VIDEO UNDERSTANDING AT A GLANCE: WHAT REMAINS AT TWO TOKENS PER FRAME?

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied and wearable systems often need a fast, low-cost understanding of ongoing visual experience before finer-grained perception or action is required. We formulate this need as *glance video understanding*: forming an overall understanding of scene context, salient events, and their temporal evolution—rather than exhaustively recovering every visual detail—at low visual cost. Preserving every frame as a distinct unit of visual evidence, we ask whether such understanding can still be sustained when the available visual-token budget is pushed to an extreme. Unlike an isolated image, a video frame carries both its *current visual state* and its *temporal contribution* relative to preceding observations. Compressing both into a single token would force these two aspects to compete within the same representational slot; we therefore set the minimum per-frame budget to two tokens—one for *what is present* and one for *what is changing*. This leads to our central question: *what remains of video understanding under this purely visual, query-independent boundary?* We investigate this question with **GlanceVLM**, which pairs an **Anchor Token** for global semantic context with an **Innovation Token** for temporally distinctive evidence. Difference-guided innovation encoding uses inter-frame changes to select informative regions while retaining their current-frame visual content, allowing both tokens to be formed before any downstream question is known. Across eight benchmarks, GlanceVLM reaches $\mathbf{63.6}\%$ on MLVU and $\mathbf{50.0}\%$ on our 2,400-question **GlanceVU**, showing that substantial video-understanding capability survives rather than collapses at 2 tokens/frame.

1 INTRODUCTION

Embodied and wearable systems often need a fast, low-cost overview of the current scene and its recent history before deciding how to act or what to inspect in detail (Yang et al., 2025; Pan et al., 2025). We term this objective **glance video understanding**: forming an overall understanding of scene context, salient events, and their temporal evolution at low visual cost, without requiring exhaustive detail recovery. For video MLLMs, we measure visual cost by the number of visual tokens supplied to the language model (LLM), which affects its computation and memory use (Tao et al., 2025). Such an overview can guide subsequent planning and direct closer inspection to a particular frame or spatial region. For this task, we seek a representation that can be formed *before a downstream question is known*. Query-independent encoding allows the same video representation to be prepared in advance and reused across later questions, rather than tailored to a single one.

Representative approaches reveal three limitations relevant to this low-cost, query-independent understanding (Fig. 1). **(i) Strong long-video performance often comes at a high visual-token cost.** Video MLLMs (Bai et al., 2025a; Wang et al., 2025b; Team et al., 2026; Chen et al., 2025) supply the LLM with large numbers of visual tokens to preserve detail across frames, making even an initial overview costly. Query-independent methods such as TimeChat-Online (Yao et al., 2025a) reduce this burden by dropping visual tokens that are highly similar across adjacent frames. However, our budget sweep shows that accuracy generally falls across VideoMME (Fu et al., 2024) subtasks as the budget shrinks, with some capabilities approaching the text-only baseline with zero visual tokens (**a**). **(ii) Question-guided encoding requires the question in advance.** LLaMA-VID (Li et al., 2024c) uses the question to construct its context token, while Tempo (Fei et al., 2026) forms question-conditioned video memories and allocates tokens according to question relevance. In our diagnostic, restricting Tempo's visual-token construction to video alone reduces VideoMME ac-

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

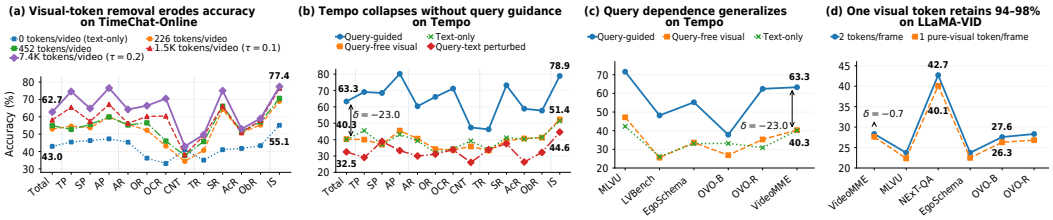


Figure 1: **Diagnostics of visual-token cost, question dependence, and text sensitivity. (a) Reducing visual tokens loses accuracy.** TimeChat-Online (Yao et al., 2025a) exhibits accuracy losses as its visual budget shrinks on VideoMME. **(b,c) Tempo relies on question guidance.** Restricting visual-token construction to video alone brings Tempo (Fei et al., 2026) close to its text-only baseline across VideoMME subtasks and 5 benchmarks. **Question text perturbations also reduce accuracy** to 32.5%. **(d) The question-conditioned token adds limited accuracy on LLaMA-VID.** LLaMA-VID (Li et al., 2024c) retains 93.8–97.5% of its original accuracy using only its purely visual content token. Together, these observations motivate a question: *can useful video understanding survive when **purely visual**, **query-independent** tokens are forced into an extreme token budget?*

curacy from 63.3% to 40.3%, near the question text-only baseline, with a similar pattern across benchmarks (**b,c**). **(iii) Question-guided performance can deteriorate under text perturbations.** Even when the question remains available, its textual form can matter: in a separate Tempo stress test, we apply synonym replacement and textual noise to the questions and answer options while keeping the original content recoverable and the answer labels unchanged. Accuracy falls to 32.5% (**b**), exposing sensitivity to the question text under this perturbation setting.

Our LLaMA-VID diagnostic provides a complementary observation: removing its question-guided context token and retaining only its purely visual content token preserves 93.8–97.5% of its original accuracy across 5 benchmarks, although absolute accuracy remains low (**d**). These limitations and observations motivate a stricter boundary question: if the visual representation must be *purely visual and query-independent*, while its LLM-facing cost is constrained to an extreme fixed token budget, how much video understanding would remain? To study this extreme boundary while retaining every frame as an individual source of visual evidence, we measure the budget at the *frame level*, treating each frame as a distinct unit rather than merging or discarding some of them. Unlike an isolated image, a video frame carries not only its *current visual state*, but also temporal information about *how that state is changing* relative to preceding observations. A minimal frame representation should therefore preserve both *what is present* and *what is changing*. If a token is regarded as an independent unit of information exposed to the LLM, collapsing these two aspects into a single token forces them to compete for the same representational slot. We therefore set the minimum budget to *two visual tokens*. This leads to our central question: ***what remains of video understanding when every observed frame must be represented by just two purely visual, query-independent tokens?***

We investigate this question with GlanceVLM, whose dual-token frame encoding fixes the visual budget and separates video representation from downstream text. **(i) Represent every observed frame with two complementary tokens.** An **Anchor Token** summarizes the current frame's global semantic context, while an **Innovation Token** gathers temporally distinctive evidence from selected regions. Rather than retaining two existing patch tokens, the model learns two role-specific summaries that exchange context while reading their respective visual evidence. **(ii) Select visual evidence without knowing the question.** Our *difference-guided innovation encoding* uses inter-frame feature changes to actively identify informative regions and distills their current-frame visual content into the Innovation Token. Temporal context guides *where to read*; the current frame supplies *what to retain*. Unlike LLaMA-VID, its image-wise token construction does not model the video's temporal structure. **(iii) Visual tokens remain unchanged.** Neither the 2 visual tokens uses the guidance of question text, audio, and LLM-side attention during its construction. So for a fixed video, changes to the question or answer options leave the encoded visual tokens unchanged.

We evaluate GlanceVLM-4B/8B across eight established benchmarks covering general video QA, long-video understanding, and timestamp-bounded understanding; each timestamped query is evaluated offline using only the video prefix available before its timestamp. The 8B model reaches **63.6**% on MLVU (Zhou et al., 2025) and **58.2**% on VideoMME (Fu et al., 2024). To complement these broad evaluations with questions targeting glance-style understanding, we curate GlanceVU, a *2,400-question benchmark* derived from existing video-QA datasets. On GlanceVU, GlanceVLM-

8B achieves $\mathbf{50.0}\%$, exceeding the matched two-token TimeChat-Online variant by 4.1 points. These results show that even under an extreme query-independent 2 tokens/frame bottleneck, substantial video-understanding capability survives rather than collapses, with GlanceVLM offering a concrete modeling approach for it. Our contributions are as follows:

- We formulate **glance video understanding and investigate its extreme boundary** at two purely visual, query-independent tokens per observed frame.
- We propose GlanceVLM, whose complementary **Anchor** and **Innovation** tokens combine global scene context with actively selected, difference-guided visual evidence, independently of downstream question text.
- Through eight established benchmarks and the curated GlanceVU, we characterize both the understanding retained and the limitations exposed by this extreme visual budget.

2 Related Work

Resource-Constrained Video Understanding. Real-world video understanding increasingly requires causal processing in streaming, on-device, and embodied settings under stringent latency, bandwidth, memory, and energy budgets, together with privacy requirements (Sharshar et al., 2025; Shinde et al., 2025; Xu et al., 2023). Long-video benchmarks increasingly demand long-horizon temporal reasoning (Yu et al., 2019; Mangalam et al., 2023; Fu et al., 2024; Wu et al., 2024; Zhou et al., 2025), while video MLLMs primarily extend temporal coverage through longer contexts, visual compression, and memory mechanisms (Song et al., 2024; Li et al., 2024c; Zhang et al., 2024a; Li et al., 2024b; Bai et al., 2025a; Yao et al., 2024; 2025b; Xu et al., 2025; Li et al., 2026; An et al., 2026). Streaming methods further couple causal observation with online management of ever-growing visual-token histories (Lin et al., 2026; Niu et al., 2025; Chen et al., 2024a; Yao et al., 2025a). Collectively, these advances in representing and managing long visual histories at varying compression levels recast resource-constrained video understanding as a new question: *when the visual-token budget is pushed to its limit, how much of the information required for video understanding can the resulting compressed representation still encode?*

Query-Guided and Query-Free Visual Compression. Long-video compression methods approach this question by suppressing visual redundancy or retaining query-relevant evidence. TimeChat-Online is query-free, removing 82.8% of streaming video tokens through inter-frame redundancy filtering (Yao et al., 2025a). Text-guided alternatives include VideoChat-Flash, which compresses 729 patch tokens to an average of 16 per frame before text-guided dropout; LongVU, which combines visual frame deduplication with text-guided feature reduction; and methods that select tokens or temporal evidence using questions, answer options, or cross-modal relevance (Li et al., 2024b; Shen et al., 2024; Huang et al., 2025; Wang et al., 2024b; Zhang et al., 2025; Wu et al., 2026). Tempo further allocates 0.5–16 tokens per frame through query-conditioned cross-modal distillation (Fei et al., 2026). At more extreme compression, LLaMA-VID pairs a visual content token with an instruction-conditioned context token per frame, whereas XComp progressively reduces each frame to one token across LLM layers and separately applies question-conditioned frame selection (Li et al., 2024c; Zhang et al., 2026). Across this spectrum, existing methods primarily optimize how visual evidence is pruned, condensed, selected, or allocated. The complementary boundary is: *what video understanding remains possible when each frame is represented by only a few purely visual, query-independent tokens that must preserve information for unknown future inference?*

3 GlanceVLM: Dual-Token Frame Encoding

Overview. Under a fixed two-token budget, we divide visual capacity between *holistic context* and *localized evidence of change*. Given a video $\mathcal{V} = \{I_t\}_{t=1}^{T}$, we use T as the LLM-facing frame count, irrespective of how each model's native visual encoder processes the sampled video. Frame I_t provides LLM-aligned spatial features $\mathbf{X}_t = [\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,N}]^\top \in \mathbb{R}^{N\times d}$, where N is the spatial-token count and d the LLM hidden size. We encode $\mathbf{X}_t$ using two role-specialized query tokens, whose final states define the **Anchor token** $\mathcal{A}_t$ and **Innovation token** $\mathcal{I}_t$: $\mathbf{Z}_t = [\mathcal{A}_t, \mathcal{I}_t]^\top \in \mathbb{R}^{2\times d}$. During refinement, the Anchor query reads the complete frame (Sec. 3.1), whereas the Innovation query reads regional descriptors routed by adjacent-frame feature differences (Sec. 3.2): the differences determine *where* to read, while current-frame features determine *what* to retain. The two query tokens exchange context before their role-specific readouts (Sec. 3.3). Finally,

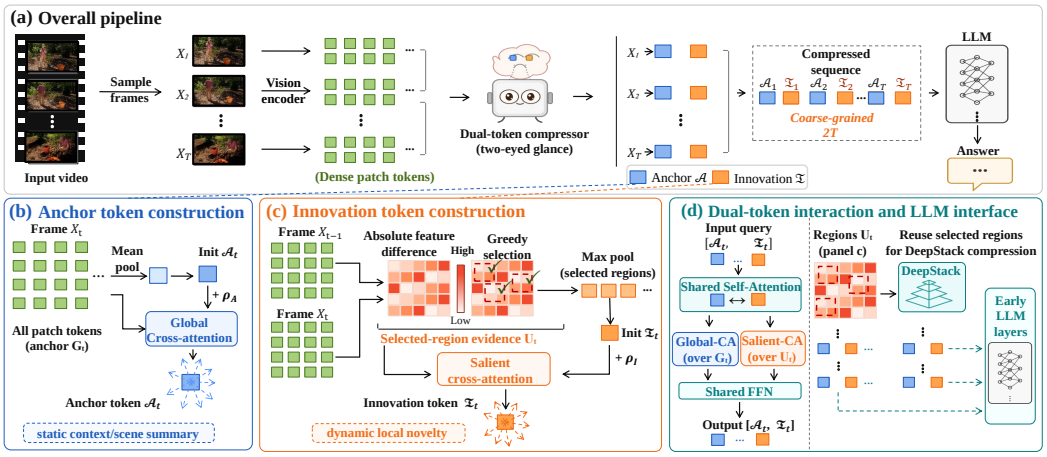


Figure 2: **Two-eyed glance tokens for fixed-budget video understanding. (a) Overall pipeline.** Dense patch tokens from T sampled frames are transformed into a role-specialized sequence of only $2T$ Anchor–Innovation tokens for LLM inference. **(b,c) Role-specialized token formation.** Mean pooling initializes the Anchor query while all frame patches constitute its full-frame memory; adjacent-frame differences localize distinctive regions, whose union-wise and region-wise pooling produce the Innovation query and region memory, respectively. **(d) Dual-token interaction.** Shared self-attention and role-specific cross-attention refine the final token pair, while the selected regions are reused for paired multi-level compression and early-layer LLM injection.

