images, and gpt-6-astra only 8.33.

General-capability trade-off. The hidden-text gain is accompanied by task-dependent degradation on the general benchmarks: AI2D changes from 86.66 to 85.33, DocVQA from 95.93 to 92.98, and OCRBench from 879 to 833. Thus, the adaptation retains substantial general vision-language capability, but introduces a measurable specialization trade-off.

5.3 Ablations

Privileged teacher context. Neither signal alone gets far: hint-only teacher input yields 41.79, view-only yields 54.68, and x-only yields 27.96. Combining the view and hint reaches 88.61, substantially above either signal alone. Independently, swapping the view for another sample's downscaled image reduces the paired score from 84.60 to 78.60, and corrupting the hint with a wrong transcription drops it to 64.00. The two privileged signals are therefore complementary, and each must be aligned with the current sample: the view contributes sample-specific visual context rather than merely extra low-resolution tokens, while the hint supplies label information beyond what the view alone provides.

Distributional supervision. Replacing the teacher distribution with its one-hot prediction drops performance from 88.61 to 68.31, below even ground-truth SFT (72.47). The gain therefore comes from information carried by the distribution rather than from better teacher hard labels. Section 6 examines this structure directly.

Where adaptation occurs. Updating only the language side reaches 29.25, whereas updating only the vision tower already reaches 71.12. Joint adaptation further improves to 88.61. Thus, visual adaptation accounts for a large part of the gain, while the strongest read-out requires both sides to adapt. The same dependence appears under supervised fine-tuning (Appendix F.4), supporting the representation change analyzed in §6.

Teacher and training variants. Entropy-based teacher selection, EMA teacher updates, and SFT warm starts likewise underperform the fixed-teacher recipe (Appendix C).

6 Analysis

We analyze three changes induced by training: glyph representation, spatial selection, and candidate-distribution structure. The probe and candidate-distribution analyses reuse the protocols established in §3.

Glyph representations become more discriminative. Re-running the probes from §3.1, glyph-identity AUC rises from 0.718 to 0.975 on Chinese and from 0.612 to 0.779 on Latin, while near-ceiling presence detection changes little. Together with the frozen-vision ablation in §5.3, this shows that training reorganizes the visual representation to make glyph identity more accessible.

Reading becomes spatially grounded. After training, answer-token attention (Vaswani et al., 2017) is more concentrated on glyph-bearing regions, with similar patterns under free generation and teacher forcing (Figure 4b). We complement this correlational evidence with a visual-token embedding intervention. Before the visual tokens enter the language model, we replace glyph-bearing embeddings with the mean embedding of non-glyph tokens from the same image, while keeping token count and positions unchanged. At dilation $d = 2$, this reduces SQUINT OPSD accuracy from 90.5% to 71.5%, whereas replacing an equal number of randomly selected background-token embeddings leaves performance nearly unchanged (90.5% to 91.0%). The result supports a causal dependence of the learned reading behavior on glyph-bearing visual regions. Full intervention details are given in Appendix E.5.

The distribution acquires visual structure. We compare non-target, high-probability candidates on the same 120 glyph positions valid for all three models. We measure ρ_{vis} as the Spearman correlation between candidate probability and pixel-IoU similarity to the target glyph, and ρ_{lang} as the correlation between candidate probability and its matched text-only probability. Visual organization increases from $\rho_{\mathrm{vis}} = 0.126$ for BASE to 0.242 for SFT and 0.286 for SQUINT OPSD. Meanwhile, SQUINT OPSD retains a text-only-prior correlation close to BASE ($\rho_{\mathrm{lang}} = 0.544$ vs. 0.555), whereas SFT drops to 0.334. Thus, SQUINT OPSD more strongly organizes high-probability alternatives by glyph compatibility while retaining language-side structure. Definitions and candidate-set sensitivity are in Appendix E.1.

7 Related Work

Hidden and squint-visible content. HC-Bench studies hidden content with downscaling (Li et al., 2025), Illusory VQA applies low-pass filtering (Rostamkhani et al., 2025), and AdvOCR/VACoT learns when to zoom in or out at inference (Xu et al., 2026); training-free alternatives steer a frozen

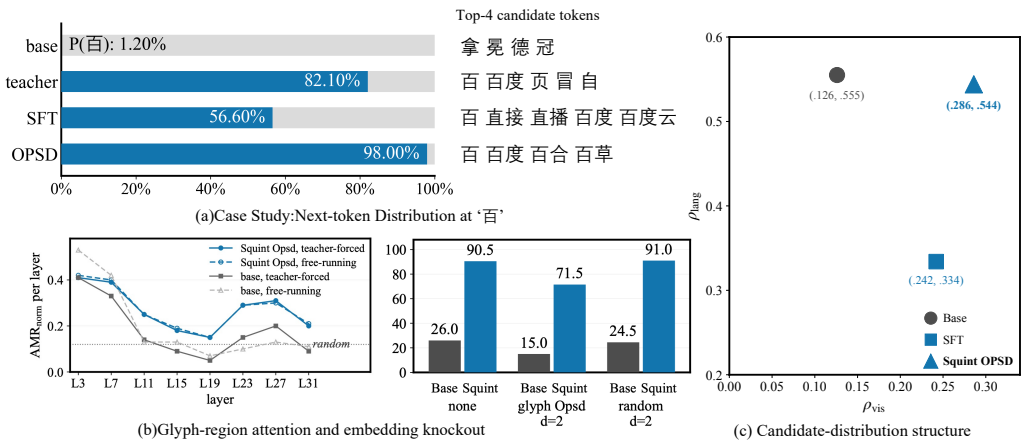


Figure 4: **What changes after training. (a)** Example next-token distributions for BASE, SFT, and SQUINT OPSD. **(b)** Attention concentration on glyph-bearing regions and a causal visual-token embedding knockout. **(c)** Candidate-distribution structure on 120 matched glyph positions.

