TOKENIZING THE UNSEEN: COMPACT VISUAL PREFIX CACHE FOR OVERCOMING VLM INPUT BANDWIDTH

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained visual understanding faces an input-bandwidth tradeoff: low-resolution inputs discard small details such as text rendered at a small size, while high-resolution inputs occupy hundreds of decoder positions, a burden that compounds when the same image must answer many questions. Existing approaches each pay a different cost: agentic methods search, zoom, or crop for every query, so latency grows with the number of questions; raising the input resolution inflates the visual sequence; and token-reduction methods lose much of their accuracy under aggressive compression. We propose a compact visual prefix that decouples a high-resolution branch from the visual input budget of a frozen language decoder. This branch reads the image at full resolution and compresses it into the prefix, so the decoder needs only a fixed number of visual positions, while the pretrained backbone and its native path are left unchanged. Computed from the image alone, the prefix can be cached once and reused across questions. Under a budget of 24 decoder positions, the learned prefix reaches 0.386 TextOCR word F1 against 0.154 for the strongest of five tested token reducers. A separately trained image-only adapter transfers without further training to an unseen scene-text set, raising ICDAR15 F1 from 0.134 to 0.424. On DocVQA, a single prefix per page serves all 4.68 questions on average. These results show that a compact visual prefix can extract and compress useful information from a high-resolution image, and that the result is reusable across questions.

1 INTRODUCTION

Many applications powered by vision–language models (VLMs) require extracting fine-grained information from images and documents (Singh et al., 2021; Mathew et al., 2021). Answering a document question may mean reading a small field on a visually dense page, or recovering a brand name or number that occupies only a small portion of a scene image. Preserving such evidence requires high-resolution input, which produces hundreds or even thousands of visual tokens in the language decoder's input sequence (Bai et al., 2025). The burden compounds when the same image is queried repeatedly, as it is in document question answering: key–value caching avoids recomputation but leaves the visual context occupying the same decoder positions for every query (Bai et al., 2025; Mu et al., 2023).

Existing work falls into three categories. Agentic approaches use visual search, zooming, or cropping to inspect fine-grained details, but re-encode each crop, so their cost grows with the number of questions (Wu & Xie, 2024a; Shen et al., 2025; Wang et al., 2025). Token-reduction methods compact the visual context through selection, pruning, or merging, but a fixed importance criterion does not know what the decoder needs, and aggressive compression discards fine-grained evidence (Bolya et al., 2023; Yang et al., 2025; Zhang et al., 2025a; Ma et al., 2026). Fine-tuning the backbone, a third route, modifies the pretrained weights for small and inconsistent gains on these tasks (Hu et al., 2022a). This raises a natural question: can we obtain *a compact visual representation* that (1) *retains fine-grained evidence* from high-resolution inputs, (2) can be *cached and reused across questions* about the same image, and (3) is learned *without fine-tuning the backbone*?

In this work, we introduce a *compact visual prefix*, a learned representation that decouples a high-resolution visual branch from the visual input budget of a frozen language decoder. Three choices follow.

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.

First, *decouple and keep the resolution*. The decoder's native path sees a low-resolution view; we add a branch that reads the image at full resolution with a frozen vision encoder. The resolution at which the image is encoded is thus separated from the number of positions the decoder receives, and this branch preserves the fine detail that downsampling would have discarded.

Second, *compress into a few reusable tokens*. A learned-query adapter attends to the dense high-resolution features and compresses them into T_{pre} soft prefix tokens, refining the prefix and the visual features alternately before readout. Because the prefix is computed from the image alone, without the query, and placed before all query-dependent inputs in the causal decoder, its key–value (KV) states are identical across questions, and one cache per image serves every question asked about it.

Third, *train only the adapter*. The adapter is optimized through answer supervision while both the vision encoder and the language decoder remain frozen, so the pretrained model is not modified and its general capabilities are not placed at risk.

At decoder visual budgets from 6 to 128 positions, a 24-token pure prefix reaches 0.386 TextOCR word F1, against 0.154 for the strongest of five tested training-free native reducers and 0.125 for a same-supervision projector. Reporting these scores alongside visual KV size characterizes the quality available at each storage budget. A separately trained prefix-first model then improves over low-resolution references on four question-answering and recognition tasks, and reuses each page's prefix to answer its associated DocVQA questions.

Our contributions are: **(1)** We introduce a compact visual prefix, a learned representation that decouples a high-resolution visual branch from the visual input budget of a frozen language decoder. A learned-query adapter compresses dense high-resolution features into a few reusable prefix tokens, while both the vision encoder and the language decoder remain frozen. **(2)** We formulate the compression interface of a frozen decoder explicitly—what must be preserved for the answer, and what the decoder must be able to read—and give sufficient conditions under which an image-only prefix preceding the question has question-independent decoder KV and admits exact reuse. **(3)** We characterize this interface empirically under extreme visual budgets, through exact decoder-position accounting, comparisons with five training-free token reducers and a same-supervision projector, quality–KV comparisons, prefix-budget and image-content controls, matched-supervision QA experiments, and evaluations of cross-question reuse.

2 Related Work

Visual search and zooming. Visual search, zooming, and cropping select and re-encode query-relevant regions (Wu & Xie, 2024b; Shen et al., 2025; Wang et al., 2025). We study the complementary setting of compressing high-resolution evidence into a fixed decoder input budget before the question is known.

Learned visual bottlenecks. Perceiver IO, Flamingo, and BLIP-2 map dense inputs to compact learned representations (Jaegle et al., 2022; Alayrac et al., 2022; Li et al., 2023). LLaVA-Mini combines compression with instruction pre-fusion (Zhang et al., 2025b); Matryoshka Multimodal Models learn nested representations with variable granularity (Cai et al., 2025). We train each prefix length independently for image-only compression at small decoder input budgets, freezing both the vision encoder and language decoder.

Token selection and merging. ToMe merges tokens within a vision transformer (Bolya et al., 2023); VisionZip, ApET, and VScan select or aggregate visual features (Yang et al., 2025; Ma et al., 2026; Zhang et al., 2025a). FastV prunes visual tokens within the decoder after initial layers process the full visual sequence (Chen et al., 2024). We compare at matched decoder input budgets; Appendix C details implementation adaptations and FastV's layerwise accounting.

Soft context and reusable memory. Prefix-tuning learns task-specific continuous parameters (Li & Liang, 2021); ICAE and gist compress context into compact states (Ge et al., 2024; Mu et al., 2023). VoCo-LLaMA uses visual compression tokens and attention distillation within the language model (Ye et al., 2024); VisCo uses the VLM as a visual autoencoder (Zheng et al., 2026). Our image-only embeddings are reusable across questions; Proposition 1 gives sufficient conditions for reusing their decoder KV when they precede all question-dependent inputs.

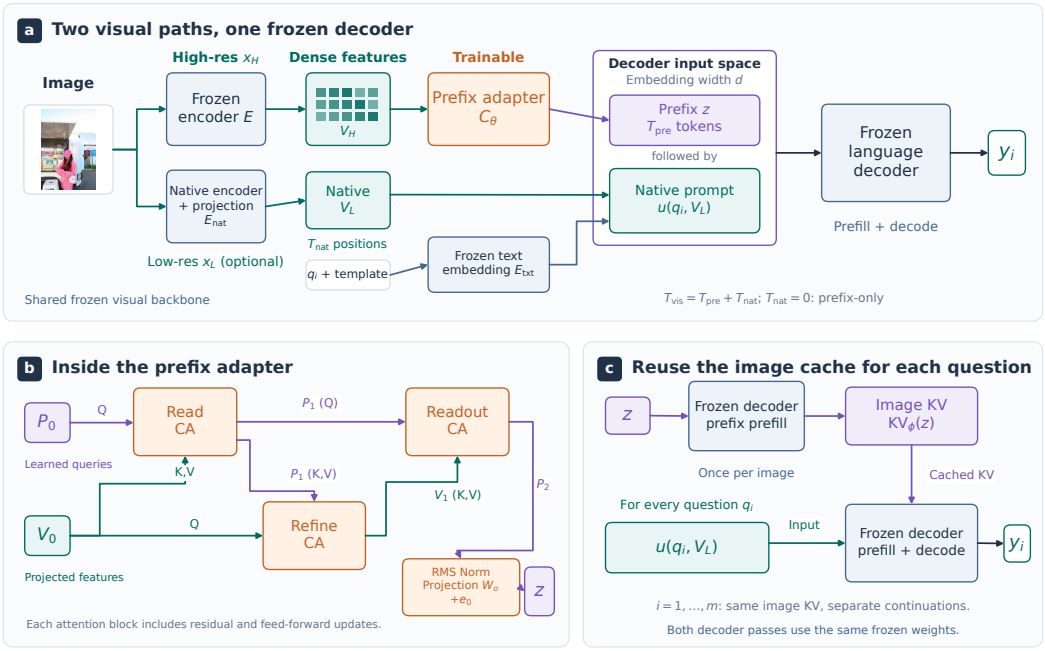

Figure 1: **High-resolution visual memory with an optional native view.** (a) The trainable, image-only adapter compresses high-resolution features from the frozen backbone into T_{pre} continuous prefix embeddings z. The optional native path supplies T_{nat} thumbnail embeddings V_L. Frozen text embeddings and V_L form $u(q_i, V_L)$ in template order. The frozen decoder receives the width-d sequence $[z; u(q_i, V_L)]$ and generates y_i. The visual budget is $T_{\text{vis}} = T_{\text{pre}} + T_{\text{nat}}$; $T_{\text{nat}} = 0$ gives the pure-prefix configuration. (b) The adapter alternates query read, visual refinement, and query readout. RMS Norm, output projection, and addition of the optional fixed embedding e_0 produce z. CA is defined in Equation 3; Q and K,V denote its query and key/value inputs. (c) Prefilling z with frozen decoder weights ϕ constructs $\text{KV}_\phi(z)$ once per image, shared by m questions with separate prompt and answer continuations. Native auxiliary visual injections are omitted; the feature grid is schematic.

3 Compact Visual Prefixes for Frozen VLMs

Our trainable, image-only adapter compresses high-resolution features from a frozen vision encoder into a compact sequence of continuous prefix embeddings for a frozen language decoder (Figure 1). An optional low-resolution path uses the same backbone to supply native visual input. We train the adapter through the decoder's answer loss to encode visual evidence in a form the frozen decoder can use. When the prefix precedes all question-dependent inputs, its decoder KV states can be computed once per image and reused across questions under the conditions in Proposition 1.

3.1 Dual-Path Visual Encoding

As shown in Figure 1(a), the frozen visual encoder E maps the high-resolution image x_H to $V_H = E(x_H) \in \mathbb{R}^{T_{\text{enc}} \times d_v}$, where T_{enc} is the number of visual feature vectors and d_v is their dimension. A trainable projection $W_v \in \mathbb{R}^{d \times d_v}$ maps each feature vector to the adapter width d, which equals the decoder embedding dimension: $V_0 = V_H W_v^\top \in \mathbb{R}^{T_{\text{enc}} \times d}$. The feature count T_{enc} depends on the image resolution, whereas the prefix length T_{pre} is chosen independently.

The prefix can supply all visual input or accompany a native low-resolution thumbnail x_L through the optional branch in Figure 1(a). Let $V_L = E_{\text{nat}}(x_L) \in \mathbb{R}^{T_{\text{nat}} \times d}$ denote its native decoder-input features, including the frozen native projection. Here x_L denotes image pixels and V_L denotes feature embeddings; both visual paths use the same frozen backbone at different resolutions. If the thumbnail occupies T_{nat} native visual positions, the total decoder visual budget is

$$T_{\text{vis}} = T_{\text{pre}} + T_{\text{nat}}. \tag{1}$$

In the pure-prefix configuration, $T_{\text{nat}} = 0$: all image information reaches the decoder through z. We remove native image placeholders and bypass the native visual and DeepStack injection paths in this case. For the primary hybrid experiments, aspect-preserving resizing and padding give exactly $T_{\text{nat}} = 12$ thumbnail positions per image (Appendix A). Here T_{vis} measures visual positions, excluding text and template tokens; high-resolution encoding and adapter computation remain additional costs.