$\mathbf{Z} = \mathrm{Concat}_{t=1}^{T} \mathbf{Z}_t \in \mathbb{R}^{2T \times d}$ reduces NT visual positions to $2T$, providing $N/2$-fold greater visual coverage under the same LLM-facing visual budget.

3.1 Anchor Token as a Holistic Frame Reference

Full-frame anchoring. Change-selected regions alone may omit ***the context needed to interpret what they depict***. We therefore form a frame-wide seed $\bar{\mathbf{x}}_t = N^{-1} \sum_{j=1}^{N} \mathbf{x}_{t,j} \in \mathbb{R}^d$, and use it as the data-dependent initialization of the Anchor query. Because every spatial position contributes equally, $\bar{\mathbf{x}}_t$ provides a holistic starting point, but it is *not* the final Anchor token $\mathcal{A}_t$. Instead, it initializes a learned query that subsequently interacts with the full frame and is iteratively refined.

Anchor initialization. Using $\bar{\mathbf{x}}_t$ as its data-dependent seed, we initialize the hidden state of the Anchor query token as

$$\mathbf{q}_{t,\mathrm{A}}^{(0)} = \bar{\mathbf{x}}_t + \mathbf{p}_{\mathrm{A}} + \boldsymbol{\rho}_{\mathrm{A}} + \boldsymbol{\tau}_t, \tag{1}$$

where $\mathbf{p}_{\mathrm{A}} \in \mathbb{R}^d$ is the learned base embedding for the Anchor slot, shared across frames; $\boldsymbol{\rho}_{\mathrm{A}} \in \mathbb{R}^d$ is its learned role embedding; and $\boldsymbol{\tau}_t \in \mathbb{R}^d$ encodes temporal position. During refinement, the evolving Anchor state $\mathbf{q}_{t,\mathrm{A}}$ serves as the query input to Anchor-specific cross-attention, while the frame $\mathbf{X}_t$ provides its key/value memory, $\mathbf{K}_{t,\mathrm{A}} = \mathbf{V}_{t,\mathrm{A}} = \mathbf{X}_t$. Thus, repeated full-frame readout can recover evidence beyond the initial average $\bar{\mathbf{x}}_t$ before producing the final Anchor token $\mathcal{A}_t$.

3.2 Innovation Token from Residual-Routed Evidence

Change-salient spatial selection. ***Once the current visual state is anchored, the complementary information that makes a video temporal lies in how that state evolves across frames.*** Under the two-token bottleneck, the Innovation query therefore focuses on visual evidence of such changes rather than forming another generic frame summary. For adjacent frames sharing an $H \times W$ feature grid ($HW = N$), we measure change at position j as $\delta_{t,j} = \|\mathbf{x}_{t,j} - \mathbf{x}_{t-1,j}\|_2$ for $t > 1$; for $t = 1$, we use $\|\mathbf{x}_{1,j}\|_2$. The scores are reshaped onto the spatial grid and smoothed with a 3×3 average filter to obtain the routing map $\widetilde{\mathbf{R}}_t$, favoring spatially coherent changes over isolated responses. ***The change signal determines where the Innovation query reads, while current-frame features determine what it retains.***

Regional evidence formation. Greedy peak selection on $\widetilde{\mathbf{R}}_t$ yields $M_t = 4$ regions, each defined by a 5×5 window clipped at the feature-grid boundary. We denote the flattened spatial-token indices of region m by $\Omega_t^m \subseteq \{1, \ldots, N\}$. After each selection, its window is suppressed from subsequent

peak-center search, although later windows may still overlap it. Rather than immediately collapsing all selected regions into the single Innovation slot, we preserve their pooled descriptors as separate candidate visual evidence:

$$\mathbf{u}_t^m = \mathrm{MaxPool}\left\{\mathbf{x}_{t,j} \mid j \in \Omega_t^m\right\}, \quad \mathbf{U}_t = [\mathbf{u}_t^1, \ldots, \mathbf{u}_t^{M_t}]^\top \in \mathbb{R}^{M_t \times d}, \quad \mathbf{K}_{t,\mathrm{I}} = \mathbf{V}_{t,\mathrm{I}} = \mathbf{U}_t. \tag{2}$$

Here, each $\mathbf{u}_t^m \in \mathbb{R}^d$ is a channel-wise max-pooled regional descriptor, and their stack $\mathbf{U}_t$ forms the key/value memory for Innovation-specific cross-attention. Importantly, residual scores only select the evidence support; the retained visual content still comes from the current-frame features.

Innovation initialization. The single Innovation query is initialized from the combined support of the selected regions, while their individual descriptors remain available for subsequent readout. Specifically, let $\Omega_t^{\cup} = \bigcup_{m=1}^{M_t} \Omega_t^m$ and define the channel-wise pooled seed $\mathbf{u}_t^{\cup} = \mathrm{MaxPool}\{\mathbf{x}_{t,j} \mid j \in \Omega_t^{\cup}\}$, equivalently $\mathbf{u}_t^{\cup} = \mathrm{MaxPool}\{\mathbf{u}_t^m \mid 1 \le m \le M_t\}$. We initialize the hidden state of the Innovation query token as

$$\mathbf{q}_{t,\mathrm{I}}^{(0)} = \mathbf{u}_t^{\cup} + \mathbf{p}_{\mathrm{I}} + \boldsymbol{\rho}_{\mathrm{I}} + \boldsymbol{\tau}_t, \tag{3}$$

where $\mathbf{p}_{\mathrm{I}} \in \mathbb{R}^d$ is the learned base embedding for the Innovation slot, shared across frames, $\boldsymbol{\rho}_{\mathrm{I}} \in \mathbb{R}^d$ is its learned role embedding, and $\boldsymbol{\tau}_t$ is the same temporal embedding. This two-level construction separates *output capacity* from *evidence diversity*: one Innovation query is initialized from the combined change-selected support, yet retains access to the individual regional descriptors during refinement. Its refined final state yields the Innovation token $\mathcal{I}_t$ (Sec. 3.3).

3.3 Role-Specialized Dual-Query Refinement

Role-preserving query exchange with asymmetric evidence. The Anchor and Innovation queries are designed from the outset for complementary visual roles through distinct initialization and evidence sources. Their representations should therefore remain role-specific while still allowing evidence acquired by one role to contextualize the other. For frame t, we stack $\mathbf{Q}_t^{(0)} = [\mathbf{q}_{t,\mathrm{A}}^{(0)}; \mathbf{q}_{t,\mathrm{I}}^{(0)}] \in \mathbb{R}^{2\times d}$ and apply $L = 2$ dual-query blocks. Each block first exchanges context between the two queries and then returns each to its role-specific visual evidence:

$$\begin{aligned} \widehat{\mathbf{Q}}_t^{(\ell)} &= \mathbf{Q}_t^{(\ell-1)} + \mathrm{MSA}^{(\ell)}\big(\mathrm{LN}_1^{(\ell)}(\mathbf{Q}_t^{(\ell-1)})\big), \quad \widetilde{\mathbf{q}}_{t,a}^{(\ell)} = \widehat{\mathbf{q}}_{t,a}^{(\ell)} + \mathrm{MCA}_a^{(\ell)}\big(\mathrm{LN}_2^{(\ell)}(\widehat{\mathbf{q}}_{t,a}^{(\ell)}), \mathbf{K}_{t,a}, \mathbf{V}_{t,a}\big), \\ \widetilde{\mathbf{Q}}_t^{(\ell)} &= [\widetilde{\mathbf{q}}_{t,\mathrm{A}}^{(\ell)}; \widetilde{\mathbf{q}}_{t,\mathrm{I}}^{(\ell)}], \quad \mathbf{Q}_t^{(\ell)} = \widetilde{\mathbf{Q}}_t^{(\ell)} + \mathrm{FFN}^{(\ell)}\big(\mathrm{LN}_3^{(\ell)}(\widetilde{\mathbf{Q}}_t^{(\ell)})\big), \qquad a \in \{\mathrm{A}, \mathrm{I}\}. \end{aligned} \tag{4}$$

Here, $\widehat{\mathbf{q}}_{t,a}^{(\ell)}$ denotes row a of $\widehat{\mathbf{Q}}_t^{(\ell)}$; $\widehat{(\cdot)}$ and $\widetilde{(\cdot)}$ denote the post-exchange and post-readout states, respectively. MSA and MCA denote multi-head self- and cross-attention, LN layer normalization, and FFN a position-wise feed-forward network. Self-attention is confined to the two queries of the same frame, while their evidence remains role-specific: $\mathbf{K}_{t,\mathrm{A}} = \mathbf{V}_{t,\mathrm{A}} = \mathbf{X}_t$ and $\mathbf{K}_{t,\mathrm{I}} = \mathbf{V}_{t,\mathrm{I}} = \mathbf{U}_t$. The two cross-attention modules are independently parameterized, whereas the normalization layers and FFN are shared across query positions. ***Each exchange lets evidence acquired by one role contextualize the other, after which both queries are repeatedly re-grounded in their designated visual sources.*** After L blocks, the final LLM-facing pair is $\mathbf{Z}_t = \mathrm{LN}_{\mathrm{out}}(\mathbf{Q}_t^{(L)}) = [\mathcal{A}_t; \mathcal{I}_t] \in \mathbb{R}^{2\times d}$, and the frame-level pairs are concatenated chronologically for long-range integration by the LLM.

4 Experiments

4.1 Experimental Setup

Implementation details. All experiments are conducted on 8× NVIDIA A100 GPUs. We train the model with a single supervised fine-tuning stage on the pure-video subset of Video-R1-Data (Feng et al., 2026), which contains 86,217 video-only training instances in our setup. Unless otherwise specified, videos are uniformly sampled at 1 fps. For Qwen-family backbones, we cap the per-frame visual resolution at 448×448. We optimize all model parameters end-to-end, using a base learning rate of 5×10^{-6}, a vision learning rate of 2×10^{-6}, and a merger learning rate of 1×10^{-5}, with a per-device batch size of 4. For Qwen3-VL (Bai et al., 2025a), we compress both the final and its DeepStack visual features to two tokens per frame and update the video RoPE grid accordingly, so the LLM observes exactly two visual positions per frame.

Model selection. Our model suite is designed to assess both generality and competitiveness. We instantiate our method on Qwen3-VL-4B/8B to test robustness across model scales, and compare

Table 1: **Long-video MLLMs with unified token-cost analysis.** *Cost*, *T/F*, and *#F* are measured on VideoMME and serve as the unified LLM-facing efficiency reference; other benchmark results retain their originally reported protocols. Rows below "Unified re-evaluation" use 1 fps with at most 768 frames. †: LLaMA-VID keeps only its per-frame visual *content token*, removing the text-conditioned *context token*. §: TimeChat-Online uses the original feature-level filtering rule with threshold 0.1, where smaller values remove more tokens. ‡: TimeChat-Online uses its original redundancy criterion but retains only the two least-redundant tokens per frame. ¶: Tempo uses a ≈2 token/frame visual-only setting: when generating 128 visual memory tokens, its compressor masks all non-visual conditioning, including system-prompt and user question. **Abbr.** LV = LVBench, LongVB = LongVideoBench, ES = EgoSchema, NQ = NExT-QA, and MLVU = MLVU-Dev.

