Query-Aware Multimodal Context Compression

Initial evidence for allocating visual context by task, time, and fidelity

1 Motivation and prior work

Long-context systems increasingly ingest PDFs, screenshots, images, and video alongside text. This improves coverage but makes the visual context expensive: dense image grids and repeated video frames create many tokens whose relevance depends on the downstream query. Text compression already treats this as an information-selection problem. LLMLingua-2, for example, distills supervision from a strong LLM and trains a small encoder to classify which text tokens should be retained, reducing context while preserving task performance [1].

Visual compression is now an active area as well. Recent methods prune redundant visual tokens, use language guidance to retain query-relevant evidence, or adapt the pruning budget by instance and layer [2, 4]. For video, PruneVid removes spatial/temporal redundancy and query-irrelevant tokens [3], while Vista-LLM explicitly introduces query-guided dynamic budgeting and temporal allocation before inference [5]. These results establish that visual inputs are highly compressible. The open opportunity is broader: a *pre-API, model-agnostic* system that decides not only which evidence to keep, but also *where and at what fidelity* to spend a limited visual-token budget across pages, regions, and frames. Modern APIs already expose useful control surfaces—for example, Gemini 3 allows per-media resolution settings that directly trade visual tokens for detail [6].

2 Initial experiment: visual budget vs. task accuracy

I tested whether visual-token allocation can be reduced without losing the information needed for a concrete multimodal task. The benchmark contains 138 visually iconic sign videos and a 10-choice meaning-identification task (chance = .10). The prompt and candidate meanings were held fixed; only the visual evidence was compressed. I evaluated Qwen3-VL-32B-Instruct and Gemini 3.6 Flash with thinking disabled.

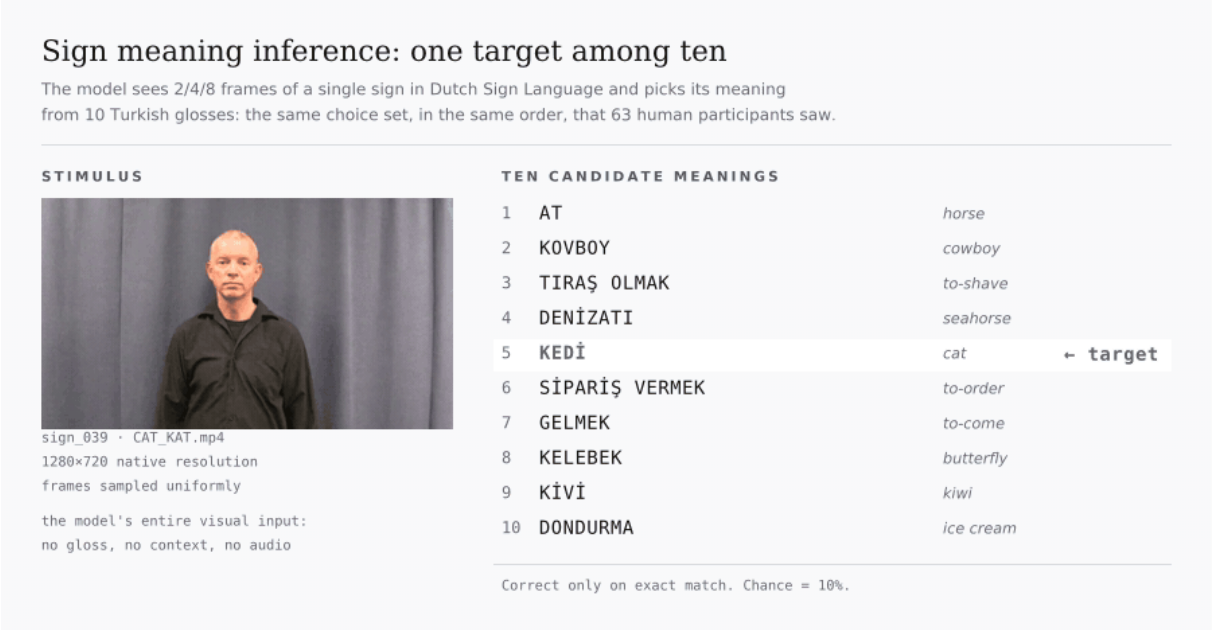


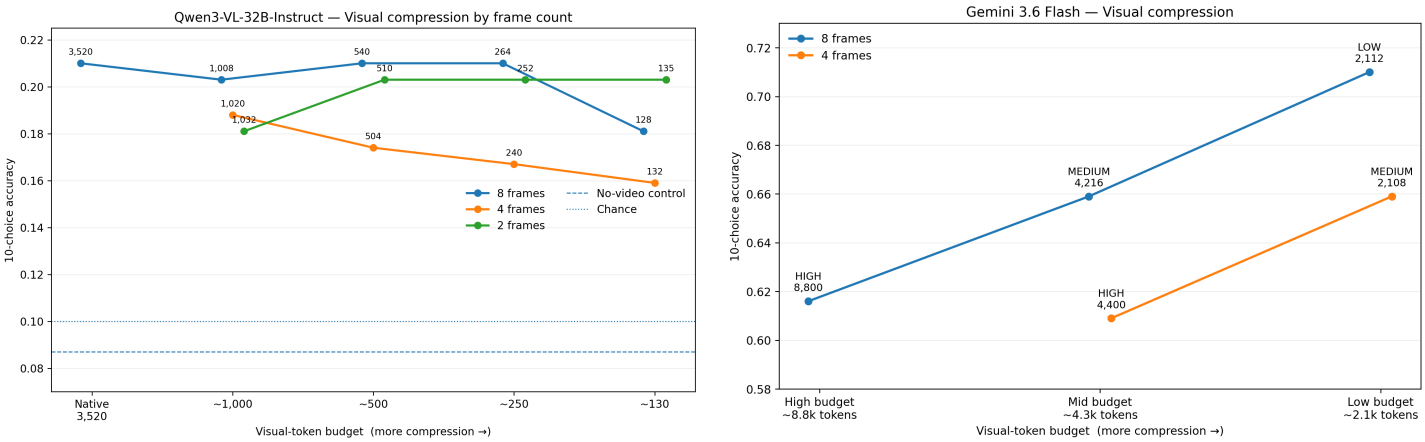
Figure 1: Task illustration for the 10-choice meaning-identification experiment. The source package includes the linked animation `task_figure.gif`; the compiled PDF shows its first frame as a static preview.

