APPENDIX

A FRAMES VISUALIZATION

This section visualizes sample input keyframes from the videos to provide a clearer understanding of the data used in our experiments. Each row in Figure 7 represents a distinct video sequence fed into the LVLMs for analysis. These examples are representative of the scenarios in our dataset, encompassing a variety of everyday actions, objects, and environments. By visualizing the raw inputs, we aim to illustrate the visual complexities, such as changes in viewpoint, object scale, and partial occlusions, that the model must handle to perform accurate semantic reasoning.

Figure 7: Visualization of Input Keyframes. Each row displays a sequence of frames provided to the model as input for a specific video. The red bounding boxes highlight the ground-truth object pertinent to the task's question (e.g., the object being picked up, kicked, or taken). It is important to note that these bounding boxes are included here for clarity and were not provided to the model during inference.

B THEORETICAL FRAMEWORK FOR VISUAL SEMANTIC CIRCUITS

In this section, we propose the theoretical framework built on three core principles that we hypothesize govern the internal computations of LVLMs: Information Localization, Progressive Semantic Refinement, and a Two-Stage Reasoning flow. We model the LVLM as a probabilistic system to formalize these principles

into specific, falsifiable predictions. This framework provides a principled foundation for understanding and predicting the model's behavior, which we then validate through targeted experiments.

B.1 Probabilistic Model Formulation

Let V be a video represented by a sequence of key frames, and Q be a textual question. The LVLM, denoted by $\mathcal{M}$, aims to generate an answer A. The process begins by encoding the video V into a set of N visual tokens, $E_V = \{e_1, e_2, \ldots, e_N\}$. The question Q is tokenized into M text tokens, $E_Q = \{q_1, \ldots, q_M\}$. The model then computes the probability of an answer A:

$$P(A|V, Q) = \mathcal{M}(E_V, E_Q). \tag{3}$$

Central to our investigation is the hypothesis that the set of visual tokens E_V can be partitioned into the subset E_O containing primary information about the specific object o, and the complementary subset E_C containing contextual information, such that $E_V = E_O \cup E_C$ and $E_O \cap E_C = \emptyset$.

B.2 Principle of Information Localization

We begin with the hypothesis that to answer a specific question, the LVLM does not treat all visual tokens equally. Instead, we propose the *Principle of Information Localization*: task-critical information is spatially concentrated in the subset of tokens corresponding to the object of interest. Let this subset be E_O, with the remainder being contextual tokens E_C.

This principle leads to a direct, testable prediction. The informational value of a token set can be quantified by the degradation in model performance upon its ablation. We model this degradation as the KL divergence between the original and ablated posterior distributions:

$$\mathcal{L}_{\text{drop}}(E_S) = D_{KL}\left(P(A|E_V, E_Q) \parallel P(A|E_V \setminus E_S, E_Q)\right). \tag{4}$$

Our principle predicts that the information is concentrated in E_O. Formally, if we ablate the object tokens E_O, the resulting information loss should be significantly greater than ablating any other random subset of tokens E_R of the same size. This leads to the following inequality, which we aim to verify experimentally:

$$\mathbb{E}_{E_R \subset E_V, |E_R| = |E_O|}\left[\mathcal{L}_{\text{drop}}(E_R)\right] \ll \mathcal{L}_{\text{drop}}(E_O). \tag{5}$$

Furthermore, we hypothesize that the model internally reasons over abstract concepts. This predicts that injecting a clean, symbolic representation of the object, $e_{w_{\text{correct}}}$, should be even more effective than the noisy visual tokens E_O. This can be formalized as:

$$P(A^*|e_{w_{\text{correct}}}, E_C, E_Q) > P(A^*|E_O, E_C, E_Q). \tag{6}$$

The ablation and injection experiments presented in Table 1 were designed to test these formal predictions.

B.3 Hypothesis of Progressive Semantic Refinement

We hypothesize that visual information is not processed into its final semantic form in a single step. Instead, we propose the model of *Progressive Semantic Refinement*, where hidden states associated with visual tokens transition from encoding low-level perceptual features in early layers to abstract, language-aligned concepts in later layers.

