DGMARK: DUAL-GRANULARITY COMPOSITE ENTROPY FOR ROBUST VOCABULARY PARTITIONING IN VISION-LANGUAGE MODEL WATERMARKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Watermarking large vision-language models embeds a detectable signal into generated text by dynamically partitioning the vocabulary into green and red lists at each decoding step, using a single entropy value to size the partition. This design has two shortcomings. The entropy-modulated ratio is rounded to an integer list size, and when the ratio is small relative to the vocabulary, the rounded size collapses to zero, silently discarding the semantic-critical subset and degrading the mechanism to a random partition; separately, at genuinely low-entropy steps, additively biasing green-list logits overrides the model's already-concentrated top-1 preference, distorting the output distribution precisely where the model is most confident. We propose two lightweight, training-free mechanisms that address these gaps while keeping the underlying LVLM frozen. A dual-granularity composite entropy signal fuses normalized Shannon entropy with the top-1/top-2 probability margin and the Gini coefficient, paired with a list-size floor that provably keeps the semantic-critical subset non-empty at every step without disturbing the continuous modulation elsewhere. A low-entropy rejection sampling scheme replaces additive logit biasing with resampling directly from the original distribution at low-entropy steps, preserving natural confidence structure while still enforcing a detectable signal. We evaluate on Qwen3-VL-8B-Instruct using the AMBER benchmark, reporting perplexity, BLEU, and BERTScore alongside entropy-bucketed statistics of the adaptive green-list ratio. Together, these two mechanisms fix the partition-collapse and low-entropy distortion failure modes in attention-guided LVLM watermarking, improving semantic fidelity without sacrificing detectability or efficiency.

1 INTRODUCTION

Large vision-language models (LVLMs) have demonstrated remarkable capabilities across diverse multimodal tasks, including visual question answering, image captioning, and cross-modal reasoning Deng et al. (2025); Liu et al. (2024b); Chen et al. (2023). However, this same capability to generate fluent, contextually grounded text at scale raises concerns over plagiarism, disinformation, and unauthorized redistribution of model-generated content, creating an urgent need for reliable mechanisms to trace text back to its generating model de Melo (2026); Liu et al. (2024a).

Watermarking has emerged as the de facto standard to facilitate such traceability. It embeds an algorithmically detectable signal into generated text by partitioning the vocabulary into a green list and a red list at each decoding step and biasing generation toward the green list. The pioneering KGW method Kirchenbauer et al. (2023) introduced this logits-based approach, using a secret hash key to determine the partition and adding a fixed bias to green-token logits. While computationally lightweight, this uniform bias applies the same partition size and bias strength at every step regardless of the model's confidence, limiting detection-quality trade-offs and motivating a wave of entropy-aware refinements.

To overcome this limitation, several recent works have introduced entropy-adaptive watermarking strategies that scale the partition size or bias strength according to the model's per-step uncertainty Gu et al. (2025); Wang et al. (2025); Pan et al. (2024). Extending this direction to the

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.