model's search toward small visual details (Wu & Xie, 2024; Zhang et al., 2025). These works alter the viewing condition or the search process at inference. To our knowledge, SQUINT OPSD is the first work on squint-visible hidden text to address this failure through post-training for original-only inference. SQUINTBENCH adds deployment-derived failures and SQUINTCTRL supports controlled diagnosis; the frequency effect is related to hybrid images (Oliva et al., 2006) and image-scaling attacks (Quiring et al., 2020), a principle recently productized by hidden-content fonts (Mixfont, 2026), while text-to-image generators render the coarse-structure channel increasingly accessible (Wu et al., 2025a).

Text-rich reading and privileged supervision. Prior work improves pixel-text understanding through reading-oriented data and architectures (Zhang et al., 2024), benchmarks text-rich comprehension (Li et al., 2024), and studies the symbolic-to-pixel modality gap (Sun et al., 2026; Fu et al., 2025). Global–local methods provide additional visual views (Lu et al., 2026), and native high-resolution input pipelines enlarge the visual budget available to small text (Guo et al., 2024b; Chen et al., 2024). Related privileged-information and on-policy distillation methods use signals available only during training (Vapnik & Vashist, 2009; Hinton et al., 2015; Agarwal et al., 2024; Zhao et al., 2026; Yuan et al., 2026; Li et al., 2026). These works largely target fine-grained visual detail retrieval.

Probing visual representations. Linear probes test information in intermediate representations (Alain & Bengio, 2018; Burns et al., 2023), while VLM work localizes object information to visual tokens (Neo et al., 2025) and identifies outlier artifact tokens that distort ViT feature maps (Darcet et al., 2024). We study coarse text whose identity becomes more accessible at the correct locations and how this changes after adaptation.

8 CONCLUSION

We introduce Squint Text, SQUINTBENCH, and SQUINTCTRL to study text embedded in coarse image structure that humans can recover but MLLMs often miss. Our diagnostics implicate spatial selection and high-frequency interference, with visual-token count as a possible additional factor. These findings motivate SQUINT OPSD, which uses a downscaled view and ground-truth transcription as privileged training context while retaining original-only inference. SQUINT OPSD reaches 88.61 ANLS on SQUINTBENCH versus 26.50 for the base model and transfers to external hidden-text benchmarks with measurable general-capability trade-offs. After training, glyph identity becomes more discriminative, reading depends on glyph-bearing regions under our visual-token embedding intervention, and high-probability alternatives become more visually structured while retaining language-side structure. The method currently relies on recoverable visual evidence and fixed privileged supervision.

ETHICS STATEMENT

SQUINTBENCH is derived from production content-moderation traffic and contains domains, contact strings and policy-violating promotional material, so it raises privacy and dual-use concerns that the fully synthetic SQUINTCTRL suite does not. The benchmark is hosted on Hugging Face and will be released openly, but because its images embed real websites, access to the production-derived portion is gated and granted on approved request; we study these patterns solely to improve the robustness of visual moderation.

AI USE STATEMENT

We used generative AI tools for code implementation, translation of the Chinese prompt templates, and language editing of the manuscript. All AI-assisted output was reviewed and verified by the authors against the raw evaluation logs and the original sources, and we take responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

Code, the evaluation protocol and the synthetic SQUINTCTRL suite are released at `https://anonymous.4open.science/r/Squint_Text-3456`; SQUINTBENCH is hosted on Hugging Face and will be open-sourced, with the production-derived portion gated behind an approved request. Training, optimisation, prompt and preprocessing details are given in Appendices.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. URL https://arxiv.org/abs/1610.01644.

Shuai Bai et al. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marcal Rusinol, Ernest Valveny, C.V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

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. 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 (CVPR)*, pp. 24185–24198, June 2024.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Pei Fu, Tongkun Guan, Zining Wang, Zhentao Guo, Chen Duan, Hao Sun, Boming Chen, Qianyi Jiang, Jiayao Ma, Kai Zhou, and Junfeng Luo. Multimodal large language models for text-rich image understanding: A comprehensive review. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19941–19958, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1023. URL https://aclanthology.org/2025.findings-acl.1023/.

Keyan Guo, Ayush Utkarsh, Wenbo Ding, Isabelle Ondracek, Ziming Zhao, Guo Freeman, Nishant Vishwamitra, and Hongxin Hu. Moderating illicit online image promotion for unsafe user generated content games using large Vision-Language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 5787–5804, Philadelphia, PA, August 2024a. USENIX Association. ISBN 978-1-939133-44-1. URL https://www.usenix.org/conference/usenixsecurity24/presentation/guo-keyan.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: An LMM perceiving any aspect ratio and high-resolution images. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXIII*, volume 15141 of *Lecture Notes in Computer Science*, pp. 390–406. Springer, 2024b. doi: 10.1007/978-3-031-73010-8_23. URL https://doi.org/10.1007/978-3-031-73010-8_23.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pp. 1–6, 2019. doi: 10.1109/ICDARW.2019.10029.

Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere de las Heras. Icdar 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1484–1493, 2013.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 235–251, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2611–2624. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1b84c4cee2b8b3d823b30e2d604b1878-Paper.pdf.