Let $h_i^{(l)}$ be the hidden state for a token i at layer l. Let $\mathcal{S}(w_{\text{correct}})$ be the semantic space associated with the correct object concept, represented by its text embedding $e_{w_{\text{correct}}}$. Our hypothesis predicts that for an object token $i \in E_O$, its representation $h_i^{(l)}$ will become progressively more aligned with this semantic space as it passes through the network. We can formalize this predicted monotonic increase in alignment for layers l beyond a critical depth l_{crit} using a similarity metric:

$$\text{sim}(h_i^{(l)}, e_{w_{\text{correct}}}) \quad \text{is a monotonically increasing function of } l \text{ for } l > l_{\text{crit}}. \tag{7}$$

To test this prediction, we employ the logit lens technique, which projects intermediate hidden states into the vocabulary space. We measure the Correspondence Rate ($C_R^{(l)}$) and Answer Probability ($A_P^{(l)}$) to track this alignment across layers. The experimental results in Figure 4 are used to validate the existence and location of the predicted critical layer depth l_{crit}.

B.4 The Two-Stage Reasoning Hypothesis

Building on the previous principles, we hypothesize that the model's reasoning is not monolithic but follows an efficient, cognitively plausible two-stage process.

1. **Stage 1 (Contextual Grounding):** In the early layers ($\mathcal{L}_{\text{early}}$), the model first processes contextual tokens (E_C) to establish a general understanding of the scene and the query.

2. **Stage 2 (Focal-Point Refinement):** In the late layers ($\mathcal{L}_{\text{late}}$), after the context is established, the model focuses its attention on the specific object tokens (E_O) to extract fine-grained details necessary for a precise answer.

This hypothesis can be formalized by considering the sensitivity of the final prediction to attention weights at different layers. Let $\nabla_{\alpha_S^{(l)}} \log P(a_1^*)$ be the gradient of the log-probability of the correct answer with respect to the attention weights from a set of tokens S at layer l. Our two-stage hypothesis predicts a shift in sensitivity:

$$\sum_{l \in \mathcal{L}_{\text{early}}} \left\| \nabla_{\alpha_C^{(l)}} \log P(a_1^*) \right\| > \sum_{l \in \mathcal{L}_{\text{early}}} \left\| \nabla_{\alpha_O^{(l)}} \log P(a_1^*) \right\|, \tag{8}$$

$$\sum_{l \in \mathcal{L}_{\text{late}}} \left\| \nabla_{\alpha_O^{(l)}} \log P(a_1^*) \right\| > \sum_{l \in \mathcal{L}_{\text{late}}} \left\| \nabla_{\alpha_C^{(l)}} \log P(a_1^*) \right\|. \tag{9}$$

These inequalities formalize the "context-first, detail-later" strategy as a testable prediction. We designed the attention masking experiments in Table 2 to directly probe these sensitivities and validate our two-stage reasoning hypothesis.

C Scaling Experiments

We conducted supplementary experiments to verify that our conclusions generalize to larger-scale models. We replicated our core analyses on the LLaVA-NeXT-34B model variants, with results that closely mirror those presented in the main body of the paper. The visual token ablation study on the 34B models reaffirms the principle of spatial localization for semantic information. Ablating object-specific tokens incurs significantly more substantial performance degradation than removing larger quantity of random tokens (in Table 3).

Furthermore, our semantic tracing analysis on the 34B architecture, depicted in Figure 8, reveals a conceptual emergence pattern consistent with our earlier observations. Both the Correspondence Rate and Answer Probability remain negligible through the initial layers before exhibiting a sharp, concurrent rise beginning around layer 40. This trend indicates that abstract, language-aligned concepts are consolidated in the deeper layers of the network, irrespective of model scale. These scaling experiments provide robust evidence that the mechanisms of semantic localization and late-stage conceptual formation are fundamental properties of the tested LVLM architectures.