3.2 Constructing the Visual Prefix

We use learned queries to aggregate the dense visual representation into a fixed number of output slots, following the learned-bottleneck approach of prior work (Jaegle et al., 2022; Alayrac et al., 2022; Li et al., 2023). Let C_θ denote the prefix adapter with trainable parameters θ, including its input and output projections. The adapter starts from T_{pre} trainable vectors $P_0 \in \mathbb{R}^{T_{\text{pre}} \times d}$, shared across images. Their updated states depend on the image through cross-attention. Each slot can combine evidence from multiple visual locations, allowing the adapter to learn how the available decoder positions should be used.

$$z = C_\theta(V_H) \in \mathbb{R}^{T_{\text{pre}} \times d}. \tag{2}$$

Here z denotes continuous visual input embeddings; their per-layer keys and values are produced by prefilling the frozen decoder.

We use a common cross-attention block $\mathrm{CA}(X, Y)$, where X is the query-state matrix and Y supplies keys and values. Both have width d but may have different sequence lengths. A tilde denotes tokenwise RMS Norm of the indicated states. The block is

$$\begin{aligned} H &= X + \mathrm{MHA}(\widetilde{X}, \widetilde{Y}, \widetilde{Y}), \\ \mathrm{CA}(X, Y) &= H + \mathrm{FFN}(\widetilde{H}). \end{aligned} \tag{3}$$

Here H is the attention residual state; MHA is multi-head scaled dot-product attention, with query, key, and value inputs in that order, and FFN is a SwiGLU feed-forward network. Each RMS Norm, attention, and FFN module has its own parameters within the block. The three calls below use independent parameter sets; the shared function name denotes their common structure. The interaction in Figure 1(b) is

$$\begin{aligned} P_1 &= \mathrm{CA}(P_0, V_0), & V_1 &= \mathrm{CA}(V_0, P_1), \\ P_2 &= \mathrm{CA}(P_1, V_1), & z &= \widetilde{P}_2 W_o^\top + \mathbf{1}_{T_{\text{pre}}} e_0^\top. \end{aligned} \tag{4}$$

The states $P_1, P_2 \in \mathbb{R}^{T_{\text{pre}} \times d}$ are updated prefix queries, and $V_1 \in \mathbb{R}^{T_{\text{enc}} \times d}$ contains refined dense features. The final RMS Norm produces $\widetilde{P}_2$. The output projection $W_o \in \mathbb{R}^{d \times d}$ follows the same row-vector convention as W_v. The fixed base embedding is $e_0 \in \mathbb{R}^d$ when configured, and $e_0 = 0$ when disabled; $\mathbf{1}_{T_{\text{pre}}} \in \mathbb{R}^{T_{\text{pre}}}$ is the all-ones vector that broadcasts it to every prefix position. The queries first read the dense visual features, which then attend to the updated queries. A final query readout carries the refined visual evidence into z. All updates occur within the adapter, with the vision encoder frozen and neither the question nor its text embeddings as input.

3.3 Decoder Integration and KV Reuse

Let q be a question, y its answer token sequence, and ϕ the frozen decoder parameters. The frozen text embedding lookup E_{txt} maps the tokenized question and template to width-d embeddings. Inserting V_L at the template's native visual positions gives the prompt embedding sequence $u(q, V_L) \in \mathbb{R}^{(T_{\text{txt}} + T_{\text{nat}}) \times d}$, where T_{txt} counts the non-visual prompt positions, including question and template tokens, excluding positions occupied by V_L. Thus u and z are represented in the same decoder input embedding space. We place the prefix before this prompt:

$$s(q, V_L, z) = [z; u(q, V_L)] \in \mathbb{R}^{(T_{\text{pre}} + T_{\text{txt}} + T_{\text{nat}}) \times d}. \tag{5}$$

Here s denotes the full input embedding sequence and $[\,;\,]$ denotes concatenation along its position axis. The embedding lookup and native visual projection are existing frozen model components.

One-time image prefilling. After image encoding and prefix construction (Figure 1(a)), we run the prefix through the frozen decoder once (Figure 1(c)):

$$z = C_\theta(E(x_H)), \qquad \mathrm{KV}_\phi(z) = \left(K^{(\ell)}_{\mathrm{pre}}(z), V^{(\ell)}_{\mathrm{pre}}(z)\right)_{\ell=1}^{L_{\mathrm{dec}}}. \tag{6}$$

Here $K^{(\ell)}_{\mathrm{pre}}(z)$ and $V^{(\ell)}_{\mathrm{pre}}(z)$ are the key and value tensors produced at decoder layer ℓ by prefilling z under the fixed attention-mask and position conventions, and L_{dec} is the number of decoder layers. Thus, $\mathrm{KV}_\phi(z)$ denotes the complete prefix cache, with T_{pre} cached positions at each layer.

Per-question generation. For each question, we initialize a separate continuation from the same prefix cache $\mathrm{KV}_\phi(z)$ (Figure 1(c)). The decoder first prefills $u(q, V_L)$ while attending to this cache, then generates the answer autoregressively. Each generated token attends to the visual prefix, the prompt, and earlier answer tokens. Prompt positions continue from the prefix boundary using the same position conventions as full recomputation. Image encoding, compression, and prefix prefilling are shared across questions; prompt processing and answer decoding are performed for each question.

Proposition 1 (Exact reuse under causal attention)**.** *For a deterministic causal decoder with tokenwise auxiliary operations, the cache* $\mathrm{KV}_\phi(z)$ *produced from an image-only prefix placed before all question-dependent inputs is question-independent, provided the prefix's positional metadata does not depend on the suffix. With identical weights, masks, and position conventions, reusing this cache gives the same continuation distributions as full recomputation in exact arithmetic.*

The result follows by induction over decoder layers: every prefix state attends only to prefix states, so introducing a new question cannot change the cached image representation. This causal isolation is related to reusable context compression (Mu et al., 2023). Appendix J.4 provides the proof and implementation conditions. Section 5 evaluates answer quality and numerical agreement under finite-precision execution.

3.4 Training a Decoder-Readable Representation

We train the adapter through the answers that the frozen decoder produces from its prefix. For training examples (x_H, q, r, y), let $z = C_\theta(E(x_H))$, where r denotes the native side input, including native visual embeddings, auxiliary visual injections, and their positional metadata when present; r is empty for the pure-prefix model. Under the prefix-first layout in Equation 5, the objective is

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x_H,q,r,y)\sim\mathcal{D}} \sum_{t=1}^{|y|} \log p_\phi(y_t \mid q, r, z, y_{<t}) \,. \tag{7}$$

Here $\mathcal{D}$ is the training-example distribution and p_ϕ the frozen decoder's conditional token distribution. For answer length $|y|$, y_t denotes the tth token and $y_{<t}$ its preceding history. The expected answer negative log-likelihood $\mathcal{L}(\theta)$ includes only answer tokens. Gradients pass through the frozen decoder to update θ; both the vision encoder and decoder remain frozen. OCR supplies transcription targets; QA examples train the prefix to answer different questions about the same image. Training recipes and evaluated checkpoint configurations appear in Section 4.1 and Appendix A.

We initialize the output projection W_o to zero. When enabled, the fixed base embedding e_0 provides a nonzero starting representation (Appendix F); configurations without it use $e_0 = 0$. Zero initialization does not make inserted positions an exact no-op. We train each prefix length independently.

Information preservation and decoder compatibility. Answer supervision addresses two requirements: the prefix must retain relevant evidence, and the frozen decoder must be able to use it. For the population analysis, let X_H, Q, R, Y denote the random image, question, native side input, and answer, with realizations x_H, q, r, y, respectively. The native side input is defined in the training objective above. Let $V = E(X_H)$ and $Z = C_\theta(V)$ denote the random dense features and prefix, with realizations V_H and z, respectively. We group the question and native side input as $S = (Q, R)$. Below, C abbreviates the fixed compressor C_θ. Define the population loss $\mathcal{R}_\phi(C) = \mathbb{E}[-\log p_\phi(Y \mid S, Z)]$ and decoder mismatch

$$\Delta_\phi(C) = \mathbb{E}_{S,Z} D_{\mathrm{KL}}(P(Y \mid S, Z) \,\|\, p_\phi(Y \mid S, Z)) \,. \tag{8}$$

Here P is the data distribution, D_{KL} is Kullback–Leibler divergence, $\mathcal{R}_\phi$ is population sequence log-loss, and Δ_ϕ is the expected decoder mismatch. We use $H(\cdot \mid \cdot)$ for conditional entropy and $I(\cdot;\cdot \mid \cdot)$ for conditional mutual information, with natural logarithms.

Proposition 2 (Compression with a frozen decoder)**.** *For deterministic $Z = C_\theta(V)$ and finite expected log-loss,*

$$\mathcal{R}_\phi(C) = H(Y \mid S, V) + I(Y; V \mid S, Z) + \Delta_\phi(C). \tag{9}$$

This specialization of the log-loss decomposition (Courtade & Weissman, 2014) separates uncertainty remaining with the full visual features, answer information lost through compression, and mismatch between the retained information and the frozen decoder. It motivates learning the prefix through the decoder's answer loss and testing its dependence on the correct image. The identity concerns expected log-loss; accuracy under word F1 and ANLS is assessed empirically. Proofs and the finite-precision capacity analysis appear in Appendix J.

4 Experiments

We evaluate recognition under fixed visual budgets, transfer, and controls on budget allocation and image content. A QA supervision study then motivates the separately trained reusable prefixes in Section 5.

4.1 Experimental Setup

Models and training. We train image-only adapters on 21,749 TextOCR training images (Singh et al., 2021), with the Qwen3-VL-4B-Instruct vision encoder and language decoder frozen (Bai et al., 2025). Each primary configuration is trained independently for 10 epochs and evaluated at epoch 10. Pure prefixes use $T_{\mathrm{pre}} \in \{6, 12, 24, 64, 128\}$; hybrid models use $T_{\mathrm{pre}} \in \{6, 12, 24\}$ with $T_{\mathrm{nat}} = 12$ thumbnail positions. Backbone, transfer, and reuse studies use same single-block adapters; Appendix A gives training recipes, metric definitions, and cohort definitions.

Datasets and metrics. The primary evaluation uses 500 held-out TextOCR validation images (indices 501–1000); the preceding 500 are reserved for development. We report image-mean word F1 and test scene-text transfer on TotalText and ICDAR15. Visual QA uses the *full standard V*Bench benchmark* (Wu & Xie, 2024a) (191 attribute and relative-position questions) and a separate *auxiliary OCR QA set* (30 questions, no shared images). *Ranking accuracy* selects the candidate with the lowest mean conditional NLL; *generation accuracy* requires the correct answer letter, counting invalid responses as incorrect. Scores use the 0–1 scale.

Comparisons and budgets. We compare five training-free reducers and an OCR-supervised VisionZip projector. Primary OCR evaluations use greedy decoding, a 256-token limit, and repetition penalty 1.0. The budget T_{vis} counts all decoder visual input positions, excluding text and generated tokens; for prefixes, $T_{\mathrm{vis}} = T_{\mathrm{pre}} + T_{\mathrm{nat}}$. We report bf16 visual KV storage separately from whole-process memory. Appendices I and K give confidence intervals and paired comparisons.

4.2 Recognition under a Fixed Visual Budget

The learned prefix achieves the highest F1 among the evaluated compressed models at every budget (Table 1). At 24 positions, it reaches 0.386, compared with 0.154 for the strongest training-free reducer, VScan-I, and 0.125 for the supervised projector. Even six learned positions reach 0.263, exceeding every tested 24-position training-free reducer.

The projector shares the prefix's data, optimizer, schedule, and 1-D interface, so this comparison supports learning the compact representation beyond re-projecting selected features. It does not isolate architecture: the 26.3M-parameter projector is smaller than the prefix adapter. We evaluate reducers through their stronger native interface, retaining spatial positions and DeepStack features (Appendix C).

The advantage over VScan-I persists at 64 and 128 positions. Increasing capacity from 24 to 64 gives a larger gain than the final increase to 128, whose paired confidence interval includes zero. Equal

Table 1: **TextOCR quality at matched visual budgets.** Image-mean word F1 on the held-out subset. The projector and prefix share OCR supervision. Each column reports bf16 visual KV only (Section 4.1). Full resolution: 0.548 F1, 746.752 mean positions, and 105.012 MiB KV. Dashes indicate unevaluated settings.