Model	Size	Cost	T/F	#F	VideoMME				LV	LongVB	ES	NQ	MLVU
					S	M	L	T					
ShareGPT4Video (Chen et al., 2024b)	8B	–	–	16	48.3	36.3	35.0	39.9	–	–	–	–	46.4
SliME (Zhang et al., 2024b)	8B	–	–	8	53.3	42.7	39.8	45.3	–	–	–	–	–
Kangaroo (Liu et al., 2024a)	8B	16384	256.0	64	66.1	55.3	46.7	56.0	39.4	54.8	62.7	–	61.0
Video-CCAM (Fei et al., 2024)	14B	1024	10.7	96	62.1	52.8	47.0	53.9	–	–	–	–	63.1
InternVL-V1.5 (Chen et al., 2024c)	20B	2560	256.0	10	60.2	46.4	45.6	50.7	–	–	–	–	50.4
VideoChat2-HD (Li et al., 2024a)	7B	–	72.0	–	–	–	39.8	45.3	–	–	65.6	79.50	44.5
LongVA (Zhang et al., 2024a)	7B	18432	144.0	128	61.1	50.4	46.2	52.6	–	–	–	68.27	56.3
LongLLaVA (Wang et al., 2024a)	13B	18432	144.0	128	62.9	52.2	46.4	53.8	–	53.5	–	–	–
LLaMA-VID (Li et al., 2024c)	7B	–	2.0	–	–	–	–	25.9	23.9	–	59.5	–	33.2
Unified re-evaluation: 1 fps, max. 768 frames													
LLaMA-VID (Li et al., 2024c)	7B	$904.17_{2.00\times}$	2.0	452.08	31.8	27.0	26.2	28.3	23.3	26.0	23.7	42.70	23.7
LLaMA-VID† (Li et al., 2024c)	7B	$452.08_{1.00\times}$	1.0	452.08	31.4	26.4	25.0	27.6	20.8	25.1	22.5	40.07	22.3
TimeChat-Online§ (Yao et al., 2025a)	7B	**1501.85**$_{3.32\times}$	**6.7**	225.89	62.2	**59.3**	**52.9**	58.1	**42.4**	53.5	39.2	72.59	61.9
TimeChat-Online‡ (Yao et al., 2025a)	7B	**451.78**$_{1.00\times}$	**2.0**	**225.89**	57.4	56.2	51.0	54.9	39.6	51.5	39.7	69.27	62.4
Tempo¶ (Fei et al., 2026)	6B	$940.66_{2.08\times}$	2.1	451.79	40.8	41.3	38.7	40.3	25.6	41.9	33.6	54.06	47.2
Ours	4B	**451.78**$_{1.00\times}$	**2.0**	**225.89**	60.6$_{+3.2}$	54.9$_{-1.3}$	48.3$_{-2.7}$	54.6$_{-0.3}$	38.5$_{-1.1}$	50.9$_{-0.6}$	48.5$_{+8.8}$	74.12$_{+4.85}$	63.0$_{+0.6}$
	8B	**451.78**$_{1.00\times}$	**2.0**	**225.89**	**64.4**$_{+7.0}$	58.7$_{+2.5}$	51.6$_{+0.6}$	**58.2**$_{+3.3}$	37.7$_{-1.9}$	**53.9**$_{+2.4}$	**52.9**$_{+13.2}$	**75.36**$_{+6.09}$	**63.6**$_{+1.2}$

against strong video MLLMs for long-context modeling and visual compression. We further select three counterparts for controlled re-evaluation: 1) LLaMA-VID compresses each frame to two tokens, 2) TimeChat-Online dynamically removes visual redundancy without language guidance, and 3) Tempo performs query-aware allocation under a fixed token budget, but in our re-evaluation, we disable all textual conditioning during visual-token computation to match our query-independent setting. As close counterparts to ours, all three are re-evaluated in Tables 1& 2 under the same protocol and comparable token costs, enabling a fair comparison under similarly constrained budgets.

4.2 GLANCEVU: TOKEN-CONSTRAINED VIDEO UNDERSTANDING

Benchmark motivation and composition. Existing benchmarks offer broad but heterogeneous evaluations and often couple visual understanding with auxiliary requirements that do not cleanly isolate the capability studied here: **recovering scene context, salient events, and coarse temporal evolution from an extremely sparse, purely visual representation**. We therefore construct **GlanceVU**, auditing 41,208 multiple-choice questions from 13 open-source benchmarks and filtering out samples requiring non-visual cues, precise temporal localization, unsupported inference, or textual shortcuts. Constraint-based selection yields *2,400 questions from 13 sources* (Fig. 4): a 1,800-question *Core* set balanced across six scene domains, four temporal horizons, three capability levels, and 12 subtasks, plus three 200-question tracks targeting *Holistic Understanding*, *Online Memory*, and *Cross-View Understanding*. Spanning egocentric, embodied, urban, sports, narrative, and documentary videos (Fig. 6), GlanceVU provides a controlled, reusable testbed for sparse-frame and token-constrained video understanding. Appendix A.1 details its construction and composition.

Complementary evaluation suites. Beyond GlanceVU, we evaluate three complementary regimes: **(i) general video QA** on NExT-QA and EgoSchema; **(ii) long-video understanding** on VideoMME, MLVU, LVBench, and LongVideoBench; and **(iii) timestamp-bounded video understanding** on StreamingBench-Real and the Real-Time/Backward subsets of OVO-Bench, where each query is evaluated offline using only frames sampled from the video prefix up to its timestamp. Together, these suites test semantic preservation, long-horizon evidence retention, and query-time reasoning without future-frame access. See Appendix A.2 for more details.

4.3 MAIN RESULTS

Competitive long-video accuracy with only two visual tokens per frame. Table 1 reports prior long-video results under their original protocols and a controlled re-evaluation of three representative

Table 2: **Accuracy (%) on GlanceVU across four duration ranges and 12 fine-grained understanding categories.** Best results within each group are **bolded**. † LLaMA-VID keeps only the visual *content token*, excluding the text-conditioned *context token*. ‡ TimeChat-Online retains the two least-redundant tokens per frame under its original redundancy criterion. § TimeChat-Online uses the original filtering rule with threshold 0.1; smaller values drop more tokens. Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), and LLaVA-OneVision-2 (An et al., 2026) denote their Instruct checkpoints. Task abbreviations are detailed in Appendix A.1. All use 1fps ($\leq$ 768 frames).

Method	LLM Size	Video Duration				Fine-Grained Task Taxonomy												Overall
		Short $\leq$ 2 min	Medium (2, 10] min	Long (10, 25] min	Extra-long $>$ 25 min	AS	CEC	CFD	CSA	EA	GTS	IGM	KVD	Ord.	PTP	SL	SCP	
Detailed-understanding baselines																		
Qwen2.5-VL	3B	45.94	53.21	54.64	60.21	49.25	56.67	38.67	53.85	43.33	73.14	52.53	70.00	51.35	44.00	**48.67**	40.47	52.92
	7B	43.28	50.51	56.54	60.38	49.25	68.00	30.67	48.56	48.00	74.57	48.64	60.67	59.46	40.67	36.67	40.47	51.83
Qwen3-VL	4B	48.74	54.05	**66.67**	67.07	**60.20**	68.67	44.67	55.77	58.00	**78.00**	**54.47**	76.00	62.70	42.00	46.67	41.14	58.04
	8B	**49.93**	**58.11**	66.24	**69.98**	59.20	**72.00**	**53.33**	**60.58**	**60.67**	77.71	**54.47**	**76.67**	**63.78**	**45.33**	47.33	44.48	**60.04**
VideoChat3	4B	49.40	55.41	64.35	67.75	59.20	**72.00**	49.33	54.33	58.67	**78.00**	53.31	72.00	54.59	44.67	46.67	**47.16**	58.29
LLaVA-OneVision-2	8B	49.67	56.93	62.66	66.72	59.70	68.00	51.33	54.81	59.33	76.86	52.53	71.33	62.70	43.33	48.00	43.48	58.17
Glance-understanding methods																		
LLaMA-VID	7B	25.57	26.86	26.16	26.93	26.87	16.67	31.33	25.48	26.00	29.43	25.68	22.67	27.03	29.33	23.33	27.42	26.33
LLaMA-VID†	7B	24.77	23.65	23.00	26.07	25.37	21.33	26.00	24.04	25.33	26.00	23.74	21.33	21.08	25.33	28.67	24.41	24.46
TimeChat-Online‡	7B	41.15	40.20	47.68	56.43	39.30	54.00	34.67	44.71	36.67	66.29	35.02	51.33	48.65	42.67	42.00	**42.14**	45.92
TimeChat-Online§	7B	41.41	39.70	50.00	**59.18**	40.80	58.00	37.33	42.31	42.00	67.43	34.63	**56.67**	**49.73**	**44.67**	**44.67**	38.80	47.00
Ours	4B	41.94	44.93	53.16	56.26	45.77	**63.33**	36.67	**51.44**	**45.33**	68.86	**45.14**	47.33	46.49	40.67	44.00	34.45	48.38
	8B	**43.54**	**48.31**	**53.38**	57.29	**50.25**	**63.33**	**42.00**	47.60	**45.33**	**71.71**	43.19	56.00	48.65	42.67	44.00	36.12	**50.00**

compression methods under a shared 1 fps protocol with at most 768 frames; VideoMME provides the common token-cost reference. GlanceVLM-4B/8B each use an average of only 451.8 visual tokens per video, achieving 54.6%/58.2% on VideoMME and 63.0%/63.6% on MLVU. Compared with Kangaroo-8B, a dedicated long-video model using 256 visual tokens per frame, GlanceVLM-8B **gains 2.2/2.6 points** on VideoMME/MLVU while reducing the VideoMME reference cost from 16,384 to 451.8 tokens (**36.3**×). On MLVU, GlanceVLM-8B retains 82.7% of the accuracy of the dense Qwen3-VL-8B backbone (63.6 vs. 76.9), while using only 1.05% as many visual tokens (494 vs. 47,090 in Appendix Table 6).

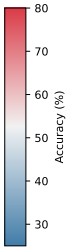
Controlled comparison with compression baselines. At essentially matched visual-token cost, GlanceVLM-8B reaches 58.2%/52.9% on VideoMME/EgoSchema, compared with 54.9%/39.7% for TimeChat-Online (↓3.3/↓13.2) and 27.6%/22.5% for the LLaMA-VID variant (↓30.6/↓30.4) (Table 1). Under our re-evaluation, which masks all non-visual conditioning during memory-token generation, Tempo reaches only 40.3%/33.6% despite using 940.7 tokens, approximately 2.1× our cost. To approach our VideoMME performance 58.2%, TimeChat-Online must **increase its budget** to 1,501.9 tokens, reaching 58.1% **at 3.3× our cost**. Overall, token count alone does not determine performance: under an extreme fixed budget, performance depends critically on what visual evidence each token preserves and how effectively it organizes that evidence.

Dense and compressed performance on GlanceVU. The strongest detailed-understanding baselines reach only 58.0–60.0% on GlanceVU, with Qwen3-VL-8B-Instruct attaining 60.0% (Table 2). GlanceVLM-4B/8B obtains 48.4/50.0%; the 8B model leads all compressed baselines and exceeds the matched two-token TimeChat-Online variant across every duration range, by 4.1 points overall. It retains 83.3% of its dense backbone's accuracy while using only 1.09% as many visual tokens (412.5 vs. 37,947 in Appendix Table 6). Together, *the best tested dense score of only* 60.0% *shows that GlanceVU remains challenging under both regimes*.

Cross-backbone robustness and controlled variants. On VideoMME, GlanceVLM consistently outperforms direct two-token pooling across all tested backbones (Table 6). Total accuracy rises by 5.5/8.1/12.3 points to 51.2%/54.6%/58.2% on Qwen2.5-VL-7B, Qwen3-VL-4B, and Qwen3-VL-8B, respectively, with gains in every duration range. On Qwen3-VL-8B, question text-only input, same-budget adaptive pooling, and a spatially dense eight-frame control trail GlanceVLM by 14.6/4.8/1.9 points, respectively; the eight-frame control still uses roughly 3× more visual tokens. These results show that the gains persist across backbone families and scales and are not explained by token reduction alone, textual priors, or concentrating more visual tokens in a few frames.

Table 3: **Performance comparison of 4B, 7B and 8B models across sub-tasks and different experimental settings on VideoMME.** 1) We substitute our proposed module with a naive frame-level adaptive pooling operation (2 tokens/frame) across Qwen2.5-VL-7B, Qwen3-VL-4&8B (1 *fps*, $\leq$ 768 *frames*). 2) For the 8B model, we further investigate a text-only input (omitting all visual tokens in GlanceVLM) and evaluate a spatial-temporal token reallocation strategy on Qwen3-VL-8B (trading temporal length for spatial resolution, *i.e.*, 8 frames, using 1438.91 tokens). Cell color encodes accuracy (25–80%), from blue (low) to red (high). TP: Temporal Perception, SP: Spatial Perception, AP: Attribute Perception, AR: Action Recognition, OR: Object Recognition, OCR: OCR Problems, CNT: Counting, TR: Temporal Reasoning, SR: Spatial Reasoning, AcR: Action Reasoning, ObR: Object Reasoning, IS: Information Synopsis.

Method	TP	SP	AP	AR	OR	OCR	CNT	TR	SR	AcR	ObR	IS
Qwen3-8B+2T	43.6	50.0	45.5	44.4	42.1	38.8	29.5	45.2	51.8	44.9	51.3	61.0
Ours-8B	65.5	53.7	65.3	59.4	60.5	54.0	40.7	51.4	66.1	56.5	57.0	76.2
Ours-8B-Only Text	40.0	37.0	46.4	40.9	35.6	33.1	36.9	45.8	35.7	43.2	47.6	59.4
Qwen3-8B-Time2Space	56.4	61.1	67.1	50.5	60.7	54.0	37.7	44.6	75.0	53.7	55.9	71.5
Ours-8B-Pool	52.7	51.9	61.3	56.2	54.0	46.8	38.4	27.7	67.9	48.1	53.3	76.8
Qwen2.5-7B+2T	47.3	55.6	46.8	49.8	43.8	36.0	32.8	35.6	57.1	46.0	44.3	61.3
Ours-7B	56.4	55.6	54.1	53.0	50.8	38.8	39.9	34.5	58.9	52.6	52.4	65.6
Qwen3-4B+2T	49.1	57.4	43.2	48.2	48.0	43.9	28.7	35.6	60.7	47.4	47.8	59.8
Ours-4B	54.5	59.3	65.3	54.6	56.2	51.8	39.2	37.3	71.4	52.3	51.3	71.8



Capability profile under the two-token bottleneck. Relative to direct two-token pooling, GlanceVLM improves 34 of 36 VideoMME subtask–backbone pairs; both Qwen3-VL scales improve across all 12 categories in Table 3. Across all three backbones, Information Synopsis and Spatial Reasoning are consistently the two strongest categories, whereas Counting and Temporal Reasoning remain among the weakest. GlanceVU reveals a complementary profile in Table 2: Global Theme Summary and Cause–Effect Condition score highest (71.7%/63.3%), whereas Key Visual Detail, Entity Attribute, Ordering, and Cross-Segment Association exhibit the largest gaps to dense Qwen3-VL-8B (20.7/15.3/15.1/13.0 points). Together, these results reveal the capability boundary of GlanceVLM: its dual-token encoding preserves global semantics and coarse spatial–temporal structure more robustly than fine-grained visual detail, counting, or exact event order.