Ablation Type	Tokens Number	LLaVA-NeXT-34B-I Open	LLaVA-NeXT-34B-I Close	LLaVA-NeXT-34B-V Open	LLaVA-NeXT-34B-V Close
Control Groups (Low Tokens)					
Baseline	0	$100.0_{\downarrow 0}$	$100.0_{\downarrow 0}$	$100.0_{\downarrow 0}$	$100.0_{\downarrow 0}$
Register	13	$55.3_{\downarrow 44.68}$	$96.9_{\downarrow 3.14}$	$55.3_{\downarrow 44.68}$	$1.8_{\downarrow 98.2}$
Object-based Ablation					
Object	304	$9.3_{\downarrow 90.75}$	$29.8_{\downarrow 70.16}$	$14.6_{\downarrow 85.37}$	$11.7_{\downarrow 88.35}$
	413	$5.8_{\downarrow 94.24}$	$21.1_{\downarrow 78.88}$	$10.7_{\downarrow 89.32}$	$11.3_{\downarrow 88.72}$
	573	$5.8_{\downarrow 94.24}$	$18.0_{\downarrow 82.02}$	$10.3_{\downarrow 89.74}$	$11.8_{\downarrow 88.21}$
Control Groups (High Tokens)					
Random	100	$54.3_{\downarrow 45.72}$	$96.9_{\downarrow 3.14}$	$93.1_{\downarrow 6.92}$	$29.7_{\downarrow 70.26}$
	350	$55.5_{\downarrow 44.5}$	$96.5_{\downarrow 3.49}$	$88.5_{\downarrow 11.54}$	$31.0_{\downarrow 68.97}$
	500	$54.5_{\downarrow 45.55}$	$96.3_{\downarrow 3.66}$	$83.1_{\downarrow 16.92}$	$30.8_{\downarrow 69.23}$
	900	$59.9_{\downarrow 40.14}$	$96.8_{\downarrow 3.17}$	$78.3_{\downarrow 21.67}$	$29.7_{\downarrow 70.33}$

Table 3: Accuracy (%) from visual token ablation on question-answering performance across 34B models. The ↓ symbol indicates the magnitude of this performance drop.

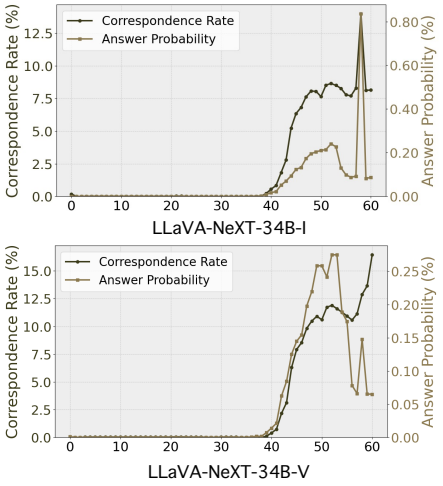

Figure 8: Quantitative analysis of semantic tracing on 34B model size.

Prompt
USER: The input consists of a sequence of key frames from a video. Question: Which object was taken by the person?
ASSISTANT: The object is the

***Answer*:** Towel

Response: Towel

Token/Layer	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5
<s>	Institution	Архив	sierp	sierp	sierp
US	ona	Архив	pert	tap	pert
ER	Session	пута	oded	庄	pel
:	Portail	пута	<s>	<s>	Sav
		пута	Carter	◆	◆
<IMG001>	sak	sak	sak	sak	sak
<IMG002>	gresql	gresql	gresql	gresql	gresql
<IMG003>	olas	olas	olas	olas	olas
<IMG004>	<<	<<	<<	<<	<<
<IMG005>	gresql	gresql	gresql	gresql	gresql
<IMG006>	sak	sak	sak	sak	sak
<IMG007>	yter	yter	yter	yter	yter
<IMG008>	nahm	nahm	nahm	nahm	nahm
<IMG009>	loop	loop	loop	loop	sak
<IMG010>	loop	loop	loop	loop	loop
<IMG011>	gresql	gresql	gresql	gresql	gresql
<IMG012>	yter	yter	yter	yter	yter
<IMG013>	izi	izi	Emp	Emp	conf
<IMG014>	alle	alle	alle	alle	conf
<IMG015>	yter	yter	yter	yter	yter
<IMG016>	alle	alle	alle	arab	anno
<IMG017>	yter	yter	yter	yter	yter

Figure 9: Qualitative example of the model correctly identifying an object. The user asks which object was taken by the person. The model correctly identifies the ”Towel”. The accompanying table shows the layer-by-layer semantic tracing for visual and text tokens.

D QUALITATIVE EXAMPLES

We provide additional qualitative examples to visually illustrate the findings from our circuit-based analysis. These examples showcase the model's process of interpreting video frames to answer specific questions about objects and actions. Each figure includes the input video frames, the posed question, the model's response, and a table showing the semantic evolution of key tokens across different layers, as analyzed through our semantic tracing circuit (Circuit ②). More examples in the anonymous interactive demo website.

Prompt
USER: The input consists of a sequence of key frames from a video. Question: Which object was thrown by the person?
ASSISTANT: The object is the

***Answer*:** Clothes

Response: Green shirt that the person is throwing.