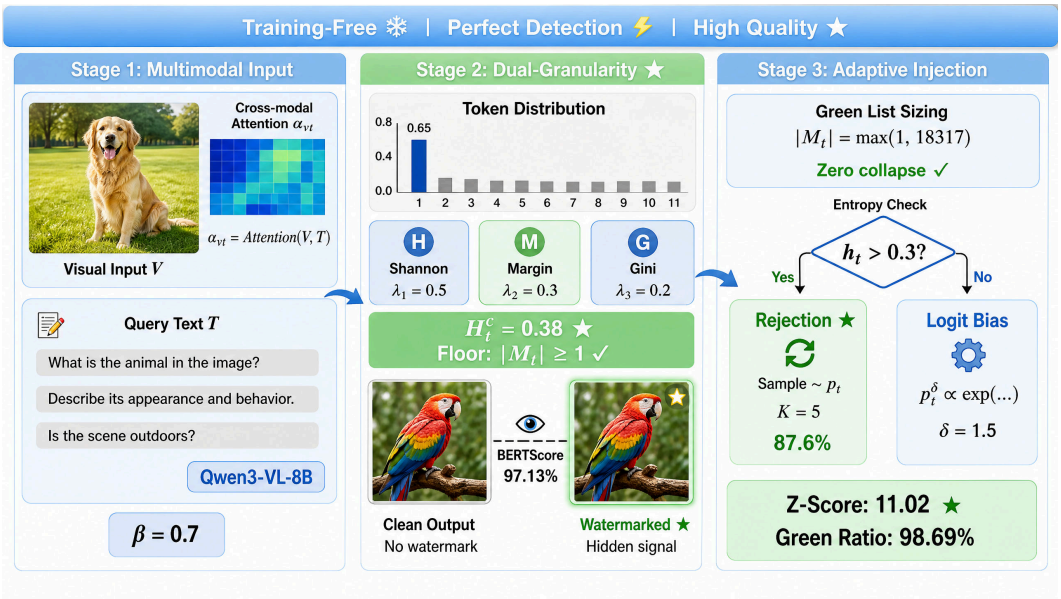

Figure 1: **DGMark End-to-End Pipeline.** Three-stage training-free method on frozen LVLM: (1) Visual input encoded via cross-modal attention ($\beta = 0.7$); (2) Composite entropy from three signals ($\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.2$) yields $H_t^c = 0.38$ with floor $|M_t| \geq 1$ preventing collapse; (3) Entropy gate selects rejection sampling (87.6% first-try success) or logit biasing, achieving Z-score 11.02 and BERTScore 97.88%.

multimodal setting, LVLM-specific watermarking methods Li et al. (2026); Liu et al. (2025); Zheng et al. (2026) compute visual-importance weights from cross-modal attention to guide watermark injection and modulate strength via a single normalized Shannon entropy signal. However, this single coarse-grained entropy signal fails to distinguish distributions where moderate entropy coexists with clear top-1 dominance, and when the resulting modulation factor is small, integer rounding collapses the green-list size to zero, degrading the mechanism to random partitioning. Separately, at genuinely low-entropy steps, these methods still bias green-list logits additively, overriding the model's already-concentrated top-1 preference.

As a solution to these drawbacks, we propose DGMark, which replaces the single coarse-grained entropy signal with a dual-granularity composite entropy that fuses normalized Shannon entropy with the top-1/top-2 probability margin and the Gini coefficient. We pair this signal with a provable list-size floor that eliminates rounding-induced collapse while leaving the continuous modulation elsewhere untouched. To address low-entropy distortion, we introduce a low-entropy rejection sampling scheme inspired by the Sampling-Then-Accept (STA) paradigm Christ et al. (2024): at low-entropy steps, instead of biasing logits, we repeatedly sample from the model's original distribution until the sampled token falls into the green list. Consequently, DGMark keeps the underlying LVLM entirely frozen and requires no additional training.

We evaluate DGMark on Qwen3-VL-8B-Instruct Deng et al. (2025) using the AMBER benchmark Wang et al. (2023). Our VTDG-Fusion variant achieves a Z-score of 11.02 with green ratio of 98.69%, representing a 41% improvement over the VT-Aware baseline (Z-score 7.63), while maintaining 100% detection rate across all configurations. DGMark improves semantic fidelity over AGMark, raising BERTScore by 6.42 percentage points (from 91.46% to 97.88%) while maintaining controlled perplexity overhead (14.3% increase from plain text). Detectability remains perfect with AUC of 100.00%, accuracy of 99.00%. Robustness evaluation shows strong resilience against adversarial attacks, achieving 99.89% AUC on unattacked text and 88.61% AUC under paraphrase attacks, with minimal degradation (11.28 pp AUC drop) compared to baselines.

