EVII: Measuring Early Visual Integration in VLM Reasoning

Hakan Muluk[1] and Ozgur S. Oguz[1]

***Abstract*— In robotics, vision-language models are increasingly used to connect visual perception with language-guided reasoning and decision making. However, in multimodal chain-of-thought reasoning, it remains unclear when visual evidence is incorporated and how this relates to final answer reliability. We introduce *Early Visual Information Integration* (EVII), which measures the divergence between image-conditioned and no-image next-token distributions over an early reasoning prefix. EVII shows a stronger relationship with correctness than several widely used confidence-based baselines. We further show that high- and low-EVII examples differ meaningfully in accuracy, that much of the useful visual information is incorporated early, and that a bounded early-prefix variant using at most the first 40 generated tokens remains informative and can be used for inference-time decisions such as routing, fallback, or sample selection.**

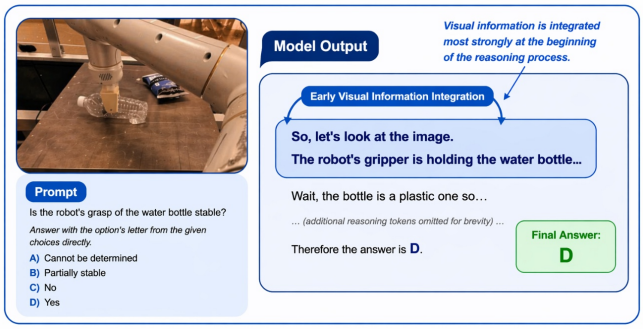

Fig. 1: Illustration of Early Visual Information Integration (EVII). Early reasoning tokens are where visual information is incorporated most strongly into the model's computation. In this example, the early tokens ground the response in the observed scene, influencing the final answer.

I. INTRODUCTION

Vision-language models (VLMs) are becoming increasingly important in robotics, where successful behavior depends on combining perception, language, and embodied state. In this setting, chain-of-thought (CoT) reasoning is useful because it can support tasks such as spatial reasoning, action understanding, and high-level planning [1], [2], [3]. However, prior work suggests that multimodal reasoning can drift toward language priors as decoding continues [4], [5], [6], raising a key question: when is visual information actually incorporated into the reasoning process, and can this help explain whether the model will be correct?

We address this question with *Early Visual Information Integration* (EVII), a visually grounded reliability metric based on the divergence between image-conditioned and no-image next-token distributions over the first few reasoning tokens. Unlike prior contrastive decoding methods that use weakened visual inputs to alter generation, EVII uses this contrast as a diagnostic signal: it measures whether the model's early reasoning trajectory is sensitive to the visual evidence. Across three MC-VQA benchmarks spanning robotics and spatial reasoning, EVII shows stronger overall correlation with correctness than average log-probability, self-certainty, negative entropy, and negative perplexity. High- and low-EVII subsets also show meaningful accuracy differences, and a bounded early-prefix version remains useful for inference-time decisions such as routing, fallback, or sample selection.

[1]Department of Computer Engineering, Bilkent University.
Corresponding author: Hakan Muluk, Department of Computer Engineering, Bilkent University, 06800 Bilkent, Ankara, Turkey. **Email:** hakan.muluk@bilkent.edu.tr
Code is available at: https://github.com/hakanmuluk/EVII

II. RELATED WORK

Recent work suggests that multimodal chain-of-thought (MCoT) is increasingly important for robotics, where successful behavior depends on reasoning jointly over perception, language, and embodied state. As noted in the recent MCoT survey [1], the value of intermediate reasoning lies in remaining connected to perceptual input, and embodied systems such as EmbodiedGPT [2] and ECoT [3] show that grounded reasoning can improve high-level planning, spatial understanding, and robotic control. At the same time, prior work shows that longer multimodal reasoning can drift toward language priors and amplify hallucination [4], [5], [6], while early token distributions already contain useful signals about response quality [7]. These findings directly motivate our focus on the first few reasoning tokens.