Method	$T_{\text{vis}} = 6$	$T_{\text{vis}} = 12$	$T_{\text{vis}} = 24$	$T_{\text{vis}} = 64$	$T_{\text{vis}} = 128$
Visual KV (MiB)	0.844	1.688	3.375	9.000	18.000
Training-free reduction; native interface					
VisionZip	0.023	0.026	0.064	–	–
Attention top-k	0.026	0.029	0.069	–	–
ToMe-post	0.024	0.046	0.093	–	–
ApET	0.019	0.040	0.077	–	–
VScan-I	0.025	0.048	0.154	0.292	0.389
OCR-supervised compression; 1-D interface					
VisionZip + projector	–	–	0.125	–	–
Learned prefix (ours)	**0.263**	**0.302**	**0.386**	**0.444**	**0.463**

position counts imply equal visual KV storage, so the quality differences reflect how the image is represented within that storage budget. A 24-position prefix uses 3.375 MiB, compared with 105.012 MiB for full-resolution input, while remaining below its recognition accuracy. Appendix K gives the size calculation, paired intervals, and adaptive-budget search comparisons.

4.3 Generalization across Backbones and Datasets

Different frozen backbones. Separately trained single-block image-only adapters improve over frozen and LoRA-tuned (Hu et al., 2022b) low-resolution baselines on Qwen3-VL-4B, Qwen3-VL-8B, and GLM-4.1V-9B. Appendix E reports these results; Appendix D gives the three-seed results.

Table 2: **Zero-shot scene-text transfer.** TextOCR-trained single-block adapters with $T_{\text{nat}} \approx 12, T_{\text{pre}} = 12$, compared with a frozen nominal-24 baseline. Budgets are approximately matched by pixel caps. All entries use image-mean word F1.

Dataset	Frozen nominal-24	Conditioned 12+12	Blind 12+12
TotalText (300 images)	0.467	0.516	0.546
ICDAR15 (496 images)	0.134	0.420	0.424

Unseen scene-text datasets. Without further training, both single-block adapter variants improve on TotalText and ICDAR15 (Table 2), supporting transfer beyond the transcription training dataset.

Reading text within visual questions. On the auxiliary OCR QA set, the single-block conditioned adapter reaches 0.900 ranking accuracy, compared with 0.100 for the frozen low-resolution reference. Figure 2 illustrates recovered brand names and inscriptions, together with a remaining recognition error. On the full standard V*Bench benchmark, the same historical comparison is 0.340 versus 0.325, without a statistically clear improvement. This distinction motivates the controlled capacity and supervision study in Section 4.5; the two QA evaluations test different kinds of visual evidence (Appendix E).

4.4 Ablation Studies

Allocating visual positions. At a fixed budget of 24 positions, the pure prefix reaches 0.386 F1, compared with 0.327 for the 12+12 allocation. Adding a thumbnail to the 24-token prefix raises F1 to 0.401 at 36 positions. Thus, learned positions are preferable to thumbnail positions under this budget. Appendix B gives the full curve; the separate single-block study in Appendix D (Table 10) shows the same pattern.

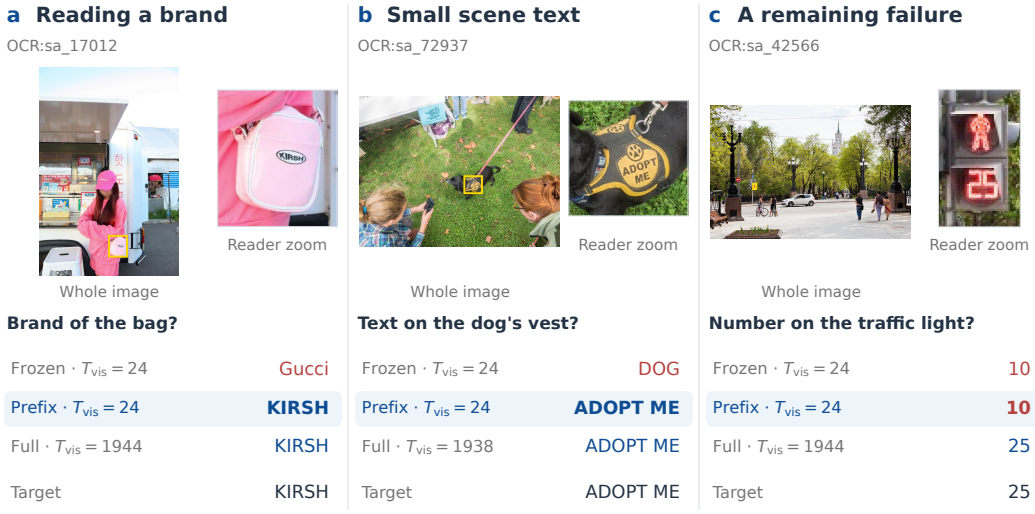

Figure 2: **Reading small text, with a remaining failure.** Selected auxiliary OCR QA examples from the context-conditioned single-block model. Both compressed models use exactly 24 visual positions on these images. Answers are selected by four-way option scoring. All models receive whole images; annotated boxes and enlarged regions are shown only to aid the reader.

Table 3: **Effect of QA supervision at $T_{\text{pre}} = 64$.** Architecture, initialization seed, and training-update budget are matched. Ranking and generation metrics follow Section 4.1.

	V*Bench (full)		Auxiliary OCR QA	
Training objective	Ranking	Generation	Ranking	Generation
OCR only	0.356	0.047	0.933	0.000
OCR + QA	0.429	0.309	0.933	0.267

Dependence on high-resolution image content. We hold the checkpoint and decoder layout fixed and replace only the compressor's image input. For the single-block pure-prefix model, TextOCR F1 changes from 0.387 with the correct image to 0.0023 with a white image and 0.0055 with a mismatched image. In the separate conditioned-model study, using a low-resolution compressor input changes F1 from 0.296 to 0.090; white- and wrong-image controls also reduce accuracy (Figure 7). These interventions show that the learned prefix uses image-specific evidence from the high-resolution path.

4.5 From Text Recognition to Visual Question Answering

A prefix trained to transcribe text need not support other visual questions equally well. Figure 10 (Appendix K) tests this by varying the prefix length while keeping OCR supervision fixed. TextOCR F1 increases with the budget, and ranking accuracy on the auxiliary OCR QA set remains above 0.86. In contrast, ranking accuracy on the full standard V*Bench benchmark stays between 0.346 and 0.372 across the same prefix lengths. Thus increasing capacity improves transcription but does not automatically transfer to attribute and spatial questions. We therefore test the training objective at a fixed length of $T_{\text{pre}} = 64$. Replacing half the OCR examples with GQA examples keeps the architecture, initialization seed, training-example count, and optimizer-update budget unchanged. Mixed supervision raises generation accuracy on the standard benchmark from 0.047 to 0.309 and ranking accuracy from 0.356 to 0.429 (Table 3). On the auxiliary OCR QA set, ranking accuracy remains 0.933 while generation accuracy rises from 0 to 0.267. Generation accuracy requires both a correct answer and the requested output format; ranking evaluates discrimination among supplied candidates. These results supports training the compact representation for the intended QA objective; it motivates the multi-task training used for the reusable prefixes in Section 5. Paired intervals and the full capacity comparison appear in Appendix K.

Table 4: **QA quality and cross-question reuse.** Qwen3-VL-4B with single-block adapters, twelve learned tokens, and a native thumbnail. (a) Equal four-task mixture, ten epochs; 500 items per task. After/First denote separately trained after-prompt/prefix-first layouts; Frozen is nominal-24. TextVQA uses a non-official soft-consensus proxy† (Appendix F). (b) Prefix-first reuse over 1,871 questions on 400 pages. ANLS is page-averaged; EM is question-weighted normalized canonical exact match. Collision is the page-mean fraction of identical predictions for question pairs with different canonical targets (↓); metric details appear in Appendix A.

(a) Multi-task quality

Task (metric)	Frozen	After	First
TextOCR (word F1)	0.178	0.200	0.236
DocVQA (ANLS)	0.139	0.194	0.244
ChartQA (relaxed acc.)	0.296	0.306	0.320
TextVQA (proxy)†	0.380	0.493	0.568

(b) DocVQA cache reuse

Model / input	ANLS	EM	Collision ↓
Frozen low-res.	0.162	0.067	0.053
LoRA SFT	0.193	0.075	0.065
Wrong-page prefix	0.230	0.114	0.051
Cached prefix	0.274	0.132	0.102
Full resolution	0.948	0.837	0.003

5 Question Answering with Reusable Prefixes

We next test the trained single-block adapter with twelve learned tokens and a small native thumbnail. Prefix-first placement makes its image KV question-independent under Proposition 1.

Answer quality with a reusable layout. The adapter is trained on an equal mixture of TextOCR, DocVQA, ChartQA, and TextVQA. Both the prefix-first and after-prompt models improve over the frozen low-resolution reference on all four tasks (Table 4(a)). The layouts are trained separately: this comparison establishes quality in a layout that permits reuse, while the following experiment tests sharing the prefix across questions.

Multiple questions from one image prefix. We encode each of 400 DocVQA pages once and answer all 1,871 associated questions, averaging 4.68 questions per page (Table 4(b)). The shared prefix reaches 0.274 page-mean ANLS, above the low-resolution and LoRA references. Substituting another page's prefix reduces ANLS to 0.230, showing a contribution from page-specific evidence. The cached prefix nevertheless remains far below full-resolution accuracy and has a higher answer-collision rate.

Constructing and reusing the image state. A separate A100 bf16 experiment measures twelve images with eight questions each (Appendix F.1). Building the 12-token cache takes 70.7 ms and stores 1.77 MB per image. Median query time including generation is 169 ms with a warm cache versus 188 ms with recomputation; greedy outputs agree on 91 of 96 pairs. Reuse avoids repeated image-prefix construction; prompt processing and answer decoding remain question-specific.

6 Limitations and Conclusion

The exact-budget experiments use one backbone and one training seed; the additional backbones and seeds use the single-block adapter. In the original $T_{\mathrm{pre}} \leq 24$ budget study, 38–50% of learned-model responses reach the 256-token output limit. The supervised projector has fewer parameters than the prefix adapter. Comparisons with longer outputs and compressors of matched capacity remain necessary, as does a controlled ablation of the final refined-vision readout. The learned compressor also adds parameters and high-resolution processing, and the tested pruning budgets cover only a small fraction of the native visual sequence.

Compact visual prefixes let a frozen decoder use high-resolution evidence through a small visual interface. The recognition comparisons, transfer results, and image-content controls support the quality of this representation under restricted visual budgets. The QA supervision control shows how the prefix can be adapted to answer prediction, and the prefix-first experiments demonstrate sharing one image state across multiple questions. Together with the causal reuse conditions, these results support compact, task-trained visual memory for frozen VLMs.

AI USE STATEMENT

Generative AI was used to assist literature retrieval and comparison, restructure and edit the manuscript, develop and check theoretical derivations, and write plotting, verification, and document-build code. Numerical figures and tables were generated from the archived experimental records; no additional model-training or benchmark runs were performed during this manuscript revision. The existing experimental provenance and the revision's source records are retained for author review.

REPRODUCIBILITY STATEMENT

Appendix A records the experimental cohorts, data partitions, training settings, and metric definitions. Appendix C documents the pruning adaptations. The accompanying manuscript repository preserves the source bundle, including per-image predictions, experiment manifests, and code snapshots. A standalone script regenerates the figures and numerical tables from these records, and verification checks their provenance and consistency. Re-running model training additionally requires the original datasets, base checkpoints, and compute environment.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2204.14198.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1612.00410.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token Merging: Your ViT But Faster. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.09461.

Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka Multimodal Models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/722fcbc1a6667f2075d75ea79a1b3552-Abstract-Conference.html.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2403.06764.

Thomas A. Courtade and Tsachy Weissman. Multiterminal source coding under logarithmic loss. *IEEE Transactions on Information Theory*, 60(1):740–761, 2014. doi: 10.1109/TIT.2013.2288257. URL https://people.eecs.berkeley.edu/~courtade/pdfs/CourtadeWeissman_TransITJan2014.pdf.

Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context Autoencoder for Context Compression in a Large Language Model. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.06945.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022a.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022b. URL https://arxiv.org/abs/2106.09685.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, et al. Perceiver IO: A General Architecture for Structured Inputs & Outputs. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2107.14795.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2301.12597.

Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of ACL-IJCNLP*, 2021. URL https://arxiv.org/abs/2101.00190.