C. Lanczos. Evaluation of noisy data. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, 1(1):76–85, 1964. doi: 10.1137/0701007. URL https://doi.org/10.1137/0701007.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension, 2024. URL https://arxiv.org/abs/2404.16790.

Pengyu Li, Zhitao Gao, Lingling Zhang, Muye Huang, Yuanming Li, Zesheng Yang, Fangzhi Xu, and Jun Liu. Visual-opsd: Cross-modal on-policy self-distillation for efficient unified multimodal reasoning, 2026. URL https://arxiv.org/abs/2606.18974.

Sifan Li, Yujun Cai, and Yiwei Wang. SemVink: Advancing VLMs' semantic understanding of optical illusions via visual global thinking. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27167–27177, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1381. URL https://aclanthology.org/2025.emnlp-main.1381/.

Cheng-Lin Liu, Fei Yin, Da-Han Wang, and Qiu-Feng Wang. Casia online and offline chinese handwriting databases. In *2011 International Conference on Document Analysis and Recognition*, pp. 37–41, 2011. doi: 10.1109/ICDAR.2011.17.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024. doi: 10.1007/s11432-024-4235-6.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Junxin Lu, Tengfei Song, Zhanglin Wu, Pengfei Li, Xiaowei Liang, Hui Yang, Kun Chen, Ning Xie, Yunfei Lu, Jing Zhao, Shiliang Sun, and Daimeng Wei. Global-local dual perception for mllms in high-resolution text-rich image translation, 2026. URL https://arxiv.org/abs/2602.21956.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2200–2209, January 2021.

Meta. Community standards enforcement report. https://transparency.meta.com/reports/community-standards-enforcement/, 2026a.

Meta. Community standards: Hateful conduct. https://transparency.meta.com/policies/community-standards/hate-speech/, 2026b.

Meta. Community standards: Restricted goods and services. https://transparency.meta.com/policies/community-standards/regulated-goods/, 2026c.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019.

Mixfont. Decoy font. https://mixfont.com, 2026. TTF font released July 15, 2026.

Monster Labs. QR Code Monster v2: ControlNet for Stable Diffusion 1.5 (control_v1p_sd15_qrcode_monster). https://huggingface.co/monster-labs/control_v1p_sd15_qrcode_monster, 2023.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chanJGoa7f.

Aude Oliva, Antonio Torralba, and Philippe G. Schyns. Hybrid images. *ACM Trans. Graph.*, 25 (3):527–532, July 2006. ISSN 0730-0301. doi: 10.1145/1141911.1141919. URL https://doi.org/10.1145/1141911.1141919.

Stephen M. Pizer, E. Philip Amburn, John D. Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B. Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, 1987. ISSN 0734-189X. doi: https://doi.org/10.1016/S0734-189X(87)80186-X. URL https://www.sciencedirect.com/science/article/pii/S0734189X8780186X.

Erwin Quiring, David Klein, Daniel Arp, Martin Johns, and Konrad Rieck. Adversarial preprocessing: Understanding and preventing Image-Scaling attacks in machine learning. In *29th USENIX Security Symposium (USENIX Security 20)*, pp. 1363–1380. USENIX Association, August 2020. ISBN 978-1-939133-17-5. URL https://www.usenix.org/conference/usenixsecurity20/presentation/quiring.

Qwen Team. Qwen3.5-9B. https://huggingface.co/Qwen/Qwen3.5-9B, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Mohammadmostafa Rostamkhani, Baktash Ansari, Hoorieh Sabzevari, Farzan Rahmani, and Sauleh Eetemadi. Illusory vqa: Benchmarking and enhancing multimodal models on visual illusions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2986–2995, 2025. doi: 10.1109/CVPRW67362.2025.00282.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pp. 1429–1434, 2017. doi: 10.1109/ICDAR.2017.233.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: A dataset for image captioning with reading comprehension. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 742–758, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58536-5.

Aaditya Singh et al. Openai gpt-5 system card, 2026. URL https://arxiv.org/abs/2601.03267.

Kaiser Sun, Xiaochuang Yuan, Hongjun Liu, Chen Zhao, Cheng Zhang, Mark Dredze, and Fan Bai. Reading, not thinking: Understanding and bridging the modality gap when text becomes pixels in multimodal llms, 2026. URL https://arxiv.org/abs/2603.09095.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. Journeydb: A benchmark for generative image understanding. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/9bc59aff4685e39e1a8175d5303248a1-Abstract-Datasets_and_Benchmarks.html.

TikTok. Community guidelines: Regulated commercial activities and regulated goods. https://www.tiktok.com/community-guidelines/en/regulated-commercial-activities, 2026.

C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pp. 839–846, 1998. doi: 10.1109/ICCV.1998.710815.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, 2009. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2009.06.042. URL https://www.sciencedirect.com/science/article/pii/S0893608009001130. Advances in Neural Networks Research: IJCNN2009.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Chenfei Wu et al. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Mengyang Wu, Yuzhi Zhao, Jialun Cao, Mingjie Xu, Zhongming Jiang, Xuehui Wang, Qinbin Li, Guangneng Hu, Shengchao Qin, and Chi-Wing Fu. Icm-assistant: Instruction-tuning multimodal large language models for rule-based explainable image content moderation, 2025b. URL https://arxiv.org/abs/2412.18216.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094, June 2024.

X Corp. Rules: Illegal or certain regulated goods and services. https://help.x.com/en/rules-and-policies/regulated-goods-services, 2026.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1eBeyHFDH.

Zhengzhuo Xu, Chong Sun, SiNan Du, Chen Li, Jing Lyu, and Chun Yuan. Vacot: Rethinking visual data augmentation with vlms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9780–9790, June 2026.