4.4 Ablation Studies

Table 4: **The roles of the 2 tokens on the GlanceVLM-8B & 4B models.** Res. = Residual region, EgoSc. = EgoSchema, Real. = Realtime, and Back. = Backward.

$\mathcal{A}_t$	$\mathcal{I}_t$	Res.	VideoMME				MLVU	EgoSc.	OVOBench		Avg.
			S	M	L	T			Real.	Back.	
GlanceVLM-8B											
✓	✓	✓	64.4	58.7	51.6	58.2	63.58	52.87	52.44	37.91	54.96
✓	✗	✓	59.0	54.6	50.6	54.7	59.00	49.42	46.28	38.80	51.55 ↓**3.41**
✗	✓	✓	64.4	58.0	50.1	57.5	61.73	52.95	52.64	41.03	54.79 ↓**0.17**
✓	✓	✗	59.0	54.9	50.1	54.7	58.83	49.49	46.74	39.11	51.61 ↓**3.35**
GlanceVLM-4B											
✓	✓	✓	60.6	54.9	48.3	54.6	63.04	48.52	51.87	41.29	52.89
✓	✗	✓	52.2	51.3	46.1	49.9	55.92	43.99	45.50	35.14	47.51 ↓**5.38**
✗	✓	✓	59.2	54.2	47.6	53.7	59.75	47.51	51.18	38.44	51.45 ↓**1.44**
✓	✓	✗	55.0	50.6	45.7	50.4	57.55	43.67	45.72	34.44	47.88 ↓**5.01**

Disentangling token roles. We conduct token ablations without retraining on GlanceVLM-4B/8B across four benchmarks in Table 4: VideoMME, MLVU, EgoSchema, and OVO-Bench. Both Anchor and Innovation tokens *undergo the complete refinement process*, only the final output token is removed immediately before LLM input. Removing Anchor token $\mathcal{A}_t$ reduces the mean of the eight reported scores by 1.44/0.17 points on the 4B/8B models, whereas removing the Innovation token $\mathcal{I}_t$ incurs substantially larger losses of 5.38/3.41 points. Whether for long-term tasks or instantaneous tasks that focus more on momentary behavior, the dynamic changes in the video contained within the $\mathcal{I}_t$ proved more helpful for the response task on both models. The marginal EgoSchema gain with $\mathcal{I}_t$ alone (+0.08 on 8B) may reflect more focused use of local interaction cues amid repeated scene content. The full pair's higher mean accuracy supports complementary token roles, while the asymmetric removal losses identify $\mathcal{I}_t$ *as the stronger contributor to prediction under the two-token bottleneck.*

An extra token or residual-guided encoding? To further examine whether the above advantage comes merely from allocating an additional token or from *how the Innovation token* $\mathcal{I}_t$ *is computed*, we retain both tokens but *replace* $\mathcal{I}_t$*'s residual-routed regional evidence with the original*

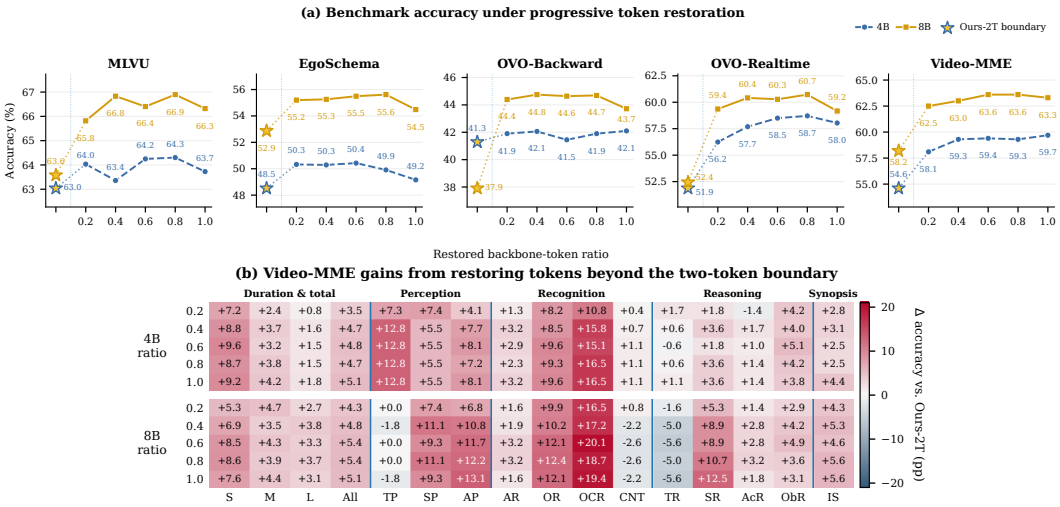


Figure 3: **Token restoration beyond the two-token boundary. (a) Cross-benchmark saturation.** The star marks our *Ours-2T* boundary, while the blue and gold curves show the 4B and 8B models after restoring 20%–100% of the backbone's native visual tokens. Most gains emerge at low restore ratios and then plateau or fluctuate. **(b) Task-selective recovery.** Each VideoMME cell reports $\mathrm{Acc}_{\mathrm{restore}} - \mathrm{Acc}_{\text{Ours-2T}}$ in percentage points, with red/blue denoting gains/losses. Additional tokens mainly improve spatial and attribute perception, object recognition, and OCR, while several temporal and counting metrics show limited or negative gains. GlanceVU is shown in Fig. 7.

frame at inference. Mean accuracy drops by $5.0/3.4$ points on the 4B/8B models in Table 4, nearly matching $\mathcal{I}_t$ removal ($5.4/3.4$). Thus, **the benefit of $\mathcal{I}_t$ lies not in an additional token slot but in its residual-guided construction**: once $\mathcal{A}_t$ establishes holistic frame context, $\mathcal{I}_t$ reads current-frame semantics from regions of pronounced inter-frame change, where answer-relevant actions and object-state transitions often concentrate. This is also shown in Fig. 8 & 9 & 10 & 11 & 12.

What do restored backbone tokens recover? At inference, we augment GlanceVLM's learned two-token representation with a fraction $r \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ of the native visual tokens from its corresponding uncompressed backbone, where r is the ratio of restored tokens to the backbone's full native visual-token count; no parameters are retrained and all other settings remain fixed (Fig. 3). For the 8B model, accuracy generally improves up to $r = 0.8$: MLVU/EgoSchema rise from $63.6/52.9\%$ to $66.9/55.6\%$, OVO-Backward/Realtime from $37.9/52.4\%$ to $44.7/60.7\%$, and VideoMME from 58.2% to 63.6%. Full restoration ($r = 1.0$) then reduces all five scores. The 4B model exhibits the same overall trend of early gains followed by saturation or fluctuation, although its optimal r varies across benchmarks. A separate analysis of GlanceVU is shown in Fig. 7.

What, then, do the restored tokens actually recover? Figure 3(b) decomposes this effect by VideoMME subtasks. At $r = 0.8$, the 8B model gains $18.7/12.4/12.2$ points on OCR, object recognition, and attribute perception, but $0.0/-5.0/-2.6$ on temporal perception, temporal reasoning, and counting. The 4B model improves all 12 categories, led by OCR ($+16.5$) and temporal perception ($+12.8$), yet gains only 0.6 on temporal reasoning. Thus, **restored backbone tokens most consistently recover perceptual and recognition detail**, whereas simply increasing token count does not reliably translate into gains on higher-level reasoning tasks such as temporal reasoning.

5 Conclusion

We formalize **glance-style video understanding** and probe its extremely question-independent boundary with GlanceVLM, which represents each frame using only complementary **Anchor** $\mathcal{A}_t$ and **Innovation** $\mathcal{I}_t$ tokens. Together with GlanceVU, our results show that useful long-horizon understanding can persist under an ultra-low, query-independent visual budget, reframing extreme efficiency as a representation problem rather than merely token removal. We view this work as a starting point for further research on glance video understanding, aiming to preserve and ultimately improve this capability without incurring high visual-token costs.

REFERENCES

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18407–18418, 2024a.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024b.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations*, volume 2025, pp. 18227–18246, 2025.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67 (12):220101, 2024c.

Yifan Du, Kun Zhou, Yuqi Huo, Yifan Li, Wayne Xin Zhao, Haoyu Lu, Zijia Zhao, Bingning Wang, Weipeng Chen, and Ji-Rong Wen. Towards event-oriented long video understanding. *arXiv preprint arXiv:2406.14129*, 2024.

Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*, 2024.

Junjie Fei, Jun Chen, Zechun Liu, Yunyang Xiong, Chong Zhou, Wei Wen, Junlin Han, Mingchen Zhuge, Saksham Suri, Qi Qian, et al. Small vision-language models are smart compressors for long video understanding. *arXiv preprint arXiv:2604.08120*, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

Yuping He, Yifei Huang, Guo Chen, Baoqi Pei, Jilan Xu, Tong Lu, and Jiangmiao Pang. Egoexobench: A benchmark for first- and third-person view video understanding in MLLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=Tpb7uKAN3H.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Xiang Yue, Bo Li, Yuanhan Zhang, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multidisciplinary professional videos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27798–27828, 2026.

Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19959–19973, 2025.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024a.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024b.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, et al. Videochat3: Fully open video mllm for efficient and generalist video understanding. *arXiv preprint arXiv:2607.14935*, 2026.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2024c.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12147–12151. IEEE, 2026.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024a.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Chang Wen Chen, and Ying Shan. E.t. bench: Towards open-ended event-level video-language understanding. In *Neural Information Processing Systems (NeurIPS)*, 2024b.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18902–18913, 2025.

Mingjie Pan, Jiyao Zhang, Tianshu Wu, Yinghao Zhao, Wenlong Gao, and Hao Dong. Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17359–17369. IEEE, 2025.

Ahmed Sharshar, Latif U Khan, Waseem Ullah, and Mohsen Guizani. Vision-language models for edge networks: A comprehensive survey. *IEEE Internet of Things Journal*, 2025.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Gaurav Shinde, Anuradha Ravi, Emon Dey, Shadman Sakib, Milind Rampure, and Nirmalya Roy. A survey on efficient vision-language models. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(3):e70036, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18992–19001. IEEE, 2025.

Kwai Keye Team, Bin Wen, Changyi Liu, Chengru Song, Chongling Rao, Guowang Zhang, Han Li, Haonan Fan, Hengrui Ju, Jiankang Chen, et al. Kwai keye-vl-2.0 technical report. *arXiv preprint arXiv:2606.10651*, 2026.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture. *arXiv preprint arXiv:2409.02889*, 2024a.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. *arXiv preprint arXiv:2405.19209*, 2024b.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024.

Zhirong Wu, Xiaodong Wang, Langling Huang, Teng Xu, and Peixi Peng. A training-free framework for long video understanding via video-query-options similarity. In *The Fourteenth International Conference on Learning Representations*, 2026.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Renjie Xu, Saiedeh Razavi, and Rong Zheng. Edge video analytics: A survey on applications, systems and enabling techniques. *IEEE Communications Surveys & Tutorials*, 25(4):2951–2982, 2023.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28885–28900. IEEE, 2025.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, et al. Timechat-online: 80% visual tokens are naturally redundant in streaming videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10807–10816, 2025a.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Chi Chen, Haoyu Li, Weilin Zhao, et al. Efficient gpt-4v level multimodal large language model for deployment on edge devices. *Nature Communications*, 16(1):5509, 2025b.

Jiashuo Yu, Yue Wu, Meng Chu, Zhifei Ren, Zizheng Huang, Pei Chu, Ruijie Zhang, Yinan He, Qirui Li, Songze Li, et al. Vrbench: A benchmark for multi-step reasoning in long narrative videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21655–21666. IEEE, 2025.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9127–9134, 2019.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024a.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22056–22065, 2025.

Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024b.

Zheyu Zhang, Ziqi Pang, Shixing Chen, Xiang Hao, Vimal Bhat, and Yu-Xiong Wang. One token per highly selective frame: Towards extreme compression for long video understanding. *Advances in Neural Information Processing Systems*, 38:57332–57358, 2026.

Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, et al. Urbanvideo-bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32400–32423, 2025.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025.

Kejian Zhu, Zhuoran Jin, Hongbang Yuan, Jiachun Li, Shangqing Tu, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Mmr-v: What's left unsaid? a benchmark for multimodal deep reasoning in videos. In *International Conference on Learning Representations*, volume 2026, pp. 67949–67991, 2026.