Related inference-time methods compare image-conditioned generation against weakened-visual variants to improve grounding or reduce hallucination [8], [9], [10], [11], [12]. Our work builds on the same intuition, but uses this contrast as a diagnostic signal rather than a decoding intervention. In parallel, recent work has proposed distribution-level metrics such as self-certainty [13] as proxies for reasoning quality, alongside common baselines such as average log-probability, negative perplexity, and negative entropy. We compare EVII against these alternatives to study whether early visual integration provides a strong signal of answer correctness.

III. METHODOLOGY

A. Preliminaries

Let x denote the textual prompt and v the visual input. Given a vision-language model (VLM), we consider autoregressive generation of a token sequence $y = (y_1, y_2, \ldots, y_T)$,

where each token y_t is generated conditioned on the prompt and the previously generated tokens $y_{<t} = (y_1, \ldots, y_{t-1})$.

At decoding step t, the VLM defines a next-token distribution over the vocabulary $\mathcal{V}$. When the image is provided, we denote this distribution by

$$p_t^{\text{img}}(\cdot) = p_\theta(\cdot \mid x, v, y_{<t}), \tag{1}$$

and when the model is run without the image, we denote it by

$$p_t^{\text{noimg}}(\cdot) = p_\theta(\cdot \mid x, y_{<t}), \tag{2}$$

where θ denotes the model parameters.

Throughout the paper, p_t^{img} and p_t^{noimg} are evaluated under the same textual context $x, y_{<t}$, so that the only difference is whether visual conditioning is present. In particular, p_t^{noimg} is computed by re-prefilling the prompt and generated prefix without the image, ensuring that the hidden states associated with earlier tokens likewise contain no visual information.

B. Early Visual Information Integration (EVII)

We define *Early Visual Information Integration (EVII)* as a token-level diagnostic that measures how strongly visual information affects the model's early reasoning. For each decoding step t, we compare the image-conditioned and no-image next-token distributions, p_t^{img} and p_t^{noimg}, using the Jensen–Shannon divergence:

$$\text{JSD}(p_t^{\text{img}} \| p_t^{\text{noimg}}) = \frac{1}{2}\text{KL}(p_t^{\text{img}} \| m_t) + \frac{1}{2}\text{KL}(p_t^{\text{noimg}} \| m_t), \tag{3}$$

where

$$m_t = \frac{1}{2}\left(p_t^{\text{img}} + p_t^{\text{noimg}}\right). \tag{4}$$

Intuitively, a larger divergence indicates that the model's next-token behavior at step t is more strongly influenced by the presence of the image, while a smaller divergence suggests greater reliance on language-only information. Given a horizon k, we define EVII as the average divergence over the first k generated tokens:

$$\text{EVII}(k) = \frac{1}{k}\sum_{t=1}^{k} \text{JSD}\left(p_t^{\text{img}} \,\|\, p_t^{\text{noimg}}\right). \tag{5}$$

Thus, EVII measures how strongly visual information shapes the model's early reasoning under the image-conditioned generation trajectory, where the generated tokens are determined by the image-conditioned run and the corresponding no-image distributions are evaluated on the same prefix. In our setting, we use EVII as a diagnostic signal to study whether stronger early visual integration is associated with higher answer correctness.

C. Dynamic Selection of k

Rather than fixing k globally, we select it adaptively for each example from the aligned no-image divergence signal. After aligning the no-image JS values to the generated token positions, we apply a light smoothing step to reduce single-token noise and then run an approximate Bayesian online changepoint detection (BOCPD) procedure [14] with a Beta-family predictive model. Since the signal is bounded in $(0, 1)$, we model each observation using a Beta distribution whose mean is estimated with prior shrinkage,

$$m_r = \frac{\kappa\mu_0 + \sum_{i \in r} x_i}{\kappa + n_r}, \tag{6}$$

and then use

$$x_t \mid r \sim \text{Beta}(m_r\phi, (1 - m_r)\phi). \tag{7}$$

where r denotes a candidate run length, n_r and $\sum_{i \in r} x_i$ are the corresponding running count and sum, μ_0 is a prior mean, κ controls the degree of shrinkage, and ϕ is a fixed concentration parameter. This is not a full Bayesian treatment of the segment parameters; instead, we use a plug-in approximation based on these simple sufficient statistics.