Qiying Yu et al. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation, 2026. URL https://arxiv.org/abs/2605.18740.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DgaY5mDdmT.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, October 2023.

Ruiyi Zhang, Yanzhe Zhang, Jian Chen, Yufan Zhou, Jiuxiang Gu, Changyou Chen, and Tong Sun. Trins: Towards multimodal language models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22584–22594, June 2024.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Jpxfof0EaS.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16(12):3848–3860, August 2023. ISSN 2150-8097. doi: 10.14778/3611540.3611569. URL https://doi.org/10.14778/3611540.3611569.

A Additional Diagnosis Controls

A.1 Per-layer localization

Figure 5 expands Figure 2a across encoder depth and script. The ordering is stable: restricting the probe to glyph-bearing tokens consistently improves glyph-identity separability, while RANDOM closely tracks GLOBAL. Both scripts peak in the middle of the encoder, but the localization gap persists across depth.

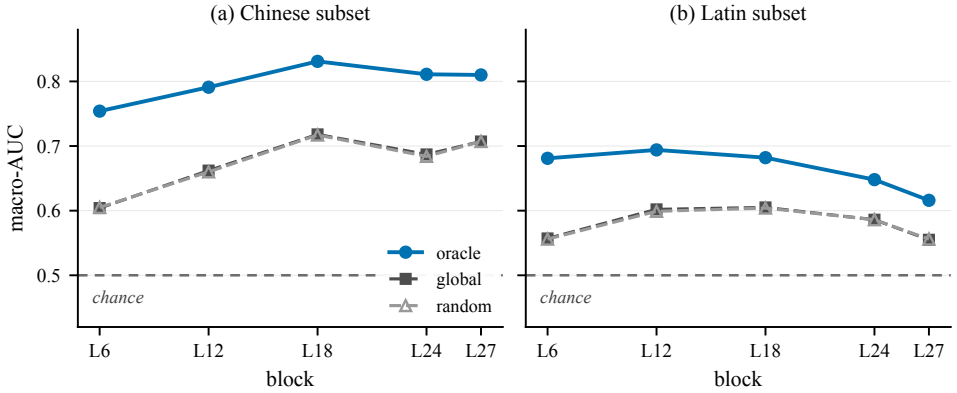


Figure 5: **Glyph identity is more accessible at the correct locations.** Glyph-identity macro-AUC on SQUINTCTRL-GLYPH by encoder block under the three poolings of §3.2. RANDOM uses the same number of tokens as ORACLE but remains close to GLOBAL across depth and script.

A.2 Patch-merger control

The probes in §3.1 read pre-merger features, whereas the language model consumes features after a 2×2 patch merger. Table 5 shows that global glyph separability changes by at most 0.018 AUC across the merger. The localization gap therefore cannot be explained by information being discarded at this architectural step.

Table 5: Glyph separability before and after the 2×2 patch merger.

script	checkpoint	before	after	Δ
Chinese	BASE	0.707	0.700	−0.007
Chinese	SQUINT OPSD	0.989	0.985	−0.004
Latin	BASE	0.555	0.573	+0.018
Latin	SQUINT OPSD	0.699	0.713	+0.014

A.3 Presence-probe control

The text-presence probe is sensitive to generation shortcuts if text-present and text-free samples follow different synthesis paths. Bypassing the ControlNet branch for text-free samples inflates AUC by roughly 0.03–0.05. All results in the paper therefore use matched generation paths, with an all-black conditioning image replacing the glyph pattern in the text-free member. Shuffled-label controls remain at 0.47–0.52 AUC.

A.4 Glyph-identity probe after training

For the before/after comparison in §6, we evaluate both checkpoints at L18, the strongest base layer shared by the two scripts. Training raises Chinese macro-AUC from 0.718 to 0.975 and Latin macro-AUC from 0.612 to 0.779, while presence detection was already near ceiling.

Table 6: Presence-probe AUC on SQUINTCTRL-PAIR. The controlled columns use the same generation path for both members of each pair.

	same pipeline		different pipelines	
block	BASE	SQUINT OPSD	BASE	SQUINT OPSD
L6	0.964	0.977	0.996	0.996
L12	0.977	0.995	0.995	0.998
L18	0.981	0.995	0.993	0.996
L24	0.967	0.991	0.991	0.995
L27	0.956	0.991	0.984	0.996

Table 7: Glyph-identity macro-AUC at L18 before and after training.

script	BASE	SQUINT OPSD	Δ
Chinese	0.718	0.975	$+0.257$
Latin	0.612	0.779	$+0.167$

B FREQUENCY AND PREPROCESSING CONTROLS

B.1 EXTERNAL-BENCHMARK AND PREPROCESSING NOTES

HC-Bench contains 56 hidden-text images. Two readers, evaluated independently with squinting permitted, recovered the hidden text in 35.7% of them on average. We therefore treat HC-Bench as an external transfer benchmark rather than a human-legibility ceiling.

For the reduced-view rows in Table 3, the model preprocessor enforces a minimum input area of 65536 pixels. Both the $1/8$ and 40×40 conditions therefore reach the encoder at approximately 256^2 pixels and 64 visual tokens. The labels in Table 3 describe the image before this preprocessing step.

The 64-token condition of §3.3 is *not* the same manipulation as these rows, even though both end at 64 visual tokens. Table 3 replaces the whole image with a reduced view and then lets the preprocessor resize it, whereas §3.3 pools the original view down to 64 tokens directly; the two therefore differ in spectrum and in score (29.2 versus 38.22 for the base model). Neither number should be read as a replication of the other.