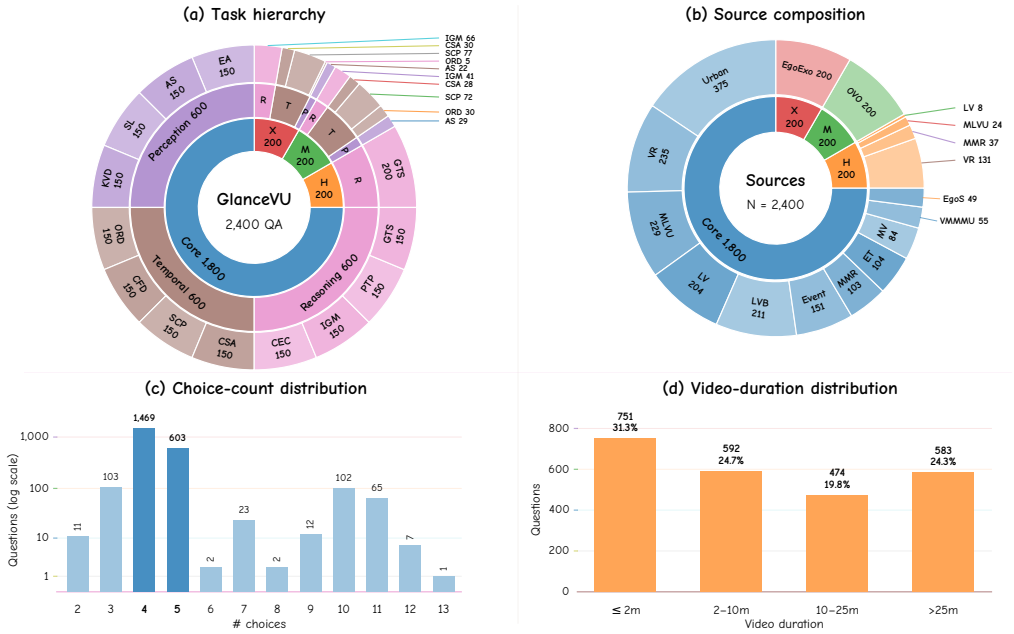


Figure 4: **Composition and statistics of GlanceVU. (a) Hierarchical task taxonomy of the 2,400-question benchmark.** The 1,800-question Core track is evenly divided into Perception, Temporal, and Reasoning, each comprising four 150-question subtasks, while the three 200-question specialized tracks target Holistic understanding (H), online Memory (M), and Cross-View reasoning (X). **(b) Upstream benchmark composition**, with the inner ring denoting the four tracks and the outer ring showing their source contributions. **(c) Distribution of the native number of answer choices** on a logarithmic count scale; 4- and 5-choice questions jointly account for 86.3% of GlanceVU. **(d) Video-duration distribution** across the predefined $\leq$2, 2–10, 10–25, and $>$25 min ranges, containing 751, 592, 474, and 583 questions, respectively; the maximum durations are 251 minutes. ***Task abbreviations:*** AS = Action State; CEC = Cause Effect Condition; CFD = Count Frequency Duration; CSA = Cross-Segment Association; EA = Entity Attribute; GTS = Global Theme Summary; IGM = Intent Goal Motivation; KVD = Key Visual Detail; Ord. = Ordering; PTP = Procedure Trajectory Planning; SL = Spatial Location; SCP = State Change Progress.

A More Details on the Experiment

A.1 GlanceVU

Overview. Existing video benchmarks are intentionally designed for broad capability assessment, combining video-level reasoning and summarization with tasks whose answers depend on recovering narrowly localized or fine-grained visual evidence, such as OCR, attribute recognition, counting, temporal grounding, and retrieval of specific details. While this breadth is desirable for general-purpose evaluation, it makes aggregate performance difficult to interpret under the extreme visual constraints studied here: *a model may preserve the scene context, salient events, and their temporal evolution yet fail because a transient frame-level detail is no longer recoverable*. Conversely, recovering an isolated local detail does not by itself imply that the representation preserves a coherent understanding of the video as a whole. Existing benchmarks therefore measure many capabilities relevant to video understanding, ***but do not cleanly isolate the capability central to our study***: how much video-level understanding remains accessible when visual cost is constrained.

We therefore introduce **GlanceVU**, a visual-only benchmark for *glance-style video understanding*, operationalized as recovering scene context, salient events, and their global temporal evolution without requiring exhaustive detail recovery. Rather than eliminating local evidence altogether, GlanceVU excludes instances whose solution is dominated by audio, subtitles, frame-accurate localization, or externally supplied evidence, while retaining diverse perceptual, temporal, and reasoning

demands that contribute to an overall understanding of the video. Starting from 41,208 questions across 16 public video benchmarks, we construct a human-validated set of 2,400 questions from 13 sources, with explicit controls over capability coverage, source concentration, and repeated visual content. GlanceVU thus complements the conventional evaluation suite by providing a controlled diagnostic of what visual understanding remains accessible under severely constrained visual tokens.

Tracks and Taxonomy. GlanceVU organizes its 2,400 questions into a 1,800-question *Core* track and three complementary 200-question tracks. Core evaluates general glance-style understanding, while *Holistic* targets global themes and full-video summaries, *Online Memory* requires retrieving evidence available before a specified query time, and *Cross View* examines associations across two ordered visual streams. Within Core, we independently balance six scene domains, a 4×3 grid of temporal horizons and capability levels, and 12 fine-grained subtasks. The temporal axis ranges from short (≤ 2 min) to extra-long (> 25 min) contexts, while the capability axis covers *Perception*, *Temporal*, and *Reasoning*, each decomposed into four subtasks; every scene, horizon–capability cell, and subtask receives 300, 150, and 150 questions, respectively. The 12 subtasks (Fig. 4 (a)) are *Action State* (AS), *Cause Effect Condition* (CEC), *Count Frequency Duration* (CFD), *Cross-Segment Association* (CSA), *Entity Attribute* (EA), *Global Theme Summary* (GTS), *Intent Goal Motivation* (IGM), *Key Visual Detail* (KVD), *Ordering* (Ord.), *Procedure Trajectory Planning* (PTP), *Spatial Location* (SL), and *State Change Progress* (SCP). Some examples and the responses of 5 models are shown in Fig. 6.

Source Coverage. To consolidate heterogeneous upstream annotations without discarding their original semantics, we standardize each item into a unified representation containing its question–answer content, duration metadata, taxonomy labels, and canonical provenance. As summarized in Fig. 4, the Core track draws from 11 public benchmarks: UrbanVideo-Bench (Urban) (Zhao et al., 2025), VRBench (VR) (Yu et al., 2025), MLVU (Zhou et al., 2025), LVBench (LV) (Wang et al., 2025a), LongVideoBench (LVB) (Wu et al., 2024), EventBench (Event) (Du et al., 2024), MMR-VBench (MMR) (Zhu et al., 2026), ETBench (ET) (Liu et al., 2024b), MVBench (MV) (Li et al., 2024a), VideoMMMU (VMMMU) (Hu et al., 2026), and EgoSchema (EgoS) (Mangalam et al., 2023). Four of these sources—VRBench, MLVU, LVBench, and MMR-VBench— also contribute to the Holistic track, while Online Memory and Cross View are sourced from OVO-Bench (OVO) (Niu et al., 2025) and EgoExoBench (EgoExo) (He et al., 2026), respectively. GlanceVU therefore draws from 13 public benchmark sources in total, with their sample proportions visualized in Fig. 4 (b).

What GlanceVU Reveals. GlanceVU remains unsaturated even without the two-token-per-frame constraint. Under the same 1-FPS setting, recent detailed-understanding baselines peak at only 60.04% overall, with Qwen3-VL-8B-Instruct achieving 60.04% and VideoChat3 and LLaVA-OneVision-2 remaining below 60% (Table 2). This gap is notable because GlanceVU deliberately excludes questions dominated by audio, subtitles, frame-accurate localization, or external evidence, while Core assigns equal weight to its 12 subtasks. ***The remaining errors therefore cannot be attributed to our extreme visual-token bottleneck or to a single overrepresented capability***; instead, they expose a broader difficulty in consistently integrating scene context, salient events, and temporal evolution into a coherent video-level representation. Our 8B model reaches 50.0% under only 2 visual tokens per frame, separating two open problems: improving the underlying capability for glance-style video understanding and preserving that capability under constrained visual budgets.

Data Screening. We first normalize annotations from 13 upstream benchmarks into a shared schema while preserving their native questions, answer choices and order, temporal metadata, and source provenance. Deterministic structural and visual-only gates then remove malformed annotations, missing media, explicit dependence on speech, audio, subtitles, transcripts, or ASR, tasks requiring unobserved or counterfactual events, and high-confidence textual or answer shortcuts. To capture semantic failure modes not readily expressed by hand-written rules, we additionally use `gpt-5.6` for text-only semantic auditing of candidate records. Its findings serve only as screening proposals: verified failure modes are converted into explicit deterministic rules and the entire pool is rebuilt, rather than using model judgments as final eligibility labels. This process retains 31,275 of the 41,208 audited questions; after explicit deterministic rules, 27,808 remain in the selection pool.

Rather than randomly subsampling the resulting pool, we *formulate benchmark assembly as a binary constrained-optimization problem* that jointly enforces the four Track quotas, the Core scene, temporal-horizon–capability, and subtask marginals, source diversity, and canonical-provenance

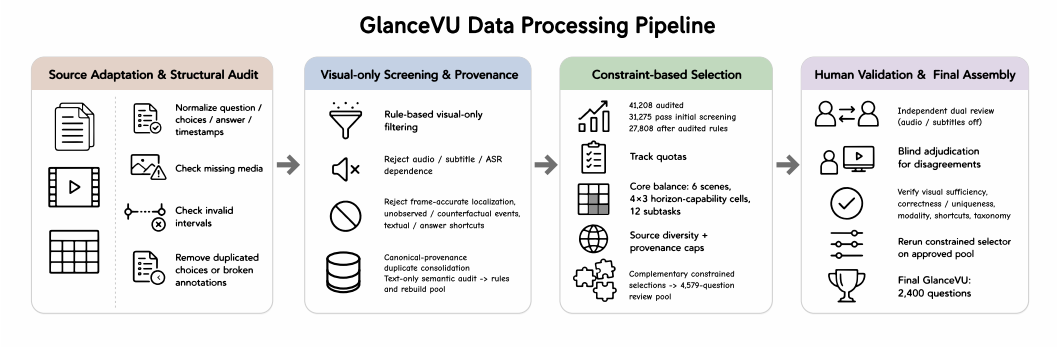


Figure 5: **GlanceVU data processing pipeline.** Annotations from 13 public benchmarks undergo schema normalization, structural validation, and deterministic filtering for visual-only eligibility. Text-only semantic auditing only proposes additional failure modes, which are verified and codified as deterministic rules before rebuilding the 27,808-question candidate pool. Two complementary quota-constrained selections form a 4,579-question review pool, followed by dual human verification, blind adjudication, and final constrained selection of 2,400 questions.

Table 5: **Comparison of GlanceVU with representative video-understanding benchmarks.** *Scale* denotes #Video/#QA. V, A, and S denote visual, audio, and subtitle evidence, with parentheses indicating auxiliary modalities. $\mathrm{V}_{\leq t}$ and $\mathrm{V}_{\mathrm{Ego+Exo}}$ denote causal visual evidence up to the query time and paired egocentric–exocentric visual streams, respectively. *Multi-Level* indicates explicit evaluation across multiple video-duration ranges.

Benchmark	Scale (#Vid./#QA)	Duration	Answer Evidence	Multi-Level	#Choices	Capability Focus
MVBench (Li et al., 2024a)	3,641 / 4,000	– (avg. 16.0 s)	V	–	3	Multi-task temporal understanding on short clips
EgoSchema (Mangalam et al., 2023)	5,063 / 5,063	3 min	V	–	5	Egocentric activity understanding
Video-MME (Fu et al., 2024)	900 / 2,700	11 s–60 min	V(+A/S)	✓	4	Multi-spectrum multimodal video analysis
MLVU$_{Dev}$ (Zhou et al., 2025)	1,122 / 2,174	3–120 min	V	✓	4	Multi-task long-video understanding
LongVideoBench (Wu et al., 2024)	3,763 / 6,678	8 s–60 min	V + S	✓	4–5	Long-context referred-detail retrieval & reasoning
OVO-Bench (Niu et al., 2025)	644 / 2,814	– (few min–30 min)	$\mathrm{V}_{\leq t}$	–	2–5	Online video understanding & temporal awareness
StreamingBench (Lin et al., 2026)	900 / 4,500	3 s–24 min	$\mathrm{V}_{\leq t}$(+A)	–	4	Comprehensive streaming video understanding
EgoExoBench (He et al., 2026)	– / 7,330	– (paired clips)	$\mathrm{V}_{\mathrm{Ego+Exo}}$	–	4	Cross-view video understanding & reasoning
GlanceVU	**2,036 / 2,400**	**1 s–251 min**	V	✓	**2–13**	**Glance video understanding under low cost**

limits while favoring higher-quality items. To provide sufficient redundancy for human validation, we construct two complementary constrained selections, discouraging reuse of items from the first selection, and merge them into a diversified 4,579-question review pool. This enlarged pool provides replacement candidates when individual items fail subsequent validation. Finally, every item in the review pool is independently inspected by two reviewers using the actual video with audio and subtitles disabled. Reviewers verify visual sufficiency, answer correctness and uniqueness, modality dependence, textual shortcuts, and the proposed taxonomy labels, while disagreements are resolved by an independent blind adjudicator. After adjudication, we rerun the constrained selector over the human-approved pool to obtain the final 2,400-question GlanceVU benchmark, ensuring that the prescribed capability coverage, source diversity, and provenance constraints remain satisfied after manual quality control.