For an example with total generated length T, we define $k_{\min} = \lfloor 0.05T \rfloor$, $\text{ERL} = \lfloor 0.1T \rfloor$, and $k_{\max} = \min(\lfloor 0.15T \rfloor, K_{\text{cap}})$, where $K_{\text{cap}} > 0$ is a fixed cap. The expected run length ERL determines the BOCPD hazard rate $H = 1/\text{ERL}$. BOCPD maintains a posterior over run lengths and updates the changepoint probability online; in particular, the probability of a changepoint at time t is proportional to

$$P(r_t = 0 \mid x_{1:t}) \propto \sum_r P(r_{t-1} = r \mid x_{1:t-1})\, p(x_t \mid r)\, H. \tag{8}$$

We then search only within this bounded early window. If $k_{\min} > k_{\max}$, we directly set $k = k_{\max}$. Otherwise, we choose the earliest $t \in [k_{\min}, k_{\max}]$ such that: 1) the changepoint probability is sufficiently high, and 2) the signal exhibits a persistent downward transition, meaning that the mean over a short post-window is substantially smaller than the mean over a short pre-window. If no such point is found, we fall back to the position with the maximum changepoint probability within the search window.

In our experiments, we set $K_{\text{cap}} = 40$, so the adaptive search is always restricted to the first 40 generated tokens. This procedure is intended to identify the point at which the early no-image divergence begins to decrease in a stable way, yielding an example-specific estimate of how long strong visual influence persists. The selected k is then used when computing EVII for that example.

IV. Experiments and Results

A. Benchmarks and Models

We evaluate our method on three MC-VQA benchmarks: a 1,000-example subset of Robo2VLM [15], a 1,000-example subset of Spatial-MM [16], and the 400-example ERQA benchmark [17], covering robot manipulation, spatial reasoning, and embodied reasoning in robotic settings. Experiments are conducted with Qwen/Qwen3-VL-30B-A3B-Thinking and Qwen/Qwen3-VL-8B-Thinking [18]. Lastly, all responses are generated with deterministic decoding.

B. Correlation with Answer Correctness

Table I reports the correlation between answer correctness and several scoring signals across the three benchmarks and two model sizes. EVII achieves the strongest average

TABLE I: **Weighted binned Pearson correlation between each evaluation metric and answer correctness.** Metric scores are first normalized to the $[0, 1]$ range and grouped into bins of width 0.01. For each bin, we compute the average accuracy, then report the Pearson correlation between bin-level metric values and accuracies, weighted by the number of examples in each bin. Higher is better. Best in each row is shown in bold.

Dataset	Log Probability	Self Certainty	Negative Entropy	Negative Perplexity	EVII (Ours)
Robo2VLM (8B)	0.6938	0.7789	0.7039	0.7182	**0.9040**
Spatial-MM (8B)	0.6923	0.4869	0.6654	0.6556	**0.7673**
ERQA (8B)	0.1129	0.2222	0.1778	0.1852	**0.5463**
Robo2VLM (30B)	0.5269	0.6971	0.5032	0.4699	**0.7370**
Spatial-MM (30B)	0.7391	0.6629	**0.7444**	0.7286	0.6246
ERQA (30B)	0.4930	0.3673	0.4627	0.4156	**0.5179**
Average	0.5430	0.5359	0.5429	0.5288	**0.6829**

TABLE II: **Selective accuracy for examples ranked by EVII.** For each benchmark–model setting, we report the overall accuracy together with the accuracy on the lowest-scoring 10%, lowest-scoring 30%, highest-scoring 10%, and highest-scoring 30% of examples according to EVII. Lower is better for the lowest-ranked subsets, and higher is better for the highest-ranked subsets.

Version	Overall Acc.	Lowest 10% EVII	Lowest 30% EVII	Highest 10% EVII	Highest 30% EVII
Robo2VLM (8B)	55.64%	32.67%	39.87%	81.19%	75.42%
Spatial-MM (8B)	67.70%	52.00%	57.00%	88.00%	79.00%
ERQA (8B)	44.50%	27.50%	35.00%	57.50%	52.50%
Robo2VLM (30B)	65.93%	56.44%	52.49%	84.16%	79.73%
Spatial-MM (30B)	69.30%	62.00%	63.00%	81.00%	77.67%
ERQA (30B)	44.75%	45.00%	35.83%	60.00%	50.83%