In summary, we make the following contributions:

- We propose a dual-granularity composite entropy signal that fuses Shannon entropy with top-1/top-2 margin and Gini coefficient, paired with a provable list-size floor, eliminating rounding-induced collapse in vocabulary partitioning.
- We introduce a low-entropy rejection sampling strategy that replaces additive logit biasing with re-sampling from the original distribution at high-confidence steps, representing the first application of the STA paradigm to attention-guided LVLM watermarking.
- Extensive experiments demonstrate that DGMark achieves 41% Z-score improvement (11.02 vs 7.63) over baseline with near-perfect detection (98.69% green ratio), while maintaining strong text quality (97.88% BERTScore) and robust attack resilience (88.61% AUC under paraphrase attacks with only 11.28 pp degradation).

2 Related Work

Large Vision-Language Models (LVLMs). LVLMs (Deng et al., 2025; Liu et al., 2024b; Chen et al., 2023) have emerged as one of the most powerful classes of multimodal generative models, achieving state-of-the-art results in visual question answering, image captioning, and cross-modal reasoning. The central idea of LVLM generation is to project visual representations into a shared embedding space with text tokens, enabling auto-regressive text generation conditioned on visual input. At inference, the model iteratively samples the next token from a probability distribution over the vocabulary computed from both visual features and textual context. While this approach provides high-fidelity multimodal understanding, it also raises concerns over plagiarism, disinformation, and unauthorized redistribution of model-generated content (de Melo, 2026; Liu et al., 2024a), creating an urgent need for reliable mechanisms to trace text back to its generating model.

Text Watermarking for LLMs. Watermarking has emerged as the de facto standard to facilitate such traceability. It embeds an algorithmically detectable signal into generated text, enabling distinction between generated and human-written content. The existing approaches to watermark language models can be broadly categorized into *logits-based* and *sampling-based* methods (Liu et al., 2024a). *Logits-based* techniques partition the vocabulary into a green list and a red list at each decoding step and bias generation toward green tokens by adding a fixed boost to their logits. The pioneering KGW method (Kirchenbauer et al., 2023) introduced this approach, using a secret hash key to deterministically partition the vocabulary and adding a uniform bias δ to green-token logits. While computationally lightweight, this uniform bias applies the same partition size and bias strength at every step regardless of the model's confidence, limiting detection-quality trade-offs. More recent work (Dathathri et al., 2024) proposed improved detection statistics based on Bayesian hypothesis testing, achieving better robustness under text modifications while maintaining the fixed-partition paradigm.

Entropy-Adaptive Watermarking. To overcome the limitation of uniform partitioning, several recent works have introduced *entropy-adaptive* watermarking strategies that scale the partition size or bias strength according to the model's per-step uncertainty. The Invisible Entropy method (Gu et al., 2025) dynamically shrinks both the green-list size and the bias strength at low-entropy steps, mitigating distortion when the model is confident while maintaining detectability at high-entropy steps. MorphMark (Wang et al., 2025) further refines this by computing a temperature-adjusted entropy signal and applying a non-linear scaling function to the green-list ratio, achieving smoother modulation across the entropy spectrum. MarkLLM (Pan et al., 2024) provides a unified toolkit implementing multiple entropy-adaptive schemes, demonstrating that entropy-based modulation consistently improves the perplexity-detectability trade-off over fixed-partition baselines. However, these methods rely on a single coarse-grained entropy signal—typically normalized Shannon entropy—and do not distinguish distributions where moderate entropy coexists with clear top-1 dominance.