B.2 SPECTRAL AUDIT OF THE REDUCED VIEWS

Because the two reduced conditions are matched in token count, their difference is primarily spectral. Table 8 shows that the 40×40 condition retains a larger fraction of energy in the lowest band and less energy in the higher bands than the $1/8$ condition. This explains why nominal scale alone is an incomplete description of the corrective view.

Table 8: Share of non-DC spectral energy by band, in cycles per pixel, for the three SQUINTBENCH viewing conditions.

band (c/p)	orig.	$1/8$	40×40
0–$1/32$	91.3%	94.3%	**96.5%**
$1/32$–$1/16$	6.5%	4.8%	3.3%
$1/16$–$1/8$	1.7%	0.87%	0.21%
$> 1/8$	0.5%	0.08%	0.03%

B.3 FREQUENCY AT CONSTANT TOKEN COUNT

A Gaussian low-pass applied at full resolution leaves the visual-token count unchanged. Every measured σ improves over the unfiltered 26.44 ANLS baseline, by 5.2–14.4 points (Figure 6). The curve is not monotonic in σ, so we use it only to establish that suppressing high-frequency content helps independently of sequence length.

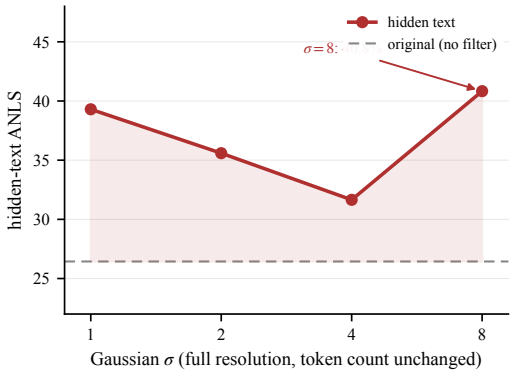


Figure 6: **Frequency at constant token count.** Hidden-text reading ANLS on all 1076 SQUINTBENCH images under Gaussian low-pass filtering. The dashed line is the unfiltered input; the visual-token count is unchanged throughout the sweep.

B.4 CONTRAST CONTROLS

Contrast-only manipulation does not reproduce the frequency effect. CLAHE (Pizer et al., 1987) leaves the spectrum approximately intact and yields essentially no gain, whereas operations that also flatten texture, such as bilateral filtering (Tomasi & Manduchi, 1998), improve more (Table 9). We therefore treat contrast as an unlikely primary explanation for the corrective-view benefit.

Table 9: Contrast and band-selective controls on SQUINTBENCH, against the unfiltered 26.44 ANLS baseline.

transform	what it does	ANLS	Δ
CLAHE	raises contrast, spectrum intact	25.91	−0.53
bilateral filter	flattens texture, keeps glyph edges	32.22	+5.79
posterise (3 levels)	compresses range and texture	37.21	+10.77
band-stop	removes mid, keeps low + high	11.22	−15.22

C TEACHER AND DISTILLATION CONTROLS

All reported default SQUINT OPSD results use a fixed teacher copied from the initial backbone. The teacher-input ablation in Table 4 separates the privileged view from the ground-truth hint. The view alone reaches 54.68, above the hint-only condition at 41.79; combining both reaches 88.61. A stronger control replaces the teacher's second view with the $1/8$ downscaled view of another sample while keeping the original image, hint, resolution, and visual-token budget unchanged. Under the paired setting, the matched view reaches 84.60 and the mismatched view 78.60; the gap remains through the third epoch. This isolates the benefit of sample-aligned corrective visual context.

The target-form ablation provides a separate control. Replacing the teacher's top-K distribution with its one-hot prediction reduces performance from 88.61 to 68.31, below ground-truth SFT at 72.47. Thus, the teacher is not simply a better source of hard labels; the distribution itself carries useful supervision.

C.1 CORRECTIVE-VIEW AND TRAINING VARIANTS

In all experiments, the teacher receives the same original image x, task instruction q, and ground-truth hint h; we vary only the corrective view v (Table 10). Replacing the default $1/8$ view with a fixed 40×40 view retains nearly identical performance (88.09 vs. 88.61), while $v = 1/4$ reaches 78.51. We also tested whether complementary scales could be combined by averaging or selecting among teacher distributions from $v \in \{1/8, 1/4, 40\times40\}$; none improves over the single default teacher, with probability averaging reaching 83.05.

Beyond the view itself, updating the teacher by EMA (rate 0.05) reaches 79.12. SFT warm starts reach 82.85 after one epoch and 81.35 after two epochs, both below direct SQUINT OPSD at 88.61. One plausible explanation is that the supervised stage first concentrates probability on a one-hot

Table 10: Teacher-view variants. All SQUINT OPSD teachers receive (x, v, q, h); only the corrective view v or its aggregation is varied. Student inference always uses (x, q) only.

teacher configuration	corrective view v	SQUINTBENCH
default	$1/8$	**88.61**
fixed view	40×40	88.09
fixed view	$1/4$	78.51
probability mean	$1/8,\ 1/4,\ 40\times40$	83.05
max-entropy selection	$1/8,\ 1/4,\ 40\times40$	80.51
min-entropy selection	$1/8,\ 1/4,\ 40\times40$	74.30

target, reducing the uncertainty over visually plausible alternatives that SQUINT OPSD is designed to transfer. Starting with distributional supervision preserves this structure throughout training.