Qiankun Ma, Ziyao Zhang, Haofei Wang, Jie Chen, Zhen Song, and Hairong Zheng. ApET: Approximation-Error Guided Token Compression for Efficient VLMs. *arXiv preprint arXiv:2602.19870*, 2026. URL https://arxiv.org/abs/2602.19870.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A Dataset for VQA on Document Images. In *IEEE Winter Conference on Applications of Computer Vision*, 2021. URL https://arxiv.org/abs/2007.00398.

Jesse Mu, Xiang Lisa Li, and Noah Goodman. Learning to Compress Prompts with Gist Tokens. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2304.08467.

Aleksandar Petrov, Philip H. S. Torr, and Adel Bibi. When do prompting and prefix-tuning work? a theory of capabilities and limitations. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.19698.

Xianhang Shen, Xingyu Liu, et al. ZoomEye: Tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://arxiv.org/abs/2105.05486.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. URL https://arxiv.org/abs/physics/0004057.

Xintong Wang et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via reinforcement learning. *arXiv preprint arXiv:2505.15436*, 2025.

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*, 2024a. URL https://arxiv.org/abs/2312.14135.

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, 2024b.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is Better but Not Necessary in Vision Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2412.04467.

Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, and Yansong Tang. VoCo-LLaMA: Towards Vision Compression with Large Language Models. *arXiv preprint arXiv:2406.12275*, 2024. URL https://arxiv.org/abs/2406.12275.

Ce Zhang, Kaixin Ma, Tianqing Fang, Wenhao Yu, Hongming Zhang, Zhisong Zhang, Haitao Mi, and Dong Yu. VScan: Rethinking Visual Token Reduction for Efficient Large Vision-Language Models. *arXiv preprint arXiv:2505.22654*, 2025a. URL https://arxiv.org/abs/2505.22654.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. LLaVA-Mini: Efficient Image and Video Large Multimodal Models with One Vision Token. In *International Conference on Learning Representations*, 2025b. URL https://arxiv.org/abs/2501.03895.

Yupeng Zheng, Kai Zou, Bin Liu, and Nenghai Yu. VisCo: Leveraging Large Language Models as Intrinsic Encoders for Visual Token Compression. *arXiv preprint arXiv:2607.12756*, 2026. URL https://arxiv.org/abs/2607.12756.

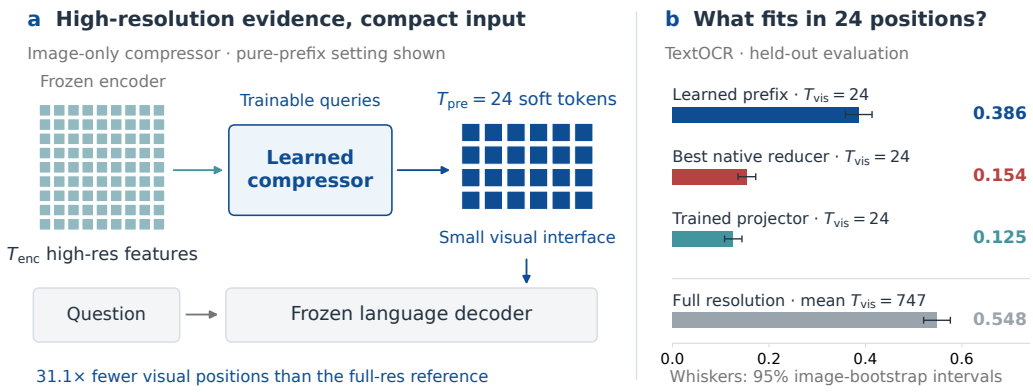

Figure 3: **High-resolution evidence through a compact visual interface.** Image-only compression into $T_{\text{vis}} = T_{\text{pre}} = 24$ positions. TextOCR word F1 is evaluated on the same 500 held-out images against the strongest tested training-free reducer and a same-supervision projector. Full resolution averages 747 positions and remains more accurate. Whiskers: pointwise 95% image-bootstrap intervals. The position ratio does not measure compute or latency.

A Experimental Protocol and Provenance

Table 5: Experimental cohorts. Rows are separate experiments, not interchangeable replications. "Legacy" means no final refined-vision readout.

Study	Model and selection	Evaluation	Budget / layout
Primary budget / pruning	Refined readout; fixed epoch 10; one seed	TextOCR val[500:1000], 500 images	Exact $T_{\text{nat}} = 0$ or 12; after prompt
Legacy allocation	Legacy blind; minimum dev filtered CER	TextOCR val[:500], 500 images	Pixel-cap thumbnail or pure 24; after prompt
Historical core / transfer	Legacy conditioned or blind; archived selected checkpoints; three backbones	TextOCR, TotalText, ICDAR15, V*Bench	Pixel-cap thumbnail + prefix; after prompt
Content controls	Legacy conditioned; locked checkpoint	2,614 TextOCR images	Same prefix budget, changed image content
Reusable memory	Legacy blind; four-task mixture; prefix-first checkpoint	400 DocVQA pages, 1,871 questions	Small native thumbnail + 12 prefix; prefix first

Primary training settings. The primary backbone is Qwen3-VL-4B-Instruct. The adapters use all 21,749 TextOCR training images, with effective batch size 64 and a 2,000,000-pixel high-resolution encoder cap. All primary OCR evaluations use greedy decoding, a 256-token output limit, and repetition penalty 1.0.

The original six-arm budget grid uses the full TextOCR training split, microbatch 1, gradient accumulation 64, 10 epochs, AdamW with first- and second-moment coefficients $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.01, learning rate 5×10^{-5}, 100 warmup steps, and a cosine schedule. Initialization and data-order seeds are 20260913. Early stopping is disabled. Out-of-memory errors fail a run instead of silently skipping samples. Each budget has a separate training run. Training, development, and heldout manifests have disjoint image paths. The source snapshot and configuration hashes are recorded in the archived budget-study manifest.

The frozen backbone revision is `ebb281ec70b05090aa6165b016eac8ec08e71b17`. The development set contains the first 500 validation images; the evaluation set contains images 501–1000. This evaluation set is excluded from checkpoint selection in the primary study. We do not claim that these images were unused in all earlier development experiments. The primary checkpoint is fixed at epoch 10; secondary results select checkpoints by development filtered CER. The 2,000,000-pixel

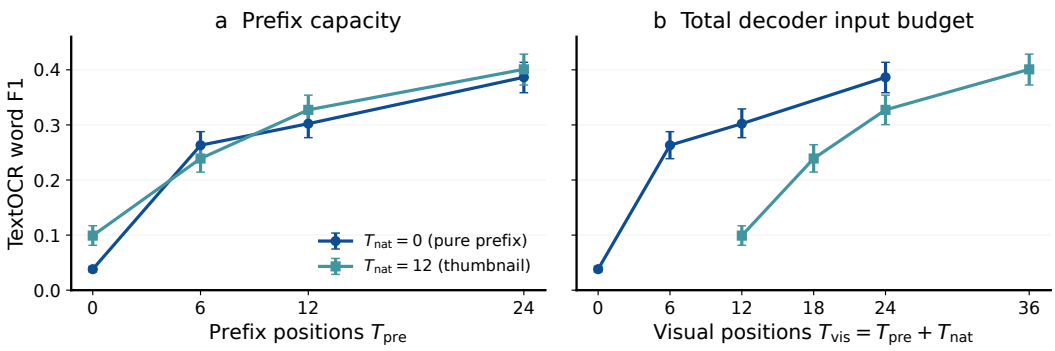

Figure 4: **OCR improves with prefix length.** Fixed epoch-10 results on the same 500 images, with pointwise 95% bootstrap intervals. Left: learned prefix length T_{pre}. Right: total visual budget $T_{\text{vis}} = T_{\text{nat}} + T_{\text{pre}}$, including the 12 thumbnail positions used by the hybrid models.

high-resolution cap produces a median of 768 and a mean of 746.752 native visual positions on this evaluation set.

Metrics. The standard OCR metric lowercases text, strips punctuation, and matches predicted and reference words one-to-one without requiring reading order. Precision, recall, and F1 are computed per image before averaging. This transcription metric does not require bounding-box localization, unlike text-spotting H-mean. Blank outputs remain blank rather than falling back to the prompt. Filtered CER is used for the stated development-checkpoint selection, not as a substitute for the primary standard word F1. DocVQA uses ANLS and ChartQA relaxed accuracy. The historical TextVQA scorer uses a simplified soft-consensus metric, defined in Appendix F. The V*Bench and auxiliary OCR QA results distinguish candidate-ranking accuracy from generation accuracy, as defined in Section 4.1. We re-score the available QA content controls with their corresponding task metrics instead of mixing their archived ANLS with another metric. In the multi-question experiment, ANLS is averaged per page and then across pages. Exact match is instead question-weighted equality to the canonical target after case and whitespace normalization; it differs from the evaluator's multi-answer exact-match summary. Collision is the page-mean fraction of identical predictions among question pairs with different canonical targets, excluding pages without such pairs.

Legacy pixel-cap accounting. A 16,384-pixel cap gives median 12 native positions but a per-image range of 10–16. The legacy 12+12 allocation therefore has 22–28 total visual positions. Its measured mean is 24.466, compared with 24.852 for the nominal 24-position frozen reference. A historical point labeled "48 tokens" has mean 55.42 and median 54 positions. We do not use those nominal labels as exact budgets. The legacy pure 24-query model uses exactly 24 positions. The primary budget study replaces pixel-cap thumbnails with exact-12 grids.

Readability calibration. On 21,749 training images, the original 16,384-pixel cap gives median rendered bounding-box-height proxy 2.161 pixels. Searching pixel caps in increments of 1,024 on training annotations gives largest caps with median heights below 4, 6, and 8 pixels of 48,128, 109,568, and 187,392. This is a description of the operating regime, not a retrospective proof that the original cap was optimal. Character width and bounding-box height are distinct proxies.

B Complete Prefix-Budget Results

For the development-selected checkpoints, pure-prefix F1 at $T_{\text{pre}} = 6, 12, 24$ is 0.2605, 0.3022, and 0.3807. The thumbnail-plus-prefix scores are 0.2335, 0.3203, and 0.4060, respectively; the archived analysis supplies full-precision values and intervals. These secondary results also increase with T_{pre}. The main text uses only fixed epoch 10 for this cohort.

Table 6: Fixed epoch-10 standard word F1. $T_{\mathrm{pre}} = 0$ uses the frozen model without a compressor. Pointwise 95% intervals appear as error bars in Figure 4; numerical intervals for every setting are archived.

T_{nat}	$T_{\mathrm{pre}} = 0$	$T_{\mathrm{pre}} = 6$	$T_{\mathrm{pre}} = 12$	$T_{\mathrm{pre}} = 24$
0	0.038	0.263	0.302	0.386
12	0.099	0.239	0.327	0.401

Table 7: Paired differences on the same 500 images. Intervals condition on one training seed and are not adjusted for multiple comparisons.

T_{nat}	Prefix step	Paired Δ F1	95% CI
0	$0 \to 6$	+0.225	[0.200, 0.250]
0	$6 \to 12$	+0.039	[0.021, 0.057]
0	$12 \to 24$	+0.084	[0.063, 0.104]
0	$0 \to 24$	+0.348	[0.319, 0.377]
12	$0 \to 6$	+0.140	[0.118, 0.163]
12	$6 \to 12$	+0.088	[0.069, 0.107]
12	$12 \to 24$	+0.074	[0.053, 0.094]
12	$0 \to 24$	+0.302	[0.275, 0.330]

C Reducer Implementations and Full Comparisons

Let T_{keep} denote the number of high-resolution feature embeddings retained or produced by a reducer. All methods use the same 500-image split and decoding protocol as the learned budget curve. The pure interface inserts $T_{\mathrm{keep}} = 6, 12, 24$ reduced embeddings after the prompt using 1-D positions. The native interface retains the same counts at their spatial rotary positions, with aligned DeepStack features. Both have $T_{\mathrm{vis}} = T_{\mathrm{keep}}$. The hybrid interface adds a 12-token thumbnail, giving $T_{\mathrm{vis}} = 12 + T_{\mathrm{keep}}$. The native-12 interface instead retains $T_{\mathrm{keep}} = 18, 24, 36$ high-resolution tokens without a separate thumbnail, so $T_{\mathrm{vis}} = T_{\mathrm{keep}}$. These four interfaces and three budgets give twelve configurations per reducer. The best-reducer curves select the highest F1 on the evaluation set at each budget, favoring the baseline methods. For native interfaces, the same reduction map is applied to the visual embeddings and all three DeepStack streams. Full-budget preflight checks compare the adapted path with the native model before evaluating reductions.