(a) Semantic Audit

Input. Versioned canonical records containing {`source/task, question, native choices, gold answer, temporal metadata, issue events, proposed taxonomy`}. The model receives *no video frames, audio, subtitles, transcripts, or ASR output*.
Audit. Using only these textual and metadata fields, `gpt-5.6` searches for semantic failure modes not fully captured by the current deterministic gates, including implicit audio/subtitle dependence, unobserved or counterfactual events, unsupported temporal localization, and textual or media-locator shortcuts. Each finding must identify its affected scope and supporting evidence.
Proposal and acceptance. The model proposes {`scope`, `failure mode`, `evidence`, `affected items`, `rationale`, `confidence`}, but its outputs never directly determine item eligibility. A proposal changes the pool only after it is reproducible from annotations, metadata, or file-level evidence and encoded as a deterministic gate.
Iterative refinement. Let $\mathcal{R}^{(v)}$, $\mathcal{P}^{(v)}$, and $\mathcal{G}^{(v)}$ denote the records, model proposals, and accumulated deterministic gates at version v, respectively:

$$\mathcal{P}^{(v)} = \mathrm{Audit}_{\mathrm{text}}(\mathcal{R}^{(v)}) \;\rightarrow\; \Delta\mathcal{G}^{(v)} = \mathrm{Verify}(\mathcal{P}^{(v)}) \;\rightarrow\; \mathcal{R}^{(v+1)} = \mathrm{Rerun}\Big(\mathcal{R}^{(0)}, \mathcal{G}^{(v)} \cup \Delta\mathcal{G}^{(v)}\Big).$$

Result. The resulting deterministic gates retain **31,275 / 41,208** questions; source-level release constraints further reduce the active optimization pool to **27,808**. These items are eligible for subsequent muted-video verification rather than yet being treated as visually verified.

(b) Quota-Constrained Review-Pool Assembly

Objective. For each candidate $i \in \mathcal{C}$, let $x_i \in \{0,1\}$ denote selection and $w_i = s_i - \epsilon h_i$, where $\epsilon = 10^{-6}$, s_i is the deterministic quality score, and h_i is a stable hash used only for tie-breaking. The base constrained selection is

$$S_1 = \arg\max_{\mathbf{x}\in\mathcal{F}} \sum_{i\in\mathcal{C}} w_i x_i.$$

Exact coverage. The feasible set $\mathcal{F}$ enforces

$$|\mathrm{Core}| = 1800, \quad |\mathrm{Holistic}| = |\mathrm{OnlineMemory}| = |\mathrm{CrossView}| = 200;$$
$$N_d^{\mathrm{scene}} = 300, \quad N_{r,c}^{\mathrm{hor}\times\mathrm{cap}} = 150, \quad N_u^{\mathrm{subtask}} = 150,$$

for all $d \in [6]$, $(r,c) \in [4]\times[3]$, and $u \in [12]$ within Core.
Diversity caps. For every applicable source a, horizon–capability cell (r,c), and canonical provenance group g,

$$N_a^{\mathrm{Core}} \le 450, \qquad N_{a,r,c}^{\mathrm{Core}} \le 90, \qquad N_g^{\mathrm{Core}} \le 3, \qquad N_g^{\mathrm{CrossView}} \le 3.$$

Review-pool diversification. To provide replacement redundancy for human validation, we construct a complementary selection over the same feasible set $\mathcal{F}$. After obtaining S_1, we set $w_i^{(2)} = w_i - 0.75$ for $i \in S_1$ ($w_i^{(2)} = w_i$ otherwise), and solve

$$S_2 = \arg\max_{\mathbf{x}\in\mathcal{F}} \sum_{i\in\mathcal{C}} w_i^{(2)} x_i.$$

Review pool and final selection. We merge the two complementary selections into $\mathcal{R} = S_1 \cup S_2$, $|\mathcal{R}| = \mathbf{4,579}$, which serves solely as a diversified candidate pool for human verification and is *not* a released benchmark split. After human validation, we rerun the base constrained selector over the approved candidates to obtain the final **2,400**-question benchmark.

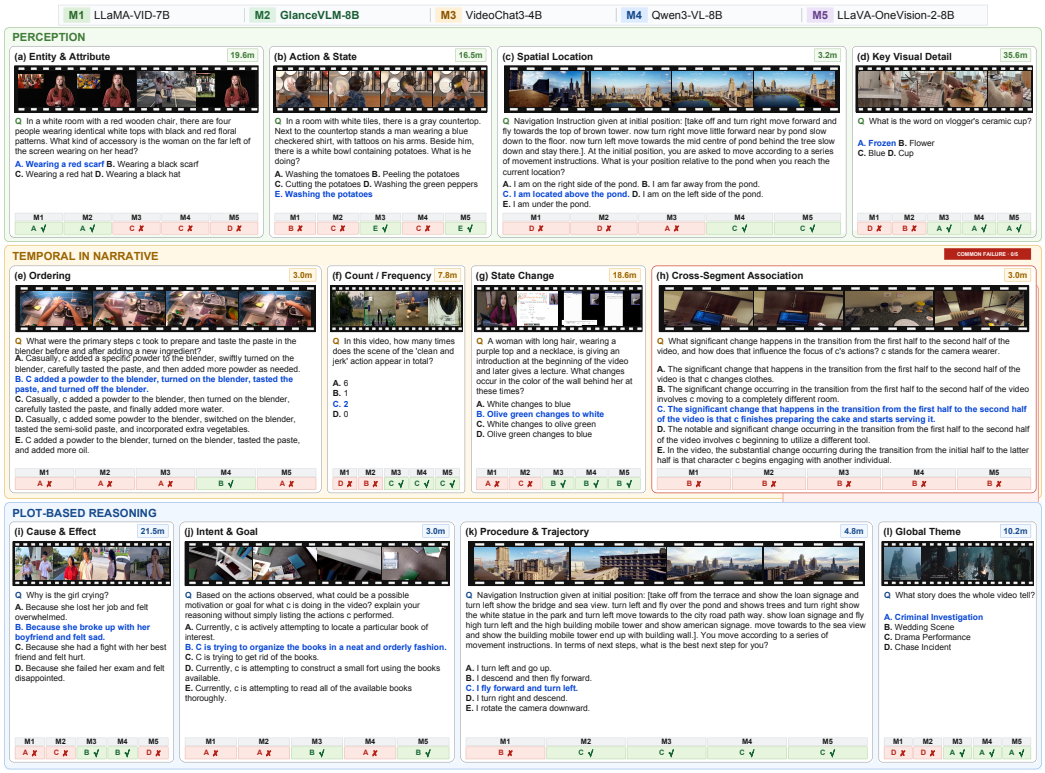


Figure 6: **Representative examples and diagnostic model behaviors on GlanceVU.** The 12 core tasks span three complementary abilities—*Perception*, *Temporal*, and *Reasoning*—with four representative cases per group. Each card presents evidence-bearing moments from the full video, its duration, the question and ground-truth answer, and predictions from five representative video MLLMs; blue marks the ground truth, while green and red indicate correct and incorrect predictions, respectively. The diverse error patterns reveal a persistent gap.

A.2 EVALUATION METRICS

Complementary Evaluation Regimes. We organize the evaluation suite into four complementary settings that progressively increase the demands on temporal coverage and response timing, ranging from preserving decision-relevant semantics in short clips, through retaining and retrieving evidence over long temporal horizons, to supporting causally grounded responses during online video streams; GlanceVU further isolates glance-style video understanding under this progression (§A.1).

(i) Short-video QA. NExT-QA and EgoSchema test whether extreme token compression still preserves answer-critical semantics, temporal relations, and action cues. NExT-QA (Xiao et al., 2021) explicitly stresses causal and temporal reasoning, with approximately 52,000 QA pairs over 5,440 videos. EgoSchema (Mangalam et al., 2023) provides a harder egocentric probe: more than 5,000 five-way questions span over 250 hours of Ego4D video, each grounded in a 3-minute first-person clip where hand–object interactions, viewpoint changes, and action-contingent evidence require integrating information over extended temporal context.

(ii) Long-video understanding. Video-MME, MLVU, LVBench, and LongVideoBench (Fu et al., 2024; Zhou et al., 2025; Wang et al., 2025a; Wu et al., 2024) shift the challenge from local perception to long-range retention and evidence recovery. Video-MME provides broad coverage across domains and temporal scales (900 videos, 2,700 QA pairs, approximately 254 hours, 6 domains and 30 categories, with durations from 11 seconds to 1 hour); MLVU extends this regime to videos spanning 3 minutes to over 2 hours and jointly probes global structure and sparse local detail through diverse long-video tasks, including reasoning, retrieval, ordering, counting, captioning, and summarization. LVBench further stresses extreme long-context understanding with 103 videos totaling approximately 117 hours, evaluating capabilities including key-information retrieval, temporal grounding, event understanding, reasoning, and summarization over hour-scale visual contexts. LongVideoBench complements these benchmarks by targeting *referring reasoning* with 6,678 human-written questions over 3,763 videos, directly exposing failures to retain or recover the correct temporal context after aggressive compression.

(iii) Timestamp-bounded video understanding. StreamingBench-Realtime (Lin et al., 2026) and the RealTime and Backward subsets of OVO-Bench (Niu et al., 2025) test whether the compressed representation remains temporally actionable under causal access to the video stream. StreamingBench contains 900 videos, 4,500 timestamped QA pairs, and 18 tasks, with its Real-Time Visual Understanding split requiring responses without access to future frames. OVO-Bench contains 644 videos and about 2,800 timestamped annotations, complementing this setting by separating Real-Time Visual Perception from Backward Tracing, which respectively probe responses to ongoing events and recovery of task-relevant past evidence from compressed causal memory.

(iv) Glance video understanding. To directly isolate the capability targeted in this work, we further construct GlanceVU, a benchmark specifically curated for glance-style video understanding by selecting and reorganizing video QA instances under a unified evaluation protocol. Unlike the preceding benchmarks, GlanceVU explicitly targets understanding from sparse, highly compressed visual observations, providing a more direct measure of whether a model can retain decision-relevant evidence under glance-level visual budgets. We detail its construction, composition, and task distribution in §A.1.

A.3 MORE RESULTS

Cross-backbone robustness and efficiency. Table 6 further examines the robustness and efficiency of our two-token representation as the complements of Table 3 in §4.3. On VideoMME, GlanceVLM improves direct two-token compression by $5.5/8.1/12.3$ points on Qwen2.5-VL-7B, Qwen3-VL-4B, and Qwen3-VL-8B, with gains across all duration ranges. For Qwen3-VL-8B, the gain decreases from $+19.4$ on short to $+4.5$ points on long videos, indicating that long-horizon information preservation remains more challenging; Table 3 provides the corresponding capability-wise breakdown. Controlled variants further show that GlanceVLM outperforms text-only, adaptive two-token pooling, and the denser eight-frame control by $14.6/4.8/1.9$ points. Across benchmarks, it retains only 1.05–1.11% of the original visual tokens. At nearly identical retained budgets, GlanceVLM also surpasses TimeChat-Online by 3.31 points on VideoMME and 4.08 points on GlanceVU, while requiring lower latency. Together, these results demonstrate that the gains generalize and arise from a more informative representation rather than the token count alone.

(a) Cross-backbone comparison on Video-MME

Bkb.	Cond.	S	M	L	Tot.
Q3-8B	+2T	45.0	45.7	47.1	45.9 ↓ **12.3**
	Ours	64.4	58.7	51.6	58.2
	Txt-only	44.0	42.0	44.7	43.6
	T2S-8U	65.4	53.9	49.7	56.3
	Pool-2T	63.0	49.3	47.9	53.4
Q2.5-7B	+2T	45.8	47.3	44.0	45.7 ↓ **5.5**
	Ours	52.8	52.6	48.2	51.2
Q3-4B	+2T	47.7	45.7	46.1	46.5 ↓ **8.1**
	Ours	60.6	54.9	48.3	54.6

(b) Benchmark-wise efficiency of our 8B model, TimeChatOnline and Qwen3-VL-8B-Instruct

Bench.	Orig.	Ret.	Lat.	Tok/s	Comp. (×)	Acc.
EgoSchema	16,201.68	179.77	1.75	228.73	90.12	52.87
NextQA	3,540.93	39.13	0.51	300.39	90.49	75.36
MLVU	47,090.18	494.00	5.63	108.01	95.32	63.58
StreamingBench	23,367.84	259.99	2.30	174.91	89.88	61.64
OVOBench R+B	37,193.08	404.76	3.39	151.50	91.89	47.32
VideoMME	40,629.35	451.78	7.62	76.22	89.93	58.20
GlanceVU	37,946.99	412.51	6.34	89.85	91.99	50.00
VMME-TimeChat	56,854.35	451.78	10.52	53.69	125.85	54.89 ↓ **3.3**
GVU-TimeChat	49,500.76	410.47	7.16	77.86	120.59	45.92 ↓ **4.1**
MLVU-Qwen3-8B	47,090.18	-	-	-	-	76.90

Table 6: **Compact side-by-side presentation of cross-backbone comparison and benchmark-wise efficiency.** All results are reported under the unified setting of **1 FPS** and a maximum of **768** frames. Left (a): Video-MME comparison across backbones and conditions. The experimental setting is exactly the same as in Table 3. Right (b): original tokens (Orig.), retained tokens (Ret.), latency (Lat.), input token throughput (Tok/s), compression ratio (Comp.), and accuracy (Acc.) of our 8B model across benchmarks. OVOBench R+B = OVOBench RealTime + Backward, for OVO-Bench R+B, accuracy is computed over all evaluated questions in the Real-Time and Backward subsets combined; VMME = VideoMME; GVU = GlanceVU.