We initially hypothesized that corrective views at different scales might provide complementary teacher signals. We therefore combine the $1/8$, $1/4$, and 40×40 teachers using probability averaging, minimum-entropy selection, and maximum-entropy selection. None exceeds the best single fixed teacher: probability averaging reaches 83.05, while maximum- and minimum-entropy selection reach 80.51 and 74.30, respectively. We think there are two reasons for this. First, the $1/8$ and 40×40 teachers are individually strong (88.61, 88.09) while the $1/4$ teacher is much weaker (78.51); averaging their distributions therefore lets the weak $1/4$ teacher drag down the two strong ones, instead of complementing them. Second, min-entropy selection (74.30) does notably worse than max-entropy selection (80.51): picking the most confident teacher discards exactly the uncertainty that a distributional target is meant to preserve — the spread over visually confusable glyphs — whereas retaining more uncertainty keeps more of the alternative candidates alive for the student to learn from.

D RECOVERABILITY BOUNDARY

SQUINTCTRL-STRENGTH varies only the ControlNet conditioning strength over the same 200 scenes. Both BASE and SQUINT OPSD improve as the glyph signal becomes more visible, but the largest useful margin appears in the partially recoverable regime. At strength 0.5, where strokes are largely destroyed, BASE and SQUINT OPSD are statistically indistinguishable (0.597 vs. 0.615, 95% CI $[-0.008, +0.063]$; paired bootstrap, 100 resamples). At strength 0.9, SQUINT OPSD improves from 0.886 to 0.944; at 1.2, both models are near ceiling. This empirically delineates the recoverability boundary of the method: training can reorganize partially recoverable evidence, but does not reconstruct glyph structure that is largely absent from the input.

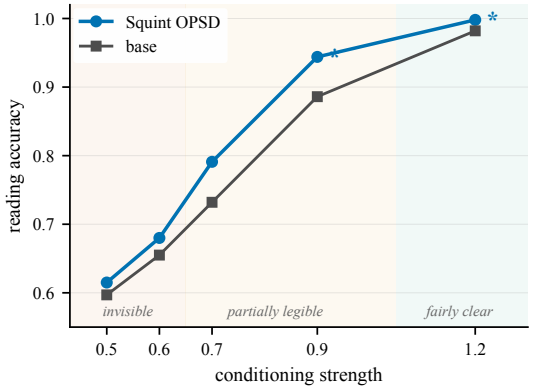


Figure 7: **Recoverability versus hiding strength.** Reading accuracy for the base model and SQUINT OPSD as conditioning strength varies, over the same 200 scenes. Stars mark strength levels at which the 95% CI of the paired base-vs-SQUINT OPSD difference excludes zero (paired bootstrap, 100 resamples); unmarked levels are not significant at this threshold, and 0.7 is borderline and is left unmarked. The method provides its clearest benefit in the partially recoverable regime; the two ends saturate for different reasons.

E Analysis Details

E.1 Candidate-distribution metrics

For §6, we analyze 200 SQUINTCTRL-READ images. We retain examples whose ground-truth characters are each represented by a single token, yielding 367 aligned glyph positions before candidate filtering. At each position, we run the model under teacher forcing and take its top-20 next-token candidates. We remove the ground-truth token, retain only tokens that decode to exactly one CJK codepoint, and keep positions with at least five valid candidates. Thus, the analysis concerns non-target high-probability alternatives rather than the full vocabulary tail.

For ρ_{vis}, each candidate and target glyph is rendered with Noto Sans CJK SC, centered by its ink bounding box, dilated with a 3×3 max filter, resized to 64×64 with Lanczos resampling (Lanczos, 1964), and binarized. Visual similarity is pixel IoU. We compute a per-position Spearman correlation between candidate probability and glyph IoU, subtract the mean correlation from 20 random permutations of the similarity scores, and average over valid positions.

For ρ_{lang}, we evaluate the same checkpoint without the image, using the identical OCR prompt and teacher-forced ground-truth prefix. We then compute the per-position Spearman correlation between the image-conditioned candidate probabilities and their probabilities under this matched text-only context. Multi-token ground-truth characters are excluded from the analysis.

E.2 Matched-position robustness

The default filtering retains different numbers of valid positions for the three checkpoints because their top-20 distributions have different sharpness: 254 for BASE, 259 for SFT, and 164 for SQUINT OPSD. To remove this selection effect, the main-text comparison uses the intersection of valid positions across all three models ($n = 120$). Table 11 shows that the ordering is unchanged on this matched set, so the main-text comparison is not an artefact of position selection. The same conclusion holds when intersecting all five analyzed conditions ($n = 106$), where ρ_{vis} is 0.117/0.247/0.298 for BASE/SFT/SQUINT OPSD, respectively.

Table 11: Candidate-distribution structure before and after matching glyph positions across checkpoints. Parentheses give the number of valid positions.

metric	BASE	SFT	SQUINT OPSD
ρ_{vis}, own valid set	0.079 (254)	0.160 (259)	0.243 (164)
ρ_{vis}, common set	0.126	0.242	**0.286**
ρ_{lang}, common set	0.555	0.334	0.544

E.3 Candidate-set sensitivity

Because a fixed top-K cutoff interacts with distribution sharpness, we sweep $K \in \{10, 12, 15, 20, 25, 30, 40, 50\}$ and recompute the matched-position comparison at each cutoff. The ordering SQUINT OPSD $>$ SFT $>$ BASE in ρ_{vis} is stable for $K = 10$–40. At $K = 50$, SFT exceeds SQUINT OPSD (0.170 vs. 0.144; BASE=0.098). This sensitivity is consistent with the substantially sharper SQUINT OPSD distribution: its top-20 contains fewer valid single-CJK alternatives on average (4.6, versus 8.4 for BASE and 7.8 for SFT), while its target confidence is substantially higher. Expanding to a large fixed K therefore adds a larger low-probability tail relative to its effective support. We accordingly scope the main claim to the organization of high-probability candidate alternatives rather than the entire vocabulary tail.