VisionZip and attention selection. VisionZip is adapted from the released Qwen2.5-VL implementation. Attention top-k supplies a simpler selection control. Both use the frozen high-resolution encoder; neither receives OCR training in the training-free tables.

Post-encoder ToMe-style merging. The baseline applies bipartite merging after encoding. This isolates a compact decoder input, but is not the original layerwise ViT acceleration algorithm. It does not measure the encoder speedups reported by Bolya et al. (2023).

ApET. The port uses farthest-point sampling for the reconstruction basis, a fixed seed rather than a random starting point, ridge parameter 10^{-5}, and retains high-reconstruction-error tokens. Because a basis of 10 is undefined for these small budgets, its size is $\min(10, \lfloor T_{\mathrm{keep}}/2 \rfloor)$. The low-error basis candidates are mixed with basis features using the released implementation's index pairing; discarded tokens are assigned by cosine nearest neighbor and averaged into retained tokens.

VScan stage I. Local and global scans each contribute approximately half the retained set. Qwen3-VL lacks the Qwen2.5-VL window-attention structure used by the original implementation, so local attention comes from the first DeepStack tap and global attention from the final block. The local scan uses 4×4 windows on the merged grid. Dropped tokens are merged by cosine-nearest-neighbor averaging. Stage II, which prunes inside the decoder, is not included in these pre-decoder comparisons.

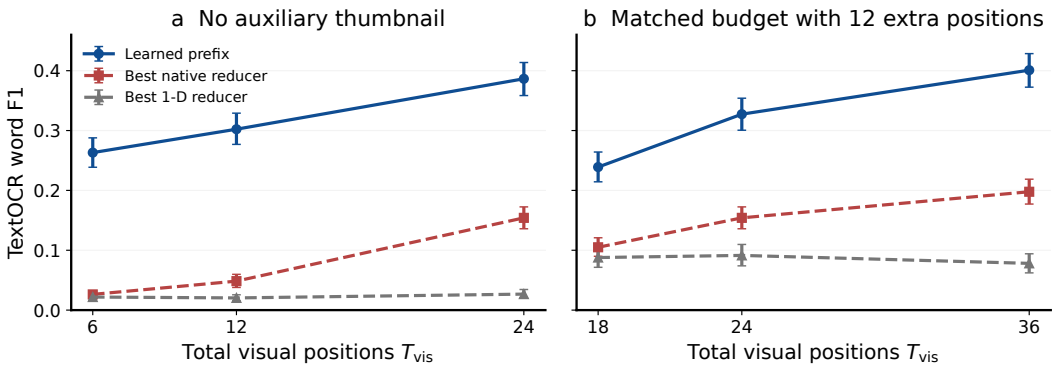

Figure 5: **Learned prefixes outperform training-free reduction at small budgets.** Each baseline curve shows the best of five reducers at each budget, with pointwise 95% intervals for the selected method. Native reductions retain spatial positions and DeepStack features; 1-D reductions use prefix insertion. The horizontal axis is total decoder visual positions T_{vis}. Panel (a) uses 6, 12, or 24 positions. In panel (b), native reductions retain 18, 24, or 36 high-resolution tokens, matching a 12-token thumbnail plus 6, 12, or 24 learned or reduced embeddings.

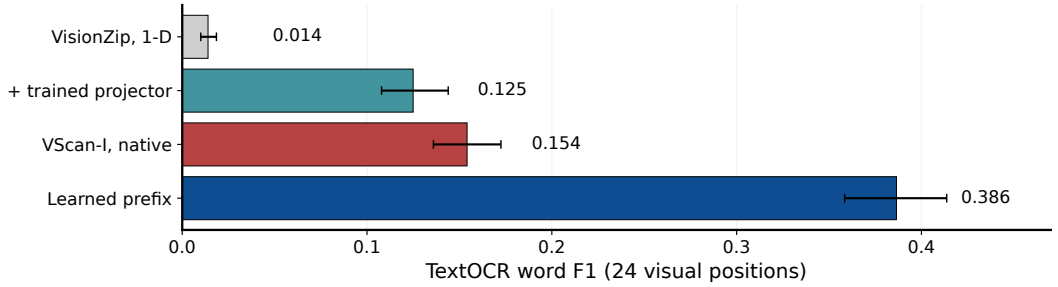

Figure 6: **A same-supervision control at 24 positions.** The trained projector improves VisionZip at the 1-D prefix interface. The strongest training-free native reducer is included separately to show the interface advantage. Error bars are pointwise 95% image-bootstrap intervals. Projector and learned-prefix parameter counts are not matched.

Supervised VisionZip projector. The residual branch is RMS Norm, Linear$(2560, 5120)$, GELU, and Linear$(5120, 2560)$, where each pair specifies input and output widths, with zero-initialized final weights and zero-initialized learnable slot biases. It has 26,286,080 parameters. The training schedule and data match the pure-prefix $T_{\text{pre}} = 24$ arm. The epoch-10 score is 0.1249; the development-selected epoch-8 checkpoint scores 0.1198. The latter selection used development filtered CER, not heldout word F1.

FastV layerwise accounting. Attention averaged over heads from the last prompt token at layer 1 ranks the visual tokens. The selected T_{keep} visual tokens are retained from layer 2 onward. The initial two layers see the full sequence: mean 746.752 image positions plus 18 text positions, or about 765 total. Later layers see $T_{\text{keep}} + 18$ positions. Keeping all tokens matches native greedy decoding for the eight-step preflight. This path cannot be treated as an input-budget-T_{keep} comparison.

D Compressor Architecture and Content Controls

The single-block adapters in the main text are labeled "legacy" in the archived experiment records and appendix tables. The legacy blind compressor has one sequential interaction block and returns the query branch before any further read of the updated visual branch. With active weights mapped to a one-way resampler, the recorded output difference is zero even with a nonzero output projection. Unused-branch gradients are absent. The audit records 251,089,920 registered parameters

Table 8: All 60 pre-decoder reducer configurations. Cells report standard word F1; full pointwise and paired intervals are retained in the archived analysis. Column headers give total decoder visual positions T_{vis}; T_{keep} counts reducer outputs. Selected comparisons appear in Appendix I. These are adaptations on Qwen3-VL at extreme budgets, not reproductions of the original benchmark settings.

Pure: T_{keep} embeddings in the after-prompt 1-D interface

Reducer	$T_{\text{vis}} = 6$	$T_{\text{vis}} = 12$	$T_{\text{vis}} = 24$
ApET	0.017	0.013	0.018
VScan-I	0.022	0.020	0.027
VisionZip	0.017	0.016	0.014
Attention top-k	0.016	0.016	0.015
ToMe-post	0.018	0.018	0.024

Native: T_{keep} high-resolution tokens, spatial positions + DeepStack

Reducer	$T_{\text{vis}} = 6$	$T_{\text{vis}} = 12$	$T_{\text{vis}} = 24$
ApET	0.019	0.040	0.077
VScan-I	0.025	0.048	0.154
VisionZip	0.023	0.026	0.064
Attention top-k	0.026	0.029	0.069
ToMe-post	0.024	0.046	0.093

Hybrid: exact-12 thumbnail + T_{keep} embeddings in the 1-D interface

Reducer	$T_{\text{vis}} = 18$	$T_{\text{vis}} = 24$	$T_{\text{vis}} = 36$
ApET	0.085	0.091	0.078
VScan-I	0.078	0.078	0.062
VisionZip	0.078	0.066	0.052
Attention top-k	0.076	0.064	0.046
ToMe-post	0.088	0.076	0.059

Native-12: $T_{\text{keep}} = 18, 24, 36$ tokens retained in place

Reducer	$T_{\text{vis}} = 18$	$T_{\text{vis}} = 24$	$T_{\text{vis}} = 36$
ApET	0.064	0.077	0.115
VScan-I	0.105	0.154	0.198
VisionZip	0.050	0.064	0.096
Attention top-k	0.051	0.069	0.097
ToMe-post	0.065	0.093	0.157

Table 9: FastV results with all visual tokens present in the first two decoder layers. Output cap is 256 new tokens.

Retained T_{keep}	Word F1	95% CI	Output-cap rate
6	0.005	[0.003, 0.008]	10.8%
12	0.007	[0.004, 0.010]	16.8%
24	0.027	[0.020, 0.034]	22.4%

Table 10: **Learned tokens versus thumbnail positions.** Single-block adapters on the first 500 TextOCR validation images, selected by development filtered CER. T_{vis} is the measured mean visual budget; thumbnail token counts vary with image dimensions.

Representation	Mean T_{vis}	Word F1
Frozen low-resolution	24.85	0.178
LoRA SFT	24.85	0.184
Learned, $T_{\text{nat}} \approx 12, T_{\text{pre}} = 12$	24.47	0.324
Learned, $T_{\text{nat}} = 0, T_{\text{pre}} = 24$	24.00	0.387
Learned, $T_{\text{nat}} \approx 24, T_{\text{pre}} = 24$	48.85	0.360

and 92,477,440 output-connected parameters for its 24-query configuration. The latter is the corresponding one-way model's parameter count. Independent initialization and training can still lead to different scores; comparing those runs does not identify a bidirectional architectural effect.

The refined-readout budget experiments instead implement Equation 4, so the updated visual features reach the returned prefix. Gradient smoke checks establish connectivity for this implementation. No accuracy-matched one-way or original-vision-readout control accompanies that budget curve, so architectural necessity remains unresolved.

Legacy allocation. The development-selected comparison favors the pure 24-token prefix (0.387 F1) over the approximately 12+12 allocation (0.324); the difference is 0.057 on the subset with exactly matched position counts. For the allocation comparison in Table 10, fixed epoch-10 scores are 0.383 for the pure 24-token prefix and 0.324 for the 12+12 hybrid, consistent with the development-selected comparison. The allocation changes both thumbnail presence and learned-query count, so it measures the choice between these representations at a similar total budget. Image-content interventions change the compressor input without retraining; white and mismatched images may also introduce distribution shift.

Table 11: Legacy architecture controls. Same nominal 12+12 budget and OCR recipe, repetition penalty 1.0. Registered parameter counts differ from output-connected counts in the audited blind implementation.

Compressor	Registered parameters	Word F1
Average-pool + MLP	38M	0.111
One-way learned-query resampler	92M	0.305
Legacy conditioned	251M	0.284
Legacy blind	251M	0.313
Conditioned + decoder LoRA	251M + 12M	0.344

Blind-minus-conditioned F1 is 0.029 [0.013, 0.045] under the historical recipe. The seed summaries are 0.3153 ± 0.0024 (three blind runs) and 0.2965 ± 0.0107 (three conditioned runs). The older summary omitted the completed third conditioned seed; these values are recomputed from all six prediction files. These observations support omitting conditioning for this recipe; they are not proof that prompt conditioning is generally harmful.

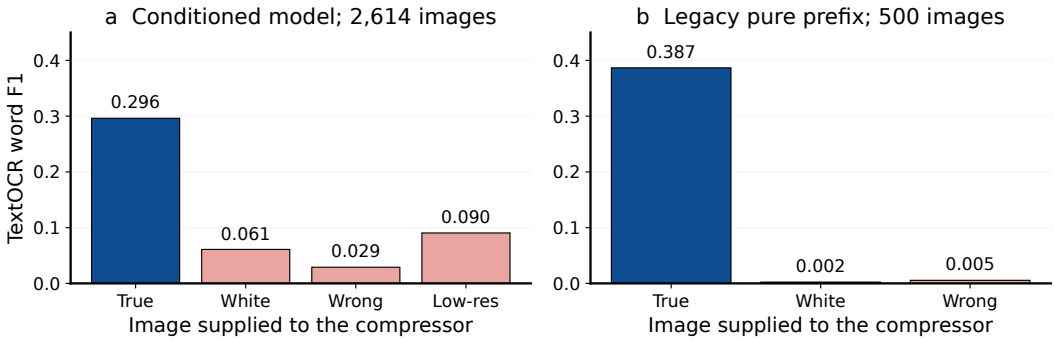

Figure 7: **Image interventions on two legacy cohorts.** Only the image sent to the compressor changes. Panel (a) uses the conditioned model on 2,614 images. Panel (b) uses the legacy pure 24-query model on the first 500 validation images. These are not interventions on the refined-readout heldout models. Bars show archived point estimates; paired intervals are collected in Appendix I.