What does relaxing the visual bottleneck reveal on GlanceVU? We examine this question under two distinct token-reduction mechanisms (Fig. 7). For TimeChat-Online, increasing the feature-level threshold τ retains more visual tokens under its original Differential Token Dropping (DTD) rule; for GlanceVLM, we directly restore a fraction $r \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ of the native backbone tokens at inference. Both models benefit from additional visual evidence, yet neither improves monotonically with token count. TimeChat-Online rises from 45.92% at its 2T setting to 49.63% at $\tau = 0.2$, before dropping to 48.38% at $\tau = 0.3$. Likewise, GlanceVLM improves from 50.00% to 54.17% at $r = 0.6$, while full restoration yields 53.25%. Notably, GlanceVLM's original 2T representation already slightly exceeds the best TimeChat-Online result in this sweep (50.00 vs. 49.63). Overall, performance on GlanceVU improves most when the added tokens carry useful visual evidence; increasing token quantity alone provides only limited returns.

Distinct representations yield different recovery profiles. The task-wise gains make this distinction more explicit. For TimeChat-Online, improvements are relatively concentrated in detail- and state-sensitive categories: Action State and Entity Attribute gain $+9.95$ and $+8.00$ points at $\tau = 0.2$, while Key Visual Detail reaches $+11.34$ at $\tau = 0.3$. GlanceVLM exhibits a broader positive response across the taxonomy: at $r = 0.6$, Entity Attribute, Ordering, Intent Goal Motivation, Global Theme Summary, and State Change Progress improve by $+10.67$, $+6.49$, $+5.84$, $+5.15$, and $+5.02$ points, respectively, with simultaneous gains across all four duration ranges. This difference is also apparent in Fig. 7, where positive gains extend across a broader set of categories for GlanceVLM.

Representation design shapes how recovered evidence is utilized. The contrasting profiles align with how the two methods construct their compact visual representations. TimeChat-Online removes tokens judged redundant across adjacent frames; relaxing DTD therefore restores evidence that its original dropping rule had filtered out, leading to particularly pronounced gains on concrete state- and detail-sensitive tasks. GlanceVLM instead builds its 2T representation from two complementary learned tokens; restoring native backbone tokens on top of this interface supplies additional evidence that can complement both components, producing gains across a wider range of perceptual, temporal, and global-understanding categories. Together, these results show that GlanceVU can diagnose not only how much visual evidence becomes available as the bottleneck is relaxed, but also how the structure of a compact representation determines where that evidence translates into improved glance-style understanding.

(a) TimeChat-Online on GlanceVU

	Horizon				Capability											Overall
	S	M	L	XL	AS	CEC	CFD	CSA	EA	GTS	IGM	KVD	Ord.	PTP	SL	SCP	All
2T	41.15 (+0.00)	40.20 (+0.00)	47.68 (+0.00)	56.43 (+0.00)	39.30 (+0.00)	54.00 (+0.00)	34.67 (+0.00)	44.71 (+0.00)	36.67 (+0.00)	66.29 (+0.00)	35.02 (+0.00)	51.33 (+0.00)	48.65 (+0.00)	42.67 (+0.00)	42.00 (+0.00)	42.14 (+0.00)	45.92 (+0.00)
$\tau = 0.1$	41.41 (+0.26)	39.70 (-0.50)	50.00 (+2.32)	59.18 (+2.75)	40.80 (+1.50)	58.00 (+4.00)	37.33 (+2.66)	42.31 (-2.40)	42.00 (+5.33)	67.43 (+1.14)	34.63 (-0.39)	56.67 (+5.34)	49.73 (+1.08)	44.67 (+2.00)	44.67 (+2.67)	38.80 (-3.34)	47.00 (+1.08)
$\tau = 0.2$	45.41 (+4.26)	41.72 (+1.52)	53.80 (+6.12)	59.69 (+3.26)	49.25 (+9.95)	60.00 (+6.00)	38.67 (+4.00)	48.08 (+3.37)	44.67 (+8.00)	69.14 (+2.85)	40.86 (+5.84)	57.33 (+6.00)	51.89 (+3.24)	43.33 (+0.66)	43.33 (+1.33)	39.46 (-2.68)	49.63 (+3.71)
$\tau = 0.3$	44.61 (+3.46)	44.43 (+4.23)	52.74 (+5.06)	53.69 (-2.74)	50.25 (+10.95)	56.67 (+2.67)	33.33 (-1.34)	44.71 (+0.00)	40.67 (+4.00)	64.00 (-2.29)	41.25 (+6.23)	62.67 (+11.34)	50.27 (+1.62)	42.00 (-0.67)	45.33 (+3.33)	41.14 (-1.00)	48.38 (+2.46)

(b) GlanceVLM on GlanceVU

	Horizon				Capability											Overall
	S	M	L	XL	AS	CEC	CFD	CSA	EA	GTS	IGM	KVD	Ord.	PTP	SL	SCP	All
2T	43.54 (+0.00)	48.31 (+0.00)	53.38 (+0.00)	57.29 (+0.00)	50.25 (+0.00)	63.33 (+0.00)	42.00 (+0.00)	47.60 (+0.00)	45.33 (+0.00)	71.71 (+0.00)	43.19 (+0.00)	56.00 (+0.00)	48.65 (+0.00)	42.67 (+0.00)	44.00 (+0.00)	36.12 (+0.00)	50.00 (+0.00)
0.2	46.60 (+3.06)	51.18 (+2.87)	59.28 (+5.90)	61.23 (+3.94)	54.73 (+4.48)	64.00 (+0.67)	42.67 (+0.67)	48.56 (+0.96)	56.00 (+10.67)	75.71 (+4.00)	50.58 (+7.39)	58.67 (+2.67)	55.14 (+6.49)	43.33 (+0.66)	44.00 (+0.00)	40.13 (+4.01)	53.79 (+3.79)
0.4	46.34 (+2.80)	52.03 (+3.72)	58.23 (+4.85)	61.58 (+4.29)	52.24 (+1.99)	65.33 (+2.00)	43.33 (+1.33)	50.48 (+2.88)	55.33 (+10.00)	76.00 (+4.29)	49.81 (+6.62)	58.00 (+2.00)	56.22 (+7.57)	42.67 (+0.00)	45.33 (+1.33)	39.46 (+3.34)	53.79 (+3.79)
0.6	46.60 (+3.06)	53.21 (+4.90)	58.44 (+5.06)	61.41 (+4.12)	53.23 (+2.98)	65.33 (+2.00)	45.33 (+3.33)	51.44 (+3.84)	56.00 (+10.67)	76.86 (+5.15)	49.03 (+5.84)	56.00 (+0.00)	55.14 (+6.49)	42.67 (+0.00)	45.33 (+1.33)	41.14 (+5.02)	54.17 (+4.17)
0.8	47.14 (+3.60)	51.01 (+2.70)	56.75 (+3.37)	61.23 (+3.94)	52.74 (+2.49)	66.00 (+2.67)	40.00 (-2.00)	50.48 (+2.88)	55.33 (+10.00)	76.57 (+4.86)	47.86 (+4.67)	56.67 (+0.67)	55.14 (+6.49)	45.33 (+2.66)	41.33 (-2.67)	40.47 (+4.35)	53.42 (+3.42)
1.0	44.74 (+1.20)	51.69 (+3.38)	57.38 (+4.00)	62.44 (+5.15)	53.23 (+2.98)	67.33 (+4.00)	49.33 (+7.33)	45.67 (-1.93)	57.33 (+12.00)	74.57 (+2.86)	47.08 (+3.89)	54.67 (-1.33)	55.14 (+6.49)	42.00 (-0.67)	41.33 (-2.67)	41.47 (+5.35)	53.25 (+3.25)

Figure 7: **Token restoration on GlanceVU. (a) TimeChat-Online.** We vary its feature-level token-dropping threshold τ, where τ controls the drop operation in this model; a smaller τ removes more visual tokens. **(b) GlanceVLM.** We progressively restore 20%–100% of the backbone's native visual tokens from the two-token-per-frame (**2T**) boundary. In both panels, each cell reports the absolute accuracy in bold and, in parentheses, the change in percentage points relative to the corresponding 2T baseline; color encodes the same change. S/M/L/XL denote video-length ranges; the remaining abbreviations denote the twelve GlanceVU capability categories.

Do Anchor and Innovation learn distinct roles? We complement component ablations with spatial, temporal, and geometric diagnostics. Spatial overlays in Fig. 9 & 10 & 11 & 12 show that Innovation token is routed toward localized regions of pronounced inter-frame change, capturing shifts in the actor's pose and ongoing actions together with nearby object-state changes, whereas Anchor token maintains broader scene-level support. Across all 900 VideoMME videos (Fig. 8), Anchor token is substantially smoother across adjacent frames ($\sim$0.89 vs. $\sim$0.73), while the within-frame role distance remains near 0.30 throughout. Innovation token also maintains $\sim$0.70 cross-attention entropy, consistent with aggregating multiple residual-selected regions rather than collapsing onto one. A 40-frame PCA example further shows coherent yet clearly separated token trajectories. Together, these diagnostics show that the pair does not collapse into redundant frame summaries: Anchor token provides a persistent contextual reference, whereas Innovation carries change-sensitive visual evidence.

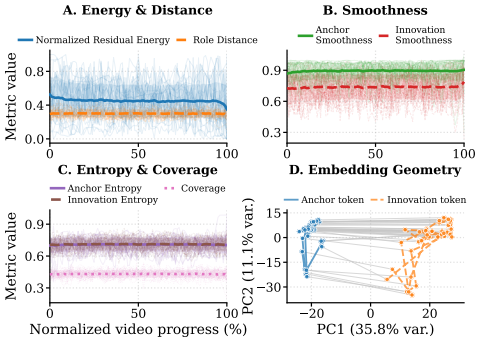


Figure 8: **Token-role diagnostics on VideoMME. (A)** Residual energy and within-frame role distance. **(B)** Adjacent-frame smoothness. **(C)** Cross-attention entropy and Innovation coverage. **(D)** PCA trajectories over a 40-frame clip. Definitions are provided in Appendix A.4.

A.4 Detailed Definitions of Token-Role Diagnostics

We provide the exact definitions underlying the four token-role diagnostics in Fig. 8. Together, they characterize how visual dynamics are decomposed into two role-specialized representations: Panel (A) relates input variation to within-frame role separation, Panel (B) measures temporal persistence, Panel (C) analyzes spatial evidence allocation, and Panel (D) visualizes their learned embedding geometry.

Notation. After passing through video encoder, frame t is represented by $\mathbf{X}_t = [\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,N}]^\top \in \mathbb{R}^{N \times D}$, where N and D denote the number and hidden dimension of spatial tokens. The two-token compressor produces an *Anchor* token $\mathcal{A}_t \in \mathbb{R}^D$ and an *Innovation* token $\mathcal{I}_t \in \mathbb{R}^D$.

Benchmark-level temporal aggregation. Panels (A)–(C) summarize frame-wise statistics over each benchmark. Videos are sampled at 1 FPS with at most 768 frames, and each trajectory is linearly resampled to 101 uniformly spaced positions on a normalized 0%–100% progress axis. Faint curves show up to 60 individual trajectories, whereas the thick curve and shaded region denote the pointwise mean and one sample standard deviation over the *full benchmark*.

Panel (A): visual change and role separation. Inter-frame change is measured in the visual feature space. For $t > 0$, the residual at spatial position i is $d_{t,i} = \|\mathbf{x}_{t,i} - \mathbf{x}_{t-1,i}\|_2$; for the first frame, we use $d_{0,i} = \|\mathbf{x}_{1,i}\|_2$. After reshaping the residuals to the spatial token grid, we apply 3×3 average pooling and spatial averaging. The resulting energy is independently min–max normalized within each video:

$$E_t = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AvgPool}_{3\times 3}(\mathbf{d}_t)_i, \qquad \widetilde{E}_t = \frac{E_t - \min_\tau E_\tau}{\max_\tau E_\tau - \min_\tau E_\tau}. \tag{5}$$

Interpretation. $\widetilde{E}_t$ measures the relative strength of visual change at time t within the same video.

We measure within-frame role separation by the cosine distance between the two token states:

$$D_t = 1 - \frac{\mathcal{A}_t^\top \mathcal{I}_t}{\|\mathcal{A}_t\|_2 \|\mathcal{I}_t\|_2}. \tag{6}$$

Interpretation. A larger D_t indicates a stronger directional distinction between *Anchor* and *Innovation*; its evolution alongside $\widetilde{E}_t$ reveals whether this separation persists as the visual content changes.

Panel (B): temporal smoothness. The persistence of each role is quantified by cosine similarity between adjacent token states:

$$S_t^A = \frac{\mathcal{A}_t^\top \mathcal{A}_{t-1}}{\|\mathcal{A}_t\|_2 \|\mathcal{A}_{t-1}\|_2}, \qquad S_t^I = \frac{\mathcal{I}_t^\top \mathcal{I}_{t-1}}{\|\mathcal{I}_t\|_2 \|\mathcal{I}_{t-1}\|_2}, \quad t \geq 1. \tag{7}$$

Interpretation. Values near one indicate a persistent temporal state, whereas lower values reflect stronger representation updates. Thus, S_t^A captures the stability of the scene-level context maintained by *Anchor*, while S_t^I reflects the responsiveness of *Innovation* to newly emerging evidence.