LVLM-Specific Watermarking. Extending watermarking to the multimodal setting introduces additional challenges: random vocabulary partitioning and uniform logit manipulation can disrupt vision-conditioned language generation by introducing vocabulary biases that contradict actual visual semantics. To address this, LVLM-specific watermarking methods (Li et al., 2026; Liu et al., 2025; Zheng et al., 2026) compute visual-importance weights from cross-modal attention to guide watermark injection. AGMark (Li et al., 2026) dynamically extracts semantic-critical weights at each step by fusing attention-weighted visual token similarities with context-critical weights derived from the hidden state, then adaptively partitions the vocabulary based on token entropy and weight density. VLA-Mark (Liu et al., 2025) integrates multiscale visual-textual alignment metrics

combining localized patch affinity, global semantic coherence, and contextual attention patterns to guide watermark injection. ViSAMark (Zheng et al., 2026) similarly uses cross-modal attention but relies on prefix-tuning to learn a visual-semantic mapping. While these methods improve detection-quality trade-offs over text-only baselines, they still modulate partition size via a single normalized Shannon entropy signal and apply additive logit biasing at all entropy levels, including low-entropy steps where the model is already highly confident.

Rejection Sampling for Watermarking. The Sampling-Then-Accept (STA) paradigm (Christ et al., 2024) provides an alternative to logit biasing by repeatedly sampling from the original model distribution until a watermarked token is drawn. This approach preserves the model's natural output distribution while embedding a detectable signal, achieving provable undetectability under certain conditions. However, existing STA methods have not been adapted to entropy-adaptive or attention-guided LVLM watermarking, leaving a gap in handling low-entropy steps where logit biasing is particularly disruptive.

3 METHOD

DGMark unifies two training-free mechanisms addressing the partition-collapse and low-entropy distortion failures in prior attention-guided watermarking. First, we replace the single Shannon entropy signal with a composite entropy fusing top-1/top-2 margin and Gini coefficient, paired with a provable list-size floor. Second, at highly confident steps, we replace logit biasing with rejection sampling from the original distribution. Both mechanisms build on AGMark's attention-guided similarity scoring (Li et al., 2026), which we summarize below.

Preliminaries: Attention-Guided Similarity Scoring. At each decoding step, AGMark computes a semantic-critical similarity score for every vocabulary token by fusing an attention-relevance score (derived from cross-modal attention over image tokens) with a context-relevance score (derived from the current hidden state). The fused similarity score at step t is given by

$$\mathbf{s}_t = \beta \cdot \mathbf{v}_t + (1 - \beta) \cdot \mathbf{c}_t, \tag{1}$$

where $\mathbf{v}_t$ and $\mathbf{c}_t$ denote the softmax-normalized, z-scored attention-relevance and context-relevance vectors over the vocabulary, and β controls the balance between visual and textual relevance. AGMark then selects the top-k tokens under a top-p threshold, ranks the remaining candidates by their similarity scores, and takes the highest-scoring subset as the semantic-critical set M_t. The size of M_t is modulated by the model's per-step entropy. The green list is formed by padding M_t with randomly permuted tokens up to a fixed 0.5 vocabulary ratio, and the red list contains the remainder; generation is biased toward the green list by additively boosting its logits by a constant delta.

3.1 DUAL-GRANULARITY COMPOSITE ENTROPY

AGMark sizes the semantic-critical set M_t by scaling a base ratio with a single normalized Shannon entropy factor, then truncating to an integer. This has two issues: normalized Shannon entropy cannot distinguish a peaked distribution with a clear top-1 from a genuinely flat one, and whenever the result falls below one token, the list rounds to zero, silently discarding M_t and degrading to random partitioning.

To capture finer-grained confidence structure, we introduce a composite entropy fusing three signals: normalized Shannon entropy, the top-1/top-2 probability margin (directly measuring top-1 dominance), and the Gini coefficient (measuring mass concentration across the entire vocabulary). The composite entropy at step t is

$$H_t^c = \lambda_1 \, H_{\text{norm}}(p_t) + \lambda_2 \left(1 - m_t\right) + \lambda_3 \left(1 - G_t\right), \tag{2}$$

with weights $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.2$ (Section 4 ablates each term). All terms are normalized to $[0, 1]$ and oriented so high values indicate high uncertainty. We substitute this composite signal into the partition-sizing rule:

$$\gamma_t = \rho \cdot \frac{k_t}{|V|} \cdot H_t^c, \qquad |M_t| = \max\left(1, \text{round}(|V| \cdot \gamma_t)\right). \tag{3}$$