Table 12: **TextOCR results across backbones.** Single-block adapters on the first 500 validation images, with approximately matched budgets. Frozen GLM uses native thinking; trained GLM uses direct-answer supervision (Appendix E). Each trained entry reports one run. Δ is the word-F1 improvement of the blind prefix over the frozen nominal-24 baseline.

Arm	Qwen3-VL-4B	Qwen3-VL-8B	GLM-4.1V-9B
Frozen nominal-12	0.113	0.117	0.074
Frozen nominal-24	0.178	0.188	0.120
LoRA SFT nominal-24	0.184	0.186	0.153
Blind 12+12	0.313	0.305	0.338
Δ vs frozen-24	+0.135	+0.117	+0.218

E Transfer and Legibility Analysis

Table 2 reports the zero-shot OCR results. The conditioned checkpoint is epoch 8; the image-only checkpoint is the archived selected model.

V*Bench and the auxiliary OCR QA set. On the fixed auxiliary OCR QA set (30 questions), the historical conditioned adapter scores 27/30, compared with 3/30 for the small frozen reference and 29/30 for the full-resolution reference. The paired adapter-minus-small reference difference is 0.800 [0.633, 0.933]. All arms see whole images, without crops or visual search. Budgets agree exactly on 26/30 images; adapter totals range from 24 to 27 and frozen totals from 24 to 30. The full standard V*Bench benchmark contains 115 attribute and 76 relative-position questions, with no image paths shared with the auxiliary OCR QA set. On that standard set, accuracy is 0.340 versus 0.325 and the archived paired interval for the difference is $[-0.052, 0.084]$. Both white- and wrong-image controls score 0.288, indicating some content sensitivity despite the inconclusive gain over the budget-matched baseline. The larger-budget follow-up is 0.361 versus 0.367. The positive auxiliary OCR result does not establish general visual-detail transfer. The newer legacy pure-prefix allocation obtains 28/30 versus 27/30 for its hybrid control, a one-example difference.

Rendered legibility. Figure 8 stratifies 18,643 reference words from three datasets by a rendered-height proxy. Matching requires consistent reference words across arms; three TotalText images with inconsistent references are excluded from this descriptive analysis. Counts are 14,371 TextOCR words, 2,176 TotalText words, and 2,096 ICDAR15 words. Word-pooled recall differs from the image-mean F1 used elsewhere. These associations do not control blur, font, perspective, or all correlations with word load. The curves therefore support a descriptive account of where gains occur, not a universal law of visual capacity.

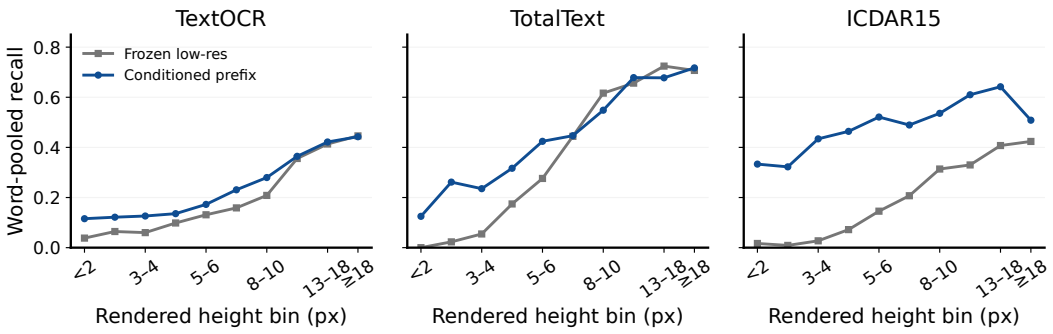

Figure 8: **Descriptive word recall by rendered height.** Legacy conditioned model versus the nominal-24 frozen view. Dataset-specific behavior remains visible instead of being hidden by a pooled curve. Markers show observed bin means; lines only connect bins.

Other backbones. Historical blind 12+12 models on Qwen3-VL-8B and GLM-4.1V-9B also improve on low-resolution references (Table 12). These are single-run legacy replications, not the exact-budget refined-readout protocol. For GLM, pixel caps of 25,088 and 50,176 are calibrated on 60 validation images to approximate the 12- and 24-position distributions. Frozen GLM arms use up to 1,280 generated tokens to allow their native thinking blocks, then extract and truncate the answer to 256 tokens. About 19% of nominal-24 and 11% of nominal-12 outputs fail to close the thinking block and are scored on that text. Trained arms use direct-answer supervision after an empty-think prefill. These protocol differences qualify comparisons across model families. Table 12 is re-scored from predictions; its intervals use 5,000 paired image resamples with seed 20260920. These intervals condition on the selected checkpoint and do not estimate training variance. Rounding is applied once after aggregation, so the 8B prefix score is 0.305 rather than rounding the four-decimal summary again to 0.306.

Negative results. Gains on dense HierText documents are small; task shift, finite output length, and representation limits remain competing explanations. A separately trained GQA small-object model scores 0.564 versus 0.586 for its approximately matched frozen reference, with weak sensitivity to image-content controls. The tested text-domain prefix variants show little useful content dependence. Later code-completion diagnostics nevertheless show that oracle-selected literal token embeddings are readable through the frozen interface. Thus failed learned compressors do not establish that the decoder cannot read text prefixes. The code branch uses a different task and protocol and is not pooled with visual evidence.

Earlier high-budget DocVQA runs further illustrate the importance of content controls: on 1,337 questions the legacy adapter has ANLS 0.9469 with the correct high-resolution image, 0.9464 with a white image, and 0.9484 with the small view. Those gains chiefly support task adaptation, not useful high-resolution compression. Early TotalText training, synthetic padding, and COCO detection runs use different budgets, targets, or generation implementations; they are retained as development history rather than combined with the final comparisons. These observations constrain the demonstrated scope without establishing impossibility under other recipes.

F Multi-Task Training and Cache Details

Table 4(a) reports the four-task mixture, trained on 15,000 examples per task for 10 epochs and evaluated on 500 items per task. The archived TextVQA score is $\min(n_{\text{match}}/3, 1)$, where n_{match} is the number of human answers matching the prediction after lowercasing, collapsing whitespace, and removing trailing periods. It does not implement the official leave-one-annotator-out scoring or its full answer normalization, and should not be compared directly with leaderboard values. ChartQA accepts numeric answers within 5% relative error (exact zero for a zero target), otherwise normalized exact matches. The archived TextVQA data-preparation audit finds no overlap between its validation images and TextOCR training images by image ID; this does not establish perceptual deduplication or globally untouched evaluation images. The OCR-only adapter scores 0.000 on a historical zero-shot

DocVQA run, frequently transcribing instead of answering. A three-task mixture omitting scene-text QA still scores only 0.044 on TextVQA. Adding same-domain QA raises the after-prompt model's score to 0.311 after five epochs and 0.493 after ten. These observations motivate including QA from the same visual domain; they do not isolate all sources of task interference.

Table 13: Prefix-first image controls re-scored with the same task metric as Table 4(a), 500 items per arm. Missing controls are shown as dashes. TextVQA retains the same soft-consensus definition (†).

Prefix-first task	Correct	White	Wrong image	Small view
DocVQA (ANLS)	0.244	0.001	0.134	0.030
ChartQA (relaxed acc.)	0.320	0.020	–	–
TextVQA (soft consensus)†	0.568	0.023	0.354	–

The prefix-first DocVQA controls score 0.001 with a white image, 0.134 with a wrong image, and 0.030 with the small view. The recovered TextVQA wrong-image score of 0.354 also shows that substantial task performance can survive an incorrect prefix; the correct-image difference is more informative than the absolute score alone.

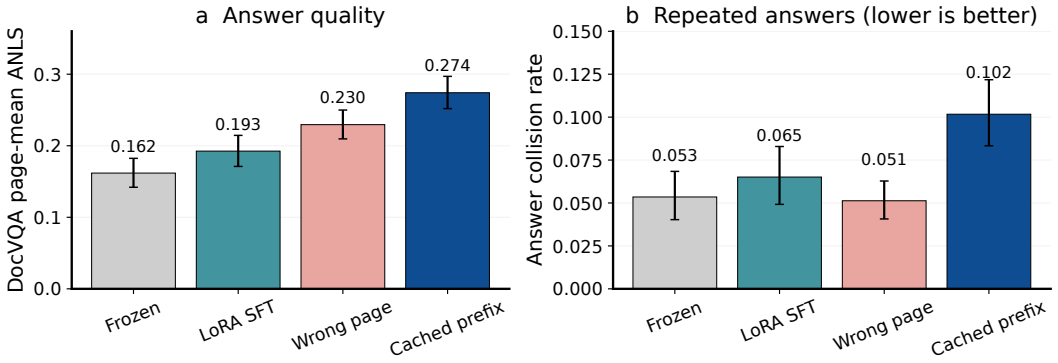

Figure 9: **Reusable memory has both gains and failure modes.** Page-clustered 95% intervals on 400 DocVQA pages. The correct cached prefix improves ANLS over a wrong-page prefix and low-resolution baselines, but repeats identical answers across distinct questions more often.

Implementation. Prefix-first rotary positions continue all three axes from a shared scalar maximum. A fixed filler-token embedding provides the base for the learned prefix, avoiding the unstable zero-vector start observed during training. The adapter uses fp32 parameters while the base model runs in bf16. At the low learning rate used here, maintaining adapter parameters directly in bf16 can suppress small optimizer updates.

F.1 Systems measurements and their scope

Table 14: Archived A100 bf16, batch-one microbenchmark. Medians over 12 TextOCR images after dropping one warmup image. "Online" includes high-resolution encoding and adapter computation. These legacy measurements do not benchmark the refined-readout model.

Arm	Median visual positions	Prefill (ms)	Peak CUDA (GB)
Frozen low-resolution	24.5	53.5	8.984
Frozen intermediate-resolution	60	56.0	9.014
Frozen full-resolution	672	82.4	9.645
Online legacy prefix	24	90.5	10.108

The recovered E5 aggregate measurements show that computing the prefix online takes 90.5 ms, exceeding the 82.4 ms full-resolution reference in this small workload, with higher peak allocated

CUDA memory (10.108 versus 9.645 GB). Thus fewer decoder positions do not imply lower end-to-end latency or peak memory. The separately measured vision-plus-adapter cost is 37.9 ms. The historical 53.2 ms "cached" figure is the median of per-image subtractions, an embedding-cache estimate rather than a measured KV restore. Native decode estimates are 28.1–28.9 ms/token. Original per-image E5 timings were not saved.

The separate prefix-first experiment does retain 12 per-image build times, 96 warm/cold generation pairs, and 96 equality flags. Its code times cache cloning plus generation on the warm path, and recomputation plus generation on the cold path; preprocessing and host-to-device input transfers are outside these timers. The source page specifies a 24-new-token cap. The raw JSON does not record the invocation, device identity, checkpoint hash, or cap, so these configuration details rely on the archived code and source-page report. Medians include the recorded first calls without a separate warmup exclusion. Let t_{build} be cache construction time, t_{warm} be warm query time, and m be the number of questions on the same image. The calculated $(t_{\mathrm{build}} + m\, t_{\mathrm{warm}})/m$ estimates are 239.88, 204.51, 186.83, and 177.99 ms for $m = 1, 2, 4, 8$. They are not separately timed multi-query workloads. The cached prefix contains 1,769,472 bytes: 12 positions, 36 layers, two KV tensors, eight KV heads, 128 dimensions, and two-byte values. The source page additionally reports fp32 logit agreement within 1.1×10^{-4}; the recovered JSON verifies only 91/96 bf16 greedy-output matches, not that fp32 claim or the cause of the five mismatches. No hardware benchmark was rerun during this revision.

G Decoding Audit

Earlier training-time OCR evaluations used repetition penalty 1.2 while standalone baselines used 1.0. Mixing those values would exaggerate the comparison. The present manuscript uses the audited 1.0 scores. For the legacy 4B frozen nominal-12/24 arms, F1 is 0.113/0.178 at 1.0 and 0.129/0.189 at 1.2. LoRA SFT-24 changes from 0.184 to 0.185; the conditioned adapter from 0.284 to 0.363; the blind adapter from 0.313 to 0.396; and the one-way resampler from 0.305 to 0.392. Repetition control therefore has a larger effect on the learned-prefix arms. The comparison must use a common decoding configuration. These differences do not support a claim that blind compression eliminates repetition.