Panel (C): spatial evidence distribution. We extract role-specific cross-attention from the final dual token layer and average the weights across attention heads. Because *Anchor* reads the complete frame-token grid, its attention is normalized over all N spatial positions as $\{p_{t,i}^A\}_{i=1}^N$.

For *Innovation*, the residual map identifies up to four peaks, each expanded into a local neighborhood of at most 5×5 spatial tokens. Previously selected locations are suppressed during subsequent peak selection, encouraging distinct change-bearing regions. The resulting attention weights are projected back onto the spatial grid, with overlapping contributions averaged and normalized as $\{p_{t,i}^I\}_{i=1}^N$. Their normalized attention entropies are

$$H_t^A = -\frac{\sum_{i=1}^N p_{t,i}^A \log p_{t,i}^A}{\log N}, \qquad H_t^I = -\frac{\sum_{i=1}^N p_{t,i}^I \log p_{t,i}^I}{\log N}. \tag{8}$$

We additionally define the spatial coverage of *Innovation* as $C_t = \left|\bigcup_{r=1}^R \Omega_{t,r}\right| / N$, where $\Omega_{t,r}$ denotes the r-th selected region and overlapping patches are counted only once.

Interpretation. Entropy measures how broadly each token distributes its visual support, whereas C_t measures how much of the visual-token grid is made available to *Innovation*. Their joint evolution therefore separates evidence dispersion from spatial support under the restricted routing budget.

Panel (D): representative embedding trajectory. Panel (D) complements the benchmark-level statistics with one representative 40-frame probe. We jointly arrange its token states as $\mathbf{Z} = [\mathcal{A}_0, \mathcal{I}_0, \ldots, \mathcal{A}_{T-1}, \mathcal{I}_{T-1}]^\top \in \mathbb{R}^{2T \times D}$, center all representations, and project them onto a shared two-dimensional PCA basis:

$$\mathbf{Y} = (\mathbf{Z} - \bar{\mathbf{Z}})\mathbf{V}_{1:2}^\top, \qquad \rho_k = \frac{\sigma_k^2}{\sum_j \sigma_j^2}. \tag{9}$$

Here, ρ_k denotes the fraction of total token-state variance explained by principal component k. The blue solid and orange dashed curves trace the trajectories of *Anchor* and *Innovation* token, respectively, while gray segments pair the two representations produced from the same frame.

Interpretation. Separated yet individually coherent trajectories indicate that the two roles occupy distinct embedding structures while preserving the temporal organization of each token stream. The same representative probe is used across benchmarks within each parameter scale; the original hidden dimensions are $D = 4096$ for the 8B model and $D = 2560$ for the 4B model.

Summary. These diagnostics connect visual change and within-frame role separation, temporal persistence, spatial evidence allocation, and embedding-space evolution. Together, they show how *Anchor* token $\mathcal{A}_t$ maintains a stable scene-level state while *Innovation* token $\mathcal{I}_t$ adaptively captures residual visual evidence.

B Ethics Statement

This work adheres to the ICLR Code of Ethics. No new human-subject studies or animal experiments were conducted. All datasets used for training are obtained from publicly available or open-source resources, and the source datasets used to construct GlanceVU are likewise publicly available. We use these resources in accordance with their stated licenses and usage conditions and do not introduce or release new private or personally identifying information. As with other research based on existing video datasets, the models and benchmark may inherit biases or limitations present in their source data. We therefore report our experimental settings and data construction procedures transparently and encourage the results to be interpreted within the scope of the evaluated datasets.

C AI Use Statement

Generative AI tools, including large language models (LLMs), were used to assist with language editing of author-written text, including grammar correction, minor sentence-level rephrasing, and improvements to clarity, coherence, and readability. During data processing, LLMs were also used to conduct a semantic audit of candidate annotations, as described in Appendix § A.1. LLMs were not used for research ideation, methodological or experimental design, data analysis, or interpretation of results. All research ideas, technical methods, experimental analyses, and conclusions were conceived, developed, and validated by the authors. All LLM-assisted outputs were reviewed by the authors, who take full responsibility for the final content of this work.



Figure 9: **Visualization of four successful examples of GlanceVLM-8B.** From top to bottom: original frames, residual-guided regions selected for the Innovation token, and Anchor-token support, with brighter colors indicating stronger responses. $\mathcal{I}_t$ targets localized object and hand–object changes, whereas $\mathcal{A}_t$ retains broad scene context, revealing complementary visual roles. The examples cover complementary video-understanding demands, including cross-frame grounding of evolving cues, holistic procedural summarization, distractor-resistant ordering in a post-processed video with interleaved irrelevant content, and fine-grained short-range temporal reasoning.



Figure 10: **Visualization of four successful examples of GlanceVLM-8B.** The examples span start–end spatial grounding, counterfactual causal reasoning, global-theme recognition, and fine-grained entity-attribute identification, covering both holistic and localized evidence demands.

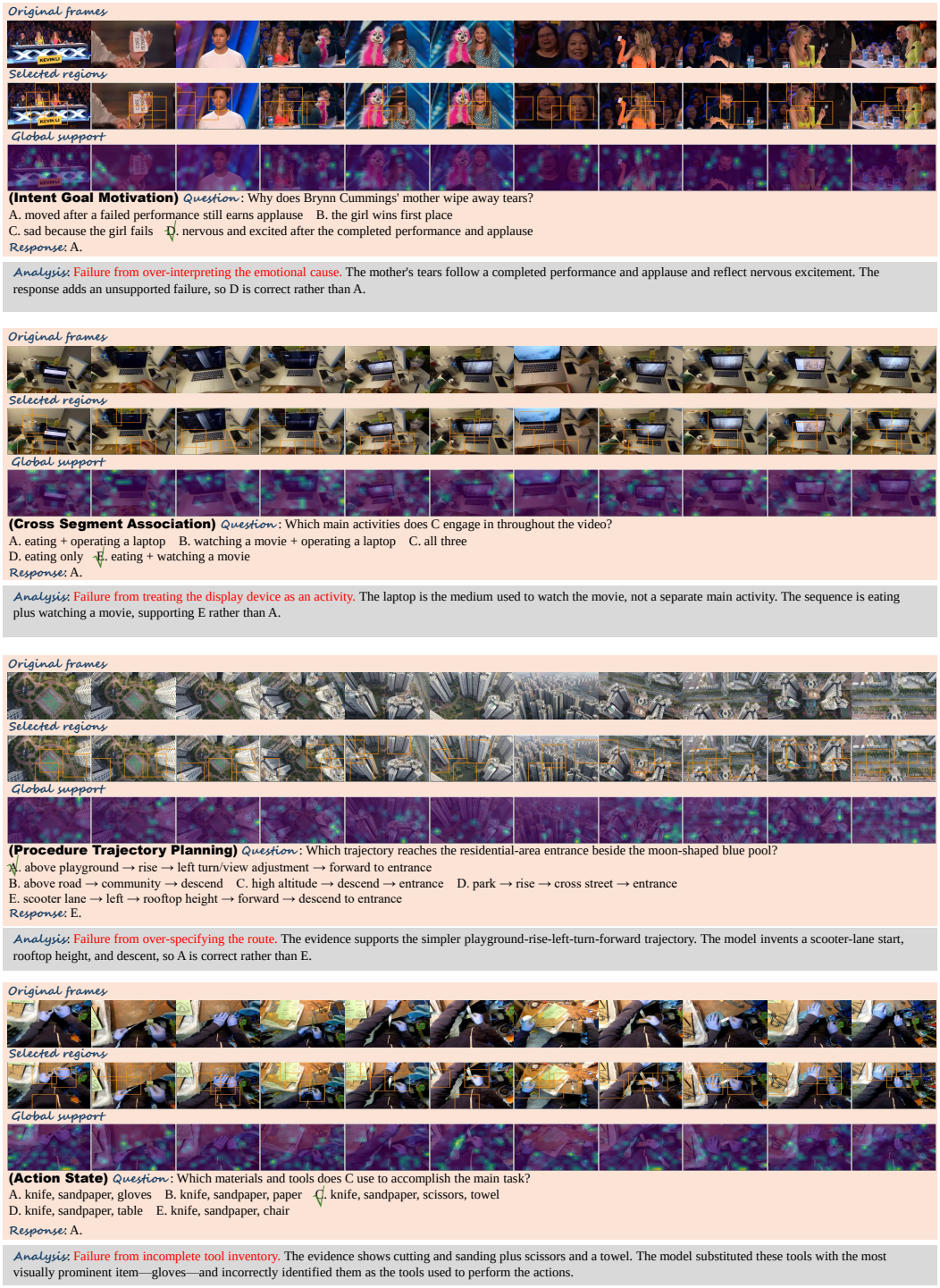


Figure 11: **Visualization of four failure examples of GlanceVLM-8B.** The cases expose distinct reasoning failures despite largely relevant visual support: over-interpreting an emotional cause, conflating a display medium with an activity, over-specifying a spatial trajectory, and incompletely aggregating the required tools. These examples suggest that the remaining errors arise less from missing salient evidence than from incorrectly interpreting or composing the grounded evidence into the final decision.

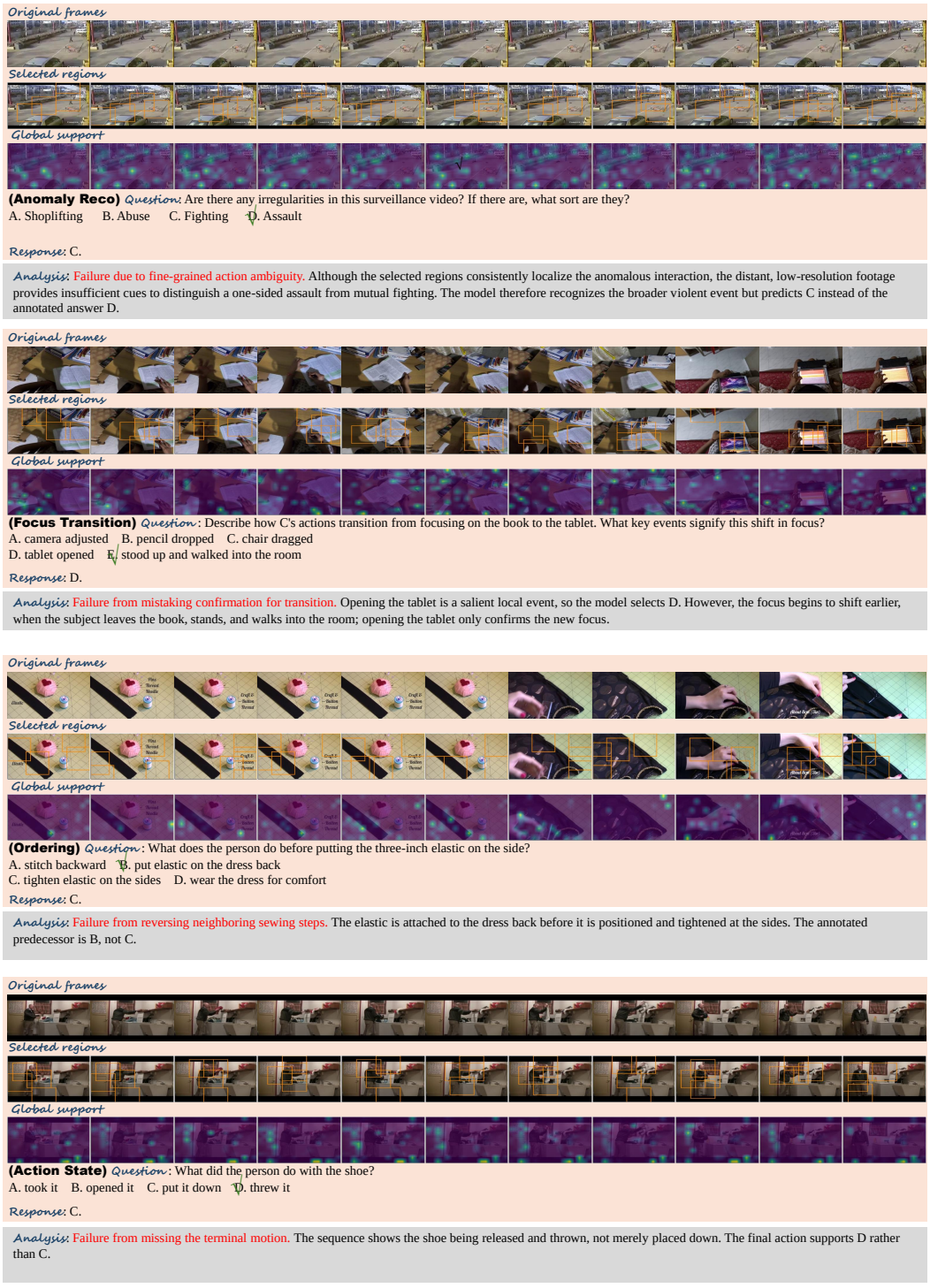


Figure 12: **Visualization of four failure examples of GlanceVLM-8B.** These cases expose fine-grained temporal ambiguities, including confusing closely related actions, misidentifying a focus-transition boundary, reversing adjacent procedural steps, and missing the terminal motion of an action. Although the relevant events are largely localized, the errors highlight the remaining difficulty of resolving precise action states and temporal relations from highly compressed visual evidence.