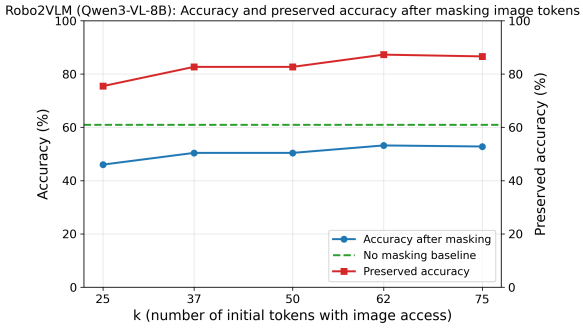

Fig. 2: Accuracy after masking image tokens following the first k generated tokens on Robo2VLM (Qwen3-VL-8B). The blue curve shows the final answer accuracy after masking further attention to image tokens at step k, while the red curve shows the preserved accuracy relative to the no-masking baseline. The green dashed line marks the baseline accuracy without masking. Results are reported on the first quarter of our Robo2VLM subset (250 examples). As k increases, a large fraction of the baseline accuracy is preserved, suggesting that much of the visually relevant information has already been integrated during the early part of the reasoning process.

correlation overall, with an average value of 0.6829, exceeding average log-probability, self-certainty, negative entropy, and negative perplexity. The pattern is especially clear for the 8B model, where EVII is the best-performing signal on all three datasets. For the 30B model, EVII remains competitive and is the strongest signal on Robo2VLM and ERQA, while Spatial-MM shows slightly higher correlation for entropy- and probability-based baselines. Overall, these results support our claim that stronger early visual integration is closely linked to answer correctness.

C. Selective Accuracy via EVII

We further examine EVII by ranking examples according to their EVII scores and measuring accuracy on both the highest- and lowest-scoring subsets. If EVII captures meaningful visual grounding, then higher-EVII examples should tend to be more likely to be correct, while lower-EVII examples should tend to be less likely to be correct. Table II largely follows this pattern across the benchmark–model settings, with one small exception on 30B ERQA for the lowest 10% subset. Overall, these results provide further evidence that EVII is a useful visually grounded reliability signal.

D. Masking the Image After the First k Tokens

To further examine the intuition behind EVII, we perform a masking-after-k experiment. The model is allowed to attend to the image for the first k generated tokens, after which image-token attention is masked and decoding continues without further direct visual access. We do not re-prefill the prompt or generated prefix, so the hidden states may still retain visual information acquired before masking. Due to computational constraints, we conduct this analysis on the 8B model using the first quarter of our Robo2VLM subset.

Figure 2 shows that, although masking the image reduces performance relative to the no-masking baseline, much of the original accuracy is preserved when the model has visual access only during the early stage of generation. For example, when only the first 37 generated tokens can attend to the image, accuracy remains at 50.4%, or about 83% of the no-masking baseline. This is notable because the responses are much longer: at $k = 37$, the mean response length is 668.95 tokens, and all outputs are longer than 37 tokens. Overall, these results are consistent with the intuition behind EVII, suggesting that much of the useful visual information may be incorporated early in the reasoning process, after which later reasoning might continue using the information already integrated into the model's state.

E. Ablation on the Choice of k

We next study how the choice of horizon k affects EVII. In addition to our per-example adaptive selection via BOCPD, we evaluate fixed-k and fixed-ratio alternatives. Table III shows that the BOCPD-based per-example adaptive version performs best overall, clearly outperforming all simpler baselines. Although the strongest fixed alternative is $k/T = 0.05$, it still remains well below the per-example adaptive version.

These results suggest that the most informative early prefix varies across examples, so a single fixed horizon is not sufficient. Per-example adaptive selection via BOCPD is therefore better able to identify the part of the reasoning trajectory most closely linked to correctness.