H Qualitative Material

The archived paper page contains five original qualitative cards with embedded images and predictions. Three TextOCR cards illustrate a poster, phone interface, and bicycle-brand inscription; two V* OCR cards concern a bag label and a dog's vest. They are illustrative selected examples, not additional evaluation samples. Their images, outputs, and low-resolution views are retained unchanged in the source bundle. Claims about improved recognition in the present paper are based on the quantitative evaluations, not on these examples alone.

I Statistical Detail

The main tables emphasize point estimates and effect sizes. Here we collect the numerical uncertainty for the principal scores and comparisons (Tables 15 and 16). All intervals are two-sided 95% bootstrap intervals. Primary TextOCR and legacy allocation analyses use 5,000 image resamples; conditioned content controls use 10,000. The backbone reanalysis uses 5,000 image resamples with seed 20260920. Multi-question DocVQA uses 10,000 page resamples, preserving dependence among questions on the same page. These cohorts are separate experiments.

Intervals condition on the selected checkpoint and measure sampling uncertainty, not variation across training runs. The $\pm$ values reported for the six legacy seed runs are instead sample standard deviations across seeds. Bootstrap intervals are not converted into standard deviations. No multiplicity correction is applied. For the best reducer at each budget, the interval is conditional on the selected arm, not adjusted for selecting among reducers. Full-precision intervals for all budget and reducer settings remain in the archived analyses; figures display the corresponding pointwise uncertainty bars. Adjacent prefix-budget differences appear in Appendix B, and the standard V*Bench null result and the auxiliary OCR result retain their intervals in Appendix E.

Table 15: Point estimates and 95% intervals for the principal reference scores, legacy allocation study, and reusable-memory study. Metrics and resampling cohorts are identified within the table.

Arm	Estimate	95% CI
Primary TextOCR, 500 images; word F1		
Pure prefix, 24 positions	0.386	[0.358, 0.414]
VScan-I native, 24 positions	0.154	[0.136, 0.172]
VisionZip + trained projector, 24 positions	0.125	[0.108, 0.144]
Full-resolution reference	0.548	[0.521, 0.576]
Legacy allocation, 500 images; word F1		
Frozen nominal-24	0.178	[0.155, 0.202]
LoRA SFT nominal-24	0.184	[0.160, 0.209]
Learned $T_{\mathrm{nat}} \approx 12, T_{\mathrm{pre}} = 12$	0.324	[0.297, 0.350]
Learned $T_{\mathrm{nat}} = 0, T_{\mathrm{pre}} = 24$	0.387	[0.360, 0.413]
Learned $T_{\mathrm{nat}} \approx 24, T_{\mathrm{pre}} = 24$	0.360	[0.332, 0.387]
Multi-question DocVQA, 400 pages; page-mean ANLS		
Frozen low-resolution	0.162	[0.142, 0.182]
LoRA SFT	0.193	[0.171, 0.214]
Prefix from wrong page	0.230	[0.210, 0.250]
Cached image prefix	0.274	[0.252, 0.297]
Full-resolution reference	0.948	[0.936, 0.959]

Table 16: Paired differences underlying the main comparisons. Each row subtracts the second arm from the first on the same evaluation items; 95% intervals use image or page resampling as specified above.

Comparison	Paired Δ	95% CI
Primary TextOCR, 500 images; paired word-F1 differences		
Learned minus best native, $T_{\mathrm{vis}} = 6$	+0.237	[0.212, 0.262]
Learned minus best native, $T_{\mathrm{vis}} = 12$	+0.254	[0.228, 0.280]
Learned minus best native, $T_{\mathrm{vis}} = 24$	+0.232	[0.204, 0.261]
Learned minus trained projector, $T_{\mathrm{vis}} = 24$	+0.262	[0.236, 0.286]
Legacy allocation; paired word-F1 differences		
Pure minus hybrid, all 500 images	+0.063	[0.045, 0.081]
Pure minus hybrid, 409 exact-budget images	+0.057	[0.036, 0.078]
Conditioned content controls, 2,614 images; word-F1 differences		
Correct minus white image	+0.235	[0.224, 0.247]
Correct minus wrong image	+0.267	[0.255, 0.279]
Legacy backbone replication, 500 images; word-F1 differences		
Qwen3-VL-4B: blind minus frozen-24	+0.135	[0.112, 0.157]
Qwen3-VL-4B: blind minus SFT-24	+0.129	[0.107, 0.152]
Qwen3-VL-8B: blind minus frozen-24	+0.117	[0.092, 0.142]
Qwen3-VL-8B: blind minus SFT-24	+0.119	[0.096, 0.144]
GLM-4.1V-9B: blind minus frozen-24	+0.218	[0.191, 0.244]
GLM-4.1V-9B: blind minus SFT-24	+0.185	[0.160, 0.210]
Multi-question DocVQA, 400 pages; page-mean ANLS differences		
Cached image prefix minus LoRA SFT	+0.082	[0.057, 0.106]
Correct-page minus wrong-page prefix	+0.045	[0.033, 0.056]

J Theoretical Analysis and Proofs

The analysis separates two questions: what answer information survives the visual interface, and when the decoder states of that interface can be reused. The first part specializes standard conditional-entropy and log-loss identities; the second formalizes causal cache equivalence. Neither assumes that training finds an optimal compressor.

J.1 Setup and relation to prior theory

Fix the trained compressor and decoder parameters and a joint distribution of image X_H, question Q, and answer Y. Their realizations are x_H, q, y; the random variables V, Z below have realizations V_H, z from Section 3. Let $V = E(X_H)$ and let $Z = C_\theta(V)$ be deterministic, including any deterministic quantization. The side input $S = (Q, R)$ contains everything else available to the decoder: R includes thumbnail embeddings, intermediate visual injections, and their image-dependent positional metadata, if any. Fixed templates are implicit. For the pure-prefix interface, R is empty. We make no independence assumption between questions and images.

Answers are discrete token sequences under a common finite vocabulary and output horizon, with an absorbing end-of-sequence convention. Thus $p_\phi(Y \mid S, Z)$ is a normalized sequence distribution and its negative logarithm is the sum of answer-token losses in Equation 7. All logarithms are natural, information quantities are in nats, and expected log-losses are finite. P denotes the data distribution; p_ϕ denotes the frozen decoder's distribution through the specified input layout. The analysis concerns population sequence log-loss, not per-token averages with example-dependent normalization, finite-sample generalization, or greedy word F1.

The information bottleneck formulation distinguishes compression of an input from preservation of information about a target (Tishby et al., 2000). Variational formulations relate predictive log-loss to that information (Alemi et al., 2017), while log-loss source coding makes the conditional-entropy relation explicit (Courtade & Weissman, 2014). We use this relation to keep track of the additional mismatch imposed by our frozen decoder; we do not optimize an information-bottleneck rate penalty. Prefix-tuning theory also identifies restrictions on how prefixes affect a fixed attention layer (Petrov et al., 2024). Those results do not directly characterize our image-conditioned, multilayer interface, so we retain the decoder mismatch instead of assuming that any information-preserving code is readable.

J.2 Loss decomposition and comparison with no prefix

Proof of Proposition 2. Conditioning on (S, Z), add and subtract the true conditional log-probability:

$$\begin{aligned}\mathcal{R}_\phi(C) &= \mathbb{E}\left[-\log P(Y \mid S, Z) + \log \frac{P(Y \mid S, Z)}{p_\phi(Y \mid S, Z)}\right] \\ &= H(Y \mid S, Z) + \Delta_\phi(C). \end{aligned} \tag{10}$$

Since Z is a deterministic function of V, it adds no information once (S, V) is given. Hence $H(Y \mid S, V, Z) = H(Y \mid S, V)$ and

$$I(Y; V \mid S, Z) = H(Y \mid S, Z) - H(Y \mid S, V). \tag{11}$$

Substitution gives Equation 9. Both the conditional mutual information and the expected KL divergence are nonnegative. □

The term $H(Y \mid S, V)$ is the optimal log-loss given these encoder features and side inputs, not the measured loss of the original full-resolution VLM. Equality with this lower bound requires both $I(Y; V \mid S, Z) = 0$ and $\Delta_\phi(C) = 0$: the code must preserve all answer-relevant information conditional on S, and the frozen decoder must realize the resulting conditional distribution. Our experiments do not establish either equality.

For a no-prefix model using the *same* side input S, let p^0_ϕ denote its decoder distribution, $\mathcal{R}_0$ its population log-loss, and Δ_0 its decoder mismatch:

$$\mathcal{R}_0 = \mathbb{E}[-\log p^0_\phi(Y \mid S)], \qquad \Delta_0 = \mathbb{E}_S D_{\mathrm{KL}}\big(P(Y \mid S) \,\|\, p^0_\phi(Y \mid S)\big). \tag{12}$$

Applying Equation 10 to both paths gives

$$\mathcal{R}_0 - \mathcal{R}_\phi(C) = I(Y; Z \mid S) + \Delta_0 - \Delta_\phi(C). \tag{13}$$

This gives the log-loss improvement condition for fixed side input. It does not directly compare allocations that change the thumbnail, since those change S. A constant task prefix has $I(Y; Z \mid S) = 0$ but can still improve the loss by reducing decoder mismatch. Conversely, preserving information alone does not ensure an improvement: take constant S, uniform binary $Y = V$, and $Z = V$. A frozen readout assigning probability 0.1 to the correct label has loss $-\log 0.1$, worse than the no-prefix uniform predictor's $\log 2$, despite preserving all of the one-bit answer information.

These observations motivate correct-image and wrong-image controls, which test whether behavior depends on image-specific content. Such interventions do not estimate either the mutual information or the KL term. Likewise, Equation 13 does not imply that F1 increases monotonically with T_{pre}, that supervised prefixes dominate pruning, or that the observed gains transfer to other tasks.

J.3 A FINITE-PRECISION CAPACITY BOUND

A position budget becomes a bit-capacity constraint only after precision is specified. Suppose T_{pre} and d are fixed, each coordinate of Z has at most 2^b possible stored values (b bits per coordinate), and no additional image-dependent code is passed outside Z and the declared side input. Thus, with $\mathcal{Z}$ denoting the set of possible stored prefixes, $|\mathcal{Z}| \le 2^{bT_{\text{pre}}d}$.

Corollary 1 (Necessary capacity for retained information)**.** *Under these assumptions,*

$$I(Y; Z \mid S) \le \min\{I(Y; V \mid S),\ bT_{\text{pre}}d \log 2\}. \tag{14}$$

For a fixed list of questions, let $A = (Y_1, \ldots, Y_m)$ be their answer vector, where m is the number of questions, and let R be shared side information. If A can be recovered exactly from (R, Z) almost surely, then

$$H(A \mid R) \le bT_{\text{pre}}d \log 2. \tag{15}$$

Proof. Determinism gives $I(Y; Z \mid S, V) = 0$. Applying the chain rule in two orders,

$$I(Y; V \mid S) = I(Y; Z \mid S) + I(Y; V \mid S, Z) \ge I(Y; Z \mid S). \tag{16}$$

Also $I(Y; Z \mid S) \le H(Z \mid S) \le H(Z) \le \log|\mathcal{Z}|$. Exact recovery gives $H(A \mid R, Z) = 0$, so $H(A \mid R) = I(A; Z \mid R) \le H(Z) \le bT_{\text{pre}}d \log 2$. □

The joint entropy in Equation 15 allows dependent answers; it is not their summed marginal entropy. This is a necessary condition, not an achievable rate for a trained adapter or a frozen decoder. At practical embedding widths it can be loose and does not select a useful T_{pre} or predict our accuracy curve. Without finite precision, counting real-valued positions alone gives no finite Shannon-capacity bound.

J.4 CAUSAL CACHE EQUIVALENCE

We state the conditions behind Proposition 1 explicitly. The decoder is in deterministic evaluation mode with fixed weights. Each layer uses causal attention and otherwise tokenwise operations, such as RMS Norm and an MLP. The input at prefix positions $0, \ldots, T_{\text{pre}} - 1$ depends only on the image; any shared prelude is fixed and included in the cached segment. Attention masks, positional indices, and positional-encoding parameters at these positions are unchanged by the continuation. In particular, suffix-dependent rotary rescaling would violate this condition. All question-dependent inputs follow the cached segment, and warm and cold executions use the same continuation embeddings, masks, and position rules.