Compression was manipulated along two independent axes: **temporal coverage** (8, 4, or 2 sampled frames) and **spatial fidelity / token budget**. For Qwen, frames were resized before the processor to target a total visual-token budget. For Gemini, the API's media-resolution setting controlled the per-frame visual allocation. Thus, matched total budgets can be spent on many low-detail frames or fewer high-detail frames.

Qwen: large redundancy without measurable loss. Native 8-frame input used 3,520 visual tokens and reached .210 accuracy. At only 264 tokens, the same 8-frame condition remained at .210—a **13.3× reduction** with no observed loss. A no-video control (.087) and mismatched-video control (.080) were significantly below native performance, confirming that the model was using visual evidence rather than only answer priors. The 2-frame condition also stayed near .203 down to roughly 135 tokens. Because Qwen's absolute accuracy is low, this is best interpreted as evidence of substantial redundancy rather than as a claim that extreme compression is universally optimal.

Gemini: compression can improve accuracy. With 8 frames, reducing observed image tokens from 8,800 (HIGH) to 2,112 (LOW) produced a **4.17× reduction** while accuracy increased from .616 to .710 (**+9.4 percentage points**; paired

McNemar $p = .011$). At nearly the same total budget (∼2.1–2.2k tokens), allocation also mattered: 8f/LOW achieved .710, 4f/MEDIUM .659, and 2f/HIGH .638. For this task, broader temporal coverage appears more valuable than maximizing per-frame spatial detail.



(a) Qwen3-VL-32B-Instruct **(b) Gemini 3.6 Flash**

Figure 2: Initial visual-context compression results. Qwen preserves accuracy under substantial compression; Gemini improves as visual allocation is reduced. Exact realized/observed token counts are annotated; no error bars are shown.

3 Moving forward: query-aware multimodal allocation

The results suggest a more useful objective than "prune as many visual tokens as possible." Visual context compression should be treated as **information allocation**: given a query and a budget, preserve the cheapest representation that retains the evidence needed for the answer.

A practical system can sit before any downstream VLM:

$$\textbf{query} + \textbf{multimodal context} \rightarrow \textbf{lightweight controller} \rightarrow \textbf{compressed context} \rightarrow \textbf{VLM}$$

The controller's action space can include: (i) KEEP/DROP pages, images, regions, or frames; (ii) LOW/MEDIUM/HIGH fidelity; (iii) query-relevant crops or regions of interest; (iv) OCR/text-only substitution when pixels add little; and (v) dynamic allocation between temporal coverage and spatial detail. This keeps the system deployable with proprietary APIs because it modifies the *input*, rather than requiring access to internal visual embeddings.

Learning signal. For each training example, generate candidate compressed views under different actions and budgets, evaluate them with a strong teacher/downstream VLM, and select the cheapest view that preserves the target answer or score. A compact encoder/controller can then learn this policy from the query plus cheap visual features. A natural objective is

$$\mathcal{L} = \mathcal{L}_{task} + \lambda \mathcal{L}_{distill} + \beta \mathcal{L}_{budget},$$

where the budget term directly rewards lower billed visual input while the task/distillation terms prevent destructive compression.

Next validation. The key test is whether the optimal allocation changes with the information need. On the same visual inputs, compare meaning recognition with handshape, location, movement/path, fine-detail/OCR, and document-style questions. Then benchmark a learned policy against fixed LOW resolution, uniform resizing, random/uniform frame sampling, and simple saliency. Success should be reported as a Pareto frontier: **task quality vs. billed visual tokens, latency, and cost**. Cross-model transfer to multiple open and proprietary VLMs would establish the strongest practical value: one query-aware compression layer that reduces multimodal inference cost while preserving—and in some regimes improving—downstream quality.

References

[1] Z. Pan et al. *LLMLingua-2: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression*. Findings of ACL, 2024.
[2] X. Ye et al. *ATP-LLaVA: Adaptive Token Pruning for Large Vision Language Models*. CVPR, 2025.
[3] X. Huang, H. Zhou, and K. Han. *PruneVid: Visual Token Pruning for Efficient Video Large Language Models*. Findings of ACL, 2025.
[4] H. Zhang et al. *TrimTokenator: Towards Adaptive Visual Token Pruning for Large Multimodal Models*. Findings of ACL, 2026.
[5] Z. Li et al. *Vista-LLM: Decoupled Query-Guided Visual Token Pruning for Efficient Long-Video Large Language Models*. ACL, 2026.
[6] Google. *Gemini API: Media resolution*. Documentation, updated July 2026.