F. Inference-Time Reliability Estimation from Early Reasoning

Our main BOCPD-based analysis uses response-length-dependent bounds, which is suitable for retrospective analysis but not directly for inference time, since the final response length is unknown beforehand. To obtain an inference-time variant, we instead fix the BOCPD parameters to $k_{\min} = 20$, $\mathrm{ERL} = 30$, and $k_{\max} = 40$. We choose $k_{\max} = 40$ to match the same upper bound used in our original BOCPD analysis. Importantly, although these bounds are now fixed,

TABLE III: **Ablation on the choice of k for EVII.** We compare fixed-k, fixed-ratio, and BOCPD-based adaptive selection using weighted binned Pearson correlation with answer correctness. For each setting, EVII scores are first normalized to the $[0, 1]$ range and grouped into bins of width 0.01. We then compute the average accuracy within each bin and report the Pearson correlation between bin-level EVII values and accuracies, weighted by the number of examples in each bin. Higher is better. Best in each row is shown in bold.

Dataset	k=10	k=30	k=50	k=70	$0.05T$	$0.10T$	$0.15T$	$0.20T$	BOCPD
Robo2VLM (8B)	0.5969	0.6716	0.3207	-0.0202	0.6707	0.4927	0.5015	0.4822	**0.9040**
Spatial-MM (8B)	0.0210	0.1776	-0.2966	-0.5159	0.5854	0.4281	0.4276	0.3319	**0.7673**
ERQA (8B)	0.1111	0.0421	0.0927	0.0246	0.2477	0.2883	0.3374	0.3969	**0.5463**
Robo2VLM (30B)	0.0367	0.4114	0.1673	-0.0480	0.5123	0.1857	0.1673	0.1653	**0.7370**
Spatial-MM (30B)	-0.6037	-0.3539	-0.5234	-0.6125	0.4328	-0.2788	-0.3868	-0.2316	**0.6246**
ERQA (30B)	0.0819	0.2588	0.1808	0.1187	0.1159	0.1695	0.2768	0.3313	**0.5179**
Average	0.0407	0.2013	-0.0098	-0.1756	0.4275	0.2143	0.2207	0.2460	**0.6829**

TABLE IV: **Inference-time reliability estimation using a bounded early-token budget.** We report the average accuracy change, relative to overall accuracy, across the six benchmark–model settings. For the top-ranked subsets, higher is better; for the bottom-ranked subsets, lower is better. EVII uses a BOCPD-based adaptive horizon with fixed constants $k_{\min} = 20$, $\mathrm{ERL} = 30$, and $k_{\max} = 40$, while the baseline metrics are computed on the first 40 generated tokens.

Metric	Bottom 10%	Bottom 30%	Top 10%	Top 30%
EVII	**-6.00pp**	**-4.50pp**	**+4.11pp**	**+5.83pp**
Log Probability	-1.86pp	-0.13pp	+2.01pp	+0.52pp
Negative Entropy	-1.61pp	-0.65pp	-0.54pp	+1.86pp
Negative Perplexity	-2.94pp	-1.43pp	+2.45pp	+2.19pp
Self Certainty	-1.34pp	-0.65pp	-0.78pp	-0.00pp

the horizon itself is still selected adaptively by BOCPD within this bounded early-token window. Thus, EVII retains its adaptive behavior while operating under an inference-time budget, whereas the baseline metrics are computed from the first 40 generated tokens.

Table IV shows that EVII provides the strongest separation between high- and low-accuracy subsets in this bounded inference-time setting. This bounded setup is also practical: one can generate only the required early tokens, stopping once k is determined, and then re-prefill the model with the prompt and generated prefix without the image to compute the EVII score. Since this second pass does not require autoregressive generation, it can be done in parallel and adds relatively little time overhead. These results suggest that EVII may be useful not only for analysis but also as an early-computable inference-time control signal, enabling decisions such as routing to a larger VLM, requesting another camera view, regenerating an answer, or selecting which partial generations to continue in a majority-voting setting.

V. Conclusion

We introduced *Early Visual Information Integration* (EVII), a visually grounded reliability metric that measures how strongly visual evidence is incorporated into the early reasoning process of a vision-language model. Across robotics-oriented and spatial MC-VQA benchmarks, EVII showed a stronger relationship with correctness than several widely used confidence-style baselines, while its high- and low-scoring subsets also reflected meaningful differences in accuracy. Our masking-after-k analysis further suggested that much of the useful image information is incorporated near the beginning of the chain of thought and can then continue to influence later reasoning even without further direct visual access. For robotics, where effective behavior depends on tightly coupling perception and reasoning, this makes EVII useful not only for analyzing how visual grounding relates to correctness, but also as a potential inference-time signal. For example, it could be used to detect when a robot should pause for re-perception before grasping or navigation, or when additional inference-time compute should be reserved for more difficult cases that appear beyond the capabilities of the current model. Because it is derived from a bounded early prefix of the reasoning trajectory, EVII can be computed early enough to support such decisions before the full CoT is completed.