E.4 Attention-mass definition

For each answer glyph position $t \in \mathcal{A}$, we renormalize attention over the visual tokens $\mathcal{V}$ and measure the mass falling on the glyph-bearing set $\mathcal{T}$, divided by its chance level:

$$\text{AMR}_{\text{norm}} = \frac{1}{|\mathcal{A}|} \sum_{t\in\mathcal{A}} \sum_{j\in\mathcal{T}} \tilde{a}_{t,j} \Big/ \frac{|\mathcal{T}|}{|\mathcal{V}|}, \qquad \tilde{a}_{t,j} = \frac{a_{t,j}}{\sum_{j'\in\mathcal{V}} a_{t,j'}}. \tag{3}$$

Here $\mathcal{A}$ includes only answer positions corresponding to transcribed glyphs, excluding formatting tokens. The known glyph mask is downsampled through the same 2×2 merger as the visual features before attention is measured. Thus $\text{AMR}_{\text{norm}} = 1$ denotes chance-level allocation within the visual tokens.

E.5 VISUAL-TOKEN EMBEDDING KNOCKOUT

To test whether reading causally depends on glyph-bearing visual regions, we intervene on the output embeddings of the vision tower before they are passed to the language model. The intervention is applied once during prefill; no language-model attention logits or weights are modified.

Let T denote visual tokens whose receptive fields overlap the known glyph pattern. For the glyph knockout, each embedding in T is replaced by the mean embedding of the non-glyph tokens from the same image. This preserves sequence length, token positions, and the subsequent attention computation. As a matched control, we sample the same number of tokens uniformly from the non-glyph region using a fixed random seed and apply the identical mean-embedding replacement.

We additionally dilate the glyph mask by $d \in \{0, 1, 2\}$ iterations of a 3×3 neighborhood operation. These interventions replace on average 169.0 (16.5%), 299.9 (29.3%), and 391.8 (38.3%) of the 1024 visual tokens, respectively. Generation is greedy with a maximum of 64 new tokens, and a prediction is counted as correct when the ground-truth string occurs in the generated transcription after whitespace removal.

Table 12 shows that the effect is strongly spatially specific. At $d = 2$, SQUINT OPSD decreases from 90.5% to 71.5%, while the matched random intervention changes accuracy only from 90.5% to 91.0%. BASE also degrades under glyph removal (26.0% to 15.0%), but begins from substantially lower reading accuracy.

Table 12: Reading accuracy after visual-token embedding replacement. Glyph-region embeddings are replaced by the mean embedding of non-glyph tokens from the same image; the random control replaces an equal number of randomly selected background-token embeddings using the same operation. d denotes the number of 3×3 dilations applied to the known glyph mask.

intervention	BASE	SQUINT OPSD
none (baseline)	26.0	90.5
replace glyph tokens, d=0	19.5	87.0
replace glyph tokens, d=1	16.0	77.0
replace glyph tokens, d=2	15.0	71.5
replace random tokens, d=0	25.0	91.5
replace random tokens, d=1	25.0	91.0
replace random tokens, d=2	24.5	91.0

F TRAINING AND REPRODUCIBILITY

F.1 TRAINING MIXTURE

Table 13: Training mixture, 5905 prompts after removing 65 samples with corrupted annotations. DocVQA contributes train-split samples only; its reported evaluation uses a disjoint test split.

group	sources	prompts
hidden text (synthetic)	AIGC collection 1143, self-synthesised ControlNet 1144, FakeTextDetect 1267, minus 65 removed	3489
open-source OCR	TextCaps 564, DocVQA 302, CASIA 150, RCTW 136, POIE 56, FUNSD 3, ICDAR13 2	1213
VQA	OCR-VQA	1203
total		5905

Dataset references: TextCaps (Sidorov et al., 2020), DocVQA (Mathew et al., 2021), CASIA (Liu et al., 2011), RCTW (Shi et al., 2017), FUNSD (Jaume et al., 2019), ICDAR13 (Karatzas et al., 2013), and OCR-VQA (Mishra et al., 2019).

F.2 HELD-OUT VALIDATION MIXTURE

We reserve a held-out split of 1045 examples drawn from the same source mixture as the training data: 616 hidden-text examples, 269 open-source OCR examples, and 160 region-conditioned OCR/VQA examples. The split is separate from the 5905 training prompts and is intended for model-selection and validation under the same task mixture as training.

F.3 Optimization and rollout settings

The default SQUINT OPSD run initializes the student from Qwen3.5-9B (Qwen Team, 2026) and uses a fixed teacher copied from the same initial backbone. All student parameters, including the vision tower, are updated. Training is configured for up to three epochs, with checkpoints evaluated every 10 optimization steps.

Table 14: Main SQUINT OPSD optimization and rollout settings.

hyperparameter	value
optimizer / learning rate	AdamW (Loshchilov & Hutter, 2019), 1×10^{-5}
schedule / warmup	constant / 10 steps
weight decay / grad clip	0.01 / 1.0
prompt batch	56 prompts per step
rollouts per prompt	n=8 (448 sequences/step)
inner epochs	1
maximum training horizon	3 epochs
teacher	fixed initial-backbone copy
trainable parameters	full model, including vision tower
distillation support	top-K=100 with tail bucket
JSD	symmetric, $\alpha = 0.5$
max prompt / response length	8000 / 1024 tokens
rollout sampling	T=1.0, top-p=1.0
validation decoding	greedy (T=0, n=1)
precision / training stack	bf16, FSDP (Zhao et al., 2023), gradient checkpointing
checkpoint evaluation	every 10 steps

Student rollouts are sampled from the current policy and treated as stop-gradient trajectories. At each sampled prefix, the teacher and student distributions are evaluated on the teacher's top-K support plus one tail bucket containing the remaining probability mass. Generation terminates at EOS or the maximum response length, and gradients are propagated only through the student logits.