Token/Layer	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5
<s>	Institution	Архив	sierp	sierp	sierp
US	ona	Архив	pert	tap	pert
ER	Session	пута	oded	庄	pel
:	Portail	пута	<s>	<s>	Sav
		пута	Carter	◆	◆
<IMG001>	yter	yter	yter	yter	yter
<IMG002>	yter	yter	yter	yter	yter
<IMG003>	middle	middle	middle	middle	middle
<IMG004>	yter	yter	yter	yter	yter
<IMG005>	yter	yter	yter	yter	yter
<IMG006>	kwiet	kwiet	kwiet	kwiet	kwiet
<IMG007>	onymes	onymes	onymes	onymes	stycz
<IMG008>	dek	dek	dek	dek	dek
<IMG009>	yter	yter	yter	yter	yter
<IMG010>	assen	assen	assen	assen	assen
<IMG011>	wheel	wheel	wheel	wheel	wheel
<IMG012>	assen	assen	assen	emann	emann
<IMG013>	alet	alet	alet	empio	empio
<IMG014>	gresql	gresql	gresql	gresql	gresql
<IMG015>	alet	alet	alet	alet	alet
<IMG016>	azon	azon	azon	azon	azon
<IMG017>	fle	fle	fle	fle	wi

Figure 10: Qualitative example where the model is prompted to identify a thrown object. The model successfully responds that a "Green shirt" was thrown, correctly identifying both the object and its color. The table illustrates the semantic trace, showing how the model processes the visual information through its layers.

E ETHICS STATEMENT

This research adheres to the ICLR Code of Ethics and aims to enhance the understanding of the internal spatiotemporal reasoning mechanisms within LVLMs through circuit-based analysis, in order to drive the development of more robust and interpretable models. The experiments are based entirely on public academic datasets (e.g., the STAR benchmark), and we acknowledge that these contain videos of human activities, which were used solely for their intended academic analytical purposes. To ensure research integrity, we explicitly state that LLMs were used only for language polishing after the manuscript was written, and that all core scientific contributions originate from the human authors.

F REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of this research. All experiments are based on publicly available models (the `LLaVA-NeXT` and `LLaVA-OV` series) and datasets (the STAR benchmark). We detail our methods for filtering and processing data from the STAR benchmark in the "Dataset Curation" part of Section 3, and provide visualization samples of keyframes in Appendix A. The core methodology of our research, including the specific settings, intervention methods, and evaluation metrics for the three analytical circuits (Visual Information Auditing, Semantic Tracing, and Attention Flow), is thoroughly elaborated in Section 3 (Subsections 3.1, 3.2, and 3.3), which includes key mathematical formulas and parameter definitions. The theoretical framework supporting our experimental design is fully formalized in Appendix B. Furthermore, an anonymous interactive demo website is provided in Appendix D for reviewers to explore additional qualitative results. We believe these detailed descriptions are sufficient to support the reproduction of this work. All code will be made available upon acceptance of the manuscript.

Prompt
USER: The input consists of a sequence of key frames from a video. Question: Which object was picked up by the person?
ASSISTANT: The object is the

***Answer*:** Box

Response: Box of shoes

Token/Layer	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5
<s>	Institution	Архив	sierp	sierp	sierp
US	ona	пута	pert	tap	pure
ER	Session	пута	Session	庄	Ori
:	Portail	пута	<s>	<s>	Sav
	ctl	пута	Carter	◆	IR
<IMG001>	мп	мп	мп	мп	aks
<IMG002>	amen	amen	amen	alberga	amen
<IMG003>	ката	ката	ката	ката	ката
<IMG004>	jes	jes	jes	jes	jes
<IMG005>	anno	мп	мп	мп	anno
<IMG006>	隆	隆	隆	Sito	隆
<IMG007>	gresql	gresql	gresql	gresql	gresql
<IMG008>	ummy	ummy	ummy	ummy	ummy
<IMG009>	osz	osz	osz	Sito	anim
<IMG010>	opf	opf	opf	opf	opf
<IMG011>	pur	pur	pur	pur	pur
<IMG012>	ode	ode	ode	ode	ode
<IMG013>	wheel	wheel	wheel	ava	置
<IMG014>	arab	arab	ode	ode	arab
<IMG015>	gresql	gresql	gresql	end	ł
<IMG016>	кал	кал	CURLOPT	CURLOPT	cub
<IMG017>	way	way	way	way	way

Figure 11: Qualitative example demonstrating the model's ability to recognize an object being picked up. The model correctly identifies the object as a "Box of shoes". The semantic tracing table displays the evolution of token representations across five layers that contribute to this accurate identification.