References

[1] Y. Wang, S. Wu, Y. Zhang, S. Yan, Z. Liu, J. Luo, and H. Fei, "Multimodal chain-of-thought reasoning: A comprehensive survey," *arXiv preprint arXiv:2503.12605*, 2025.

[2] Y. Mu, Q. Zhang, M. Hu, W. Wang, M. Ding, J. Jin, B. Wang, J. Dai, Y. Qiao, and P. Luo, "EmbodiedGPT: Vision–language pre-training via embodied chain of thought," *arXiv preprint arXiv:2305.15021*, 2023.

[3] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine, "Robotic control via embodied chain-of-thought reasoning," *arXiv preprint arXiv:2407.08693*, 2024.

[4] C. Liu, Z. Xu, Q. Wei, J. Wu, J. Zou, X. E. Wang, Y. Zhou, and S. Liu, "More thinking, less seeing? Assessing amplified hallucination in multimodal reasoning models," *arXiv preprint arXiv:2505.21523*, 2025.

[5] A. Favero, L. Zancato, M. Trager, S. Choudhary, P. Perera, A. Achille, A. Swaminathan, and S. Soatto, "Multi-modal hallucination control by visual information grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[6] L. Parcalabescu and A. Frank, "Do vision and language decoders use images and text equally? How self-consistent are their explanations?" *arXiv preprint arXiv:2404.18624*, 2024.

[7] Q. Zhao, M. Xu, K. Gupta, A. Asthana, L. Zheng, and S. Gould, "The first to know: How token distributions reveal hidden knowledge in large vision–language models?" in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[8] X. Xiao, B. Wu, J. Wang, C. Li, X. Zhou, and H. Guo, "Seeing the image: Prioritizing visual correlation by contrastive alignment," *arXiv preprint arXiv:2405.17871*, 2024.

[9] L. Zhu, D. Ji, T. Chen, P. Xu, J. Ye, and J. Liu, "IBD: Alleviating hallucinations in large vision–language models via image-biased decoding," *arXiv preprint arXiv:2402.18476*, 2024.

[10] S. Leng, H. Zhang, G. Chen, X. Li, S. Lu, C. Miao, and L. Bing, "Mitigating object hallucinations in large vision–language models through visual contrastive decoding," *arXiv preprint arXiv:2311.16922*, 2023.

[11] D. Wan, J. Cho, E. Stengel-Eskin, and M. Bansal, "Contrastive region guidance: Improving grounding in vision–language models without training," *arXiv preprint arXiv:2403.02325*, 2024.

[12] S. Liu, X. Wen, Z. Lan, A. Wang, and J. Su, "Countering the over-reliance trap: Mitigating object hallucination for LVLMs via a self-validation framework," *arXiv preprint arXiv:2601.22451*, 2026.

[13] Z. Kang, X. Zhao, and D. Song, "Scalable best-of-N selection for large language models via self-certainty," *arXiv preprint arXiv:2502.18581*, 2025.

[14] R. P. Adams and D. J. C. MacKay, "Bayesian online changepoint detection," *arXiv preprint arXiv:0710.3742*, 2007.

[15] K. Chen, S. Xie, Z. Ma, P. R. Sanketi, and K. Goldberg, "Robo2VLM: Visual question answering from large-scale in-the-wild robot manipulation datasets," *arXiv preprint arXiv:2505.15517*, 2025.

[16] F. Shiri, X.-Y. Guo, M. Golestan Far, X. Yu, G. Haffari, and Y.-F. Li, "An empirical analysis on spatial reasoning capabilities of large multimodal models," *arXiv preprint arXiv:2411.06048*, 2024.

[17] Gemini Robotics Team, "Gemini Robotics: Bringing AI into the physical world," technical report, 2025.

[18] S. Bai *et al.*, "Qwen3-VL technical report," *arXiv preprint arXiv:2511.21631*, 2025.