Proof of Proposition 1. Let $h_i^{(\ell)}$ be the state at position i after layer ℓ; j indexes a position attended to by i. Here i indexes sequence positions, rather than questions as in Figure 1(c). At layer zero, $h_i^{(0)}$ for $i < T_{\text{pre}}$ is fixed by the image-only prefix and its positional metadata, independently of the question. Assume that the states $h_i^{(\ell-1)}$ at all prefix positions are independent of the suffix. The layer's key and value projections at these positions are functions only of these states and the fixed

Table 17: **Recognition quality and visual KV storage on TextOCR.** BF16 KV tensor storage attributable to visual positions, derived from actual position counts and decoder cache geometry. T_{vis} counts decoder visual input positions. Text and generated-token KV, model weights, and activations are excluded. Full-resolution storage is averaged across images.

Method	TextOCR word F1	Visual KV (MiB)
Full-resolution input	0.548	105.012
VScan-I ($T_{\text{vis}} = 24$)	0.154	3.375
Learned prefix ($T_{\text{vis}} = 24$)	0.386	3.375
VScan-I ($T_{\text{vis}} = 64$)	0.292	9.000
Learned prefix ($T_{\text{vis}} = 64$)	0.444	9.000
VScan-I ($T_{\text{vis}} = 128$)	0.389	18.000
Learned prefix ($T_{\text{vis}} = 128$)	0.463	18.000

positional metadata. For $i < T_{\text{pre}}$, causal attention reads only positions $j \le i < T_{\text{pre}}$; its output therefore cannot depend on the suffix. Residual connections and tokenwise operations preserve this property. Induction over layers proves that every prefix key and value is question-independent, and that prefilling the prefix alone yields exactly the prefix slice of a full prefill.

Now consider a fixed continuation. At the first layer, its warm and cold input states coincide, and the stored prefix keys and values equal those from recomputation. Each continuation position consequently attends to identical keys and values under an identical mask, giving identical attention outputs and next-layer states. Induction over layers yields identical continuation logits. Applying the same argument after any fixed generated history establishes identical next-token distributions, hence the same autoregressive distribution. Greedy outputs coincide with fixed tie-breaking; sampled outputs have the same distribution, without requiring uncoupled random samples to be identical. □

Why an image-only embedding is not enough after the prompt. A two-layer scalar causal-attention example suffices. Let the scalar question and prefix inputs be $[q, z]$ with $z = 0$, zero query and key projections (uniform attention), identity value and output projections, a residual connection, and no other transformations. At the prefix position after the first layer,

$$h_z^{(1)} = z + \tfrac{1}{2}(q + z) = \tfrac{1}{2}q. \tag{17}$$

The second layer's cached value at that position is therefore $q/2$. Changing q changes the prefix KV despite identical z and identical positional indices. Special masks or weights can remove this dependency; prefix-first placement is a sufficient guarantee, not a necessary condition for every possible decoder.

Implementation and numerical scope. Applying the proposition requires the prefix to be computed without question-dependent inputs. Prefix attention masks and positional metadata must match full recomputation, and continuation inputs must use matching masks and position conventions. Each question must start from the same unmodified prefix cache.

The proposition establishes equivalence in exact arithmetic. Floating-point execution can introduce numerical differences. In our bf16 experiment, 91 of 96 image–question pairs produced identical greedy outputs. This observation neither establishes bitwise equivalence nor determines the cause of the remaining differences. Additional numerical checks on a small causal decoder supplement the proof but do not establish equivalence for the evaluated VLM.

K Additional Capacity, Supervision, and KV Results

Visual-KV accounting. For Qwen3-VL-4B, the bf16 decoder cache has 36 layers, 8 KV heads, and head dimension 128. Both keys and values are stored at 2 bytes per element, so each decoder position occupies

$$2 \times 36 \times 8 \times 128 \times 2 = 147{,}456 \text{ bytes} = 0.140625 \text{ MiB}.$$

Multiplying by the actual visual-position count T_{vis} gives $0.140625\, T_{\text{vis}}$ MiB for the visual component reported in Table 17. The full-input mean is 746.752 positions; compressed models use exactly 24,

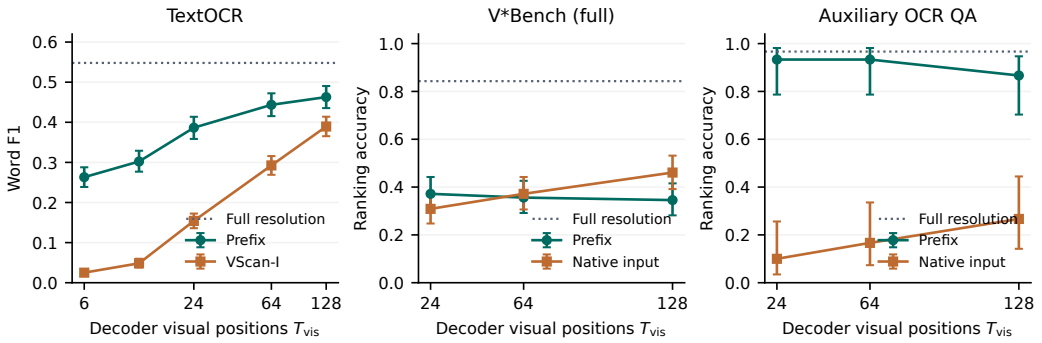

Figure 10: **Effect of prefix capacity across tasks.** The OCR-trained prefix is evaluated on TextOCR, the full standard V*Bench benchmark, and the auxiliary OCR QA set. TextOCR uses word F1; the QA panels use candidate-ranking accuracy. Each prefix length is trained independently. Native input denotes the frozen model's resolution-controlled visual interface; dotted lines show full-resolution references. Error bars are pointwise 95% intervals (image bootstrap for F1 and Wilson intervals for ranking accuracy). Generation accuracy is reported separately in Tables 3 and 18.

64, or 128. These derived tensor sizes exclude text and generated-token KV, model weights, and activations. They are not whole-process GPU memory or a measurement of cross-question retained state. Cache geometry was checked against the returned cache tensors for all 500 images in each of the seven reported arms. Quality and position counts use the same held-out TextOCR subset.

Capacity and matched supervision. The added OCR models use independently trained pure-prefix checkpoints fixed at epoch 10. Candidate ranking selects the answer with the lowest mean conditional token NLL. Generation is scored with the strict answer-letter parser, counting invalid responses as incorrect. The full standard V*Bench benchmark and the auxiliary OCR QA set are evaluated separately, as defined in Section 4.1. Table 18 reports ranking and generation accuracy at every tested capacity.

The mixed $T_{\text{pre}} = 64$ control uses 10,875 TextOCR and 10,874 GQA examples per epoch, for 21,749 examples and 10 epochs, matching the OCR-only run's example and update counts. Architecture, seeds, optimizer, schedule, gradient accumulation, encoder cap, and fixed-epoch rule are unchanged. Source examples are selected by a fixed hash rule without model scores; V* examples are not used for training or checkpoint selection. Table 19 gives paired differences; these intervals do not measure variation across training seeds.

Search comparisons. Table 20 reports generation accuracy on the full standard V*Bench benchmark for direct input and ZoomEye search. Timing includes cue, search, and answer calls; cumulative decoder positions include all calls. The search arm has an open visual budget, so it is not a matched-position compression control. Table 21 separately reports OCR quality and visual budgets. The default OCR cue parser falls back to full-image answering on all 500 images. The compatible parser enables search; sharing the task-level cue produces identical transcriptions on all 500 images and is reported as a separate execution variant. These OCR measurements use the same transcription scorer and cohort as Table 1.

Exposure sensitivity. One image in the standard V*Bench benchmark appears in an official demonstration. Excluding it leaves 190 images; generated-answer counts are 151/190 for full input, 175/190 for ZoomEye, 9/190 for OCR-only $T_{\text{pre}} = 64$, and 59/190 for mixed $T_{\text{pre}} = 64$. The full standard benchmark remains the primary evaluation; this single-image sensitivity check does not establish absence of prior exposure elsewhere.

Table 18: **QA accuracy across visual budgets and training objectives.** The full standard V*Bench benchmark and the auxiliary OCR QA set are evaluated separately. All prefix models here use $T_{\mathrm{nat}} = 0$, so their total visual budget is $T_{\mathrm{vis}} = T_{\mathrm{pre}}$. Ranking and generation follow the definitions in Section 4.1; all accuracies are on the 0–1 scale.

	V*Bench (full)		Auxiliary OCR QA	
Model	Ranking	Generation	Ranking	Generation
Native input ($T_{\mathrm{vis}} = 24$)	0.309	0.356	0.100	0.333
OCR prefix ($T_{\mathrm{vis}} = 24$)	0.372	0.026	0.933	0.000
Native input ($T_{\mathrm{vis}} = 64$)	0.372	0.408	0.167	0.333
OCR prefix ($T_{\mathrm{vis}} = 64$)	0.356	0.047	0.933	0.000
Native input ($T_{\mathrm{vis}} = 128$)	0.461	0.466	0.267	0.400
OCR prefix ($T_{\mathrm{vis}} = 128$)	0.346	0.010	0.867	0.000
Full-resolution input	0.843	0.791	0.967	0.933
Mixed prefix ($T_{\mathrm{vis}} = 64$)	0.429	0.309	0.933	0.267

Table 19: Paired differences in TextOCR word F1 or V*Bench accuracy, on the 0–1 scale. Intervals use paired image bootstrap and do not cover training-seed variation.

Comparison	Difference	Paired 95% interval
Prefix vs VScan-I ($T_{\mathrm{vis}} = 64$)	+0.1511	[+0.1269, +0.1765]
Prefix vs VScan-I ($T_{\mathrm{vis}} = 128$)	+0.0734	[+0.0493, +0.0977]
Prefix $T_{\mathrm{pre}} = 64$ minus 24	+0.0570	[+0.0372, +0.0758]
Prefix $T_{\mathrm{pre}} = 128$ minus 64	+0.0193	[-0.0002, +0.0386]
QA supervision, ranking	+0.0733	[+0.0000, +0.1466]
QA supervision, generation	+0.2618	[+0.1937, +0.3298]

Table 20: **Direct answering and adaptive search on V*Bench.** Generation accuracy on the full standard benchmark. Valid is the fraction of responses parsed as answer letters; invalid responses count as incorrect. T_{vis} counts final decoder visual inputs; all prefix models here use $T_{\mathrm{nat}} = 0$. Enc. counts vision-encoder calls, and T_{cum} counts cumulative decoder input positions. All cue, search, and answer calls contribute to the reported costs.

Model	Gen. acc.	Valid	s/query	Enc.	T_{cum}
Full-resolution input	0.791	1.000	0.386	1.00	2012
Native input ($T_{\mathrm{vis}} = 24$)	0.356	1.000	0.148	1.00	102
Native input ($T_{\mathrm{vis}} = 64$)	0.408	1.000	0.148	1.00	142
Native input ($T_{\mathrm{vis}} = 128$)	0.466	1.000	0.154	1.00	206
OCR prefix ($T_{\mathrm{vis}} = 24$)	0.026	0.110	0.861	1.00	114
OCR prefix ($T_{\mathrm{vis}} = 64$)	0.047	0.126	0.899	1.00	156
OCR prefix ($T_{\mathrm{vis}} = 128$)	0.010	0.031	0.799	1.00	216
ZoomEye	0.921	1.000	7.095	33.85	28666

Table 21: **Adaptive visual search on TextOCR.** The same held-out subset and word-F1 metric as Table 1. T_{vis} is the mean final decoder visual positions, T_{cum} the mean cumulative decoder input positions, and Enc. the mean vision-encoder calls. Times include cue, search, and answer processing; parser and cue-sharing variants are reported separately.

Method	Word F1	T_{vis}	T_{cum}	Enc.	s/image
Full-resolution input	0.548	746.8	868	1.00	3.386
ZoomEye (default parser)	0.548	746.8	1553	1.00	8.478
ZoomEye (compatible parser)	0.491	549.3	7771	13.27	8.340
ZoomEye (shared cue)	0.491	549.3	7087	13.27	3.152