Supervised and RL baselines. SFT updates the full model for 3 epochs with learning rate 1×10^{-5}, cosine decay, warmup ratio 0.03, effective batch size 64, and maximum sequence length 2048. DAPO (Yu et al., 2025), which extends the clipped policy-gradient objective of GRPO (Shao et al., 2024), is also trained for up to 3 epochs. It uses learning rate 1×10^{-5}, rollout n=8, clipping $(\epsilon_{\text{low}}, \epsilon_{\text{high}}) = (0.2, 0.28)$, and ANLS reward with threshold $\tau = 0.5$.

F.4 Visual adaptation under supervised fine-tuning

To test whether the importance of visual adaptation is specific to the OPSD objective, we repeat supervised fine-tuning while changing which modules are trainable. With the vision tower frozen, SFT reaches only 27.29 on SQUINTBENCH; updating both the vision tower and language model raises this to 72.47. The same qualitative dependence therefore appears under a standard supervised objective.

Table 15: Effect of visual adaptation under three-epoch supervised fine-tuning.

trainable modules	SQUINTBENCH	AI2D	DocVQA	OCRBench
LM only; ViT frozen	27.29	86.66	50.39	741
ViT + LM	72.47	85.52	64.73	722

F.5 Training prompts

The student receives the original image and the task instruction. The teacher receives the original image together with its $1/8$ downscaled view, followed by the same task instruction and output format; the ground-truth output is provided only as privileged teacher context.

Student template.

```
<image>
{task instruction}
{task-specific format instruction}
```

Teacher template.

```
<image>
<image>
第一张图为原图,第二张图为同一张图的 1/8 下采样视图。
{task instruction}
{task-specific format instruction}
该图的正确输出为{gt}。
```

For OCR-style samples, the output format is a JSON list of {"text_content": "..."} objects, with [] used when no text is present. OCR-VQA and detection samples use the same schema with task-specific wording.

F.6 SQUINTCTRL SYNTHESIS

The main synthetic branch uses Stable Diffusion 1.5 with QR Code Monster-style conditioning. Images are generated at 512 pixels and upscaled to 1024, with ControlNet scale 1.2, two-stage CFG 5.0/7.5, and img2img strength 0.5. We use seed 42 and generate 10 samples per prompt. The conditioning image is retained to derive the exact glyph mask used by the controlled diagnostic experiments.

G QUALITATIVE DEPLOYMENT EXAMPLES

Figure 9 shows four production-derived SQUINTBENCH examples. Given the transcription instruction, two frontier commercial MLLMs report that no text is present, while SQUINT OPSD, still reading only the original image, recovers the embedded promotional strings; a human viewer can read the same strings when the page is viewed at reduced scale.



Figure 8: Two examples from each of the four SQUINTCTRL subsets. The embedded strings are carried in the images' coarse structure: they are hard to notice at full size but become legible when the page is squinted at or viewed at reduced scale. SQUINTCTRL-PAIR: a text-present image and its text-free counterpart, regenerated through the same ControlNet branch with an all-black conditioning image. SQUINTCTRL-GLYPH: conditioning on the closed vocabularies of ten Chinese glyphs and ten Latin letters. SQUINTCTRL-STRENGTH: the same type of content under weak versus strong conditioning. SQUINTCTRL-READ: arbitrary open-vocabulary strings.



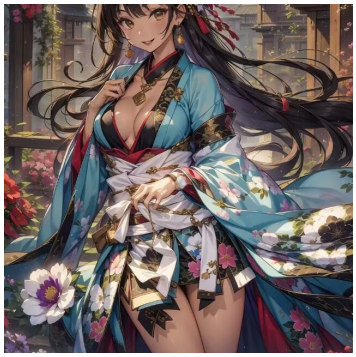
User	What text is written in this image?
gemini-3.1-pro	“There is no text in this image.”
claude-opus-5	“There is no text in this image.”
gpt-6-astra	“There is no text in this image.”
SQUINT OPSD	浏览器搜78wm.cc



User	What text is written in this image?
gemini-3.1-pro	“There is no text in this image.”
claude-opus-5	“There is no text in this image.”
gpt-6-astra	“There is no text in this image.”
SQUINT OPSD	浏览器搜58hx.cc



User	What text is written in this image?
gemini-3.1-pro	“There is no text in this image.”
claude-opus-5	“There is no text in this image.”
gpt-6-astra	“There is no text in this image.”
SQUINT OPSD	Q群198943973



User	What text is written in this image?
gemini-3.1-pro	“There is no text in this image.”
claude-opus-5	“There is no text in this image.”
gpt-6-astra	“There is no text in this image.”
SQUINT OPSD	搜索ov91.cc

Figure 9: **SQUINTBENCH deployment examples that commercial models miss.** Four production-derived moderation images, shown together with the answers of two frontier commercial MLLMs and SQUINT OPSD to the transcription instruction. Both commercial models report that no text is present on the original image, whereas SQUINT OPSD reads out the embedded promotional strings; the same strings become legible to a human when the page is squinted at or viewed at reduced scale. These examples come from the production-derived portion of SQUINTBENCH and contain no personal data.