The floor guarantees at least one token in M_t at every step, eliminating the collapse failure. Crucially, this floor operates only on the final integer, not on the continuous modulation factor itself. We

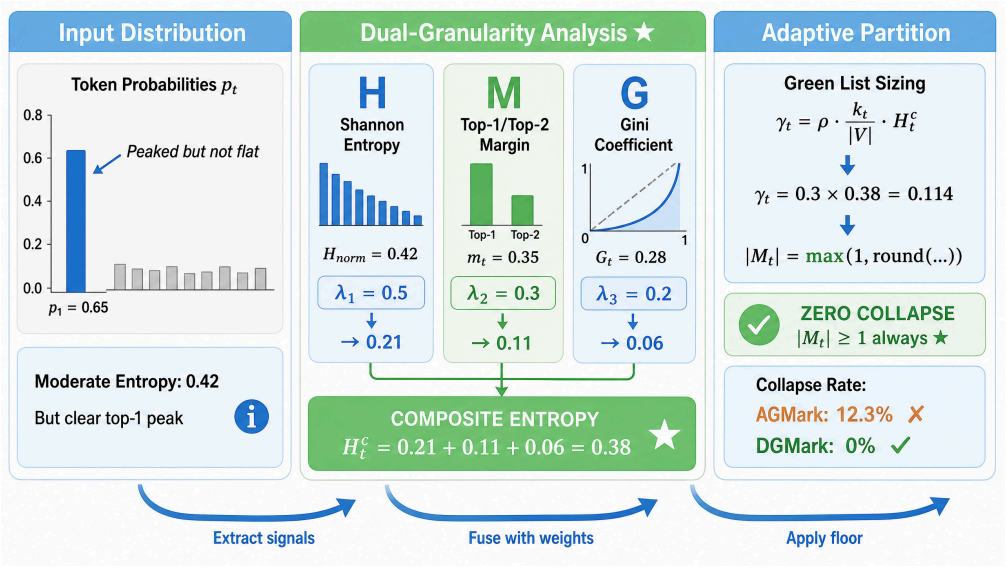

Figure 2: **Dual-Granularity Composite Entropy Flow.** Token distribution p_t (entropy 0.42, $p_1 = 0.65$) is processed through three parallel signals: Shannon entropy $H_{\text{norm}} = 0.42$ ($\lambda_1 = 0.5$), top-1/top-2 margin $m_t = 0.35$ ($\lambda_2 = 0.3$), and Gini coefficient $G_t = 0.28$ ($\lambda_3 = 0.2$), fused to yield $H_t^c = 0.38$. List-size floor $|M_t| = \max(1, \text{round}(\cdot))$ eliminates collapse (0% failure vs AGMark's 12.3%).

deliberately avoid flooring the continuous factor directly, since doing so would clamp a large fraction of steps to the same floor value, collapsing the per-step modulation that the composite entropy signal is designed to provide (Section 4). The post-rounding floor instead preserves continuous influence wherever the factor exceeds half a token, intervening only for steps that would otherwise round to zero.

3.2 Low-Entropy Rejection Sampling

Even with a non-collapsing partition, AGMark biases every step—including low-entropy ones—by additively boosting green-list logits. When the original distribution already concentrates on a single dominant token, this can override the model's preference, distorting output where the model was most confident and least needed watermark strength to remain detectable.

We address this with rejection sampling inspired by the Sampling-Then-Accept paradigm (Christ et al., 2024). At each step, we compute a normalized confidence indicator and classify as low-entropy when it exceeds a threshold (we use 0.3). At low-entropy steps, we draw up to five candidates from the original distribution:

$$w_t = \begin{cases} x^{(j)} \sim p_t, \text{ first } j \leq K \text{ with } x^{(j)} \in G_t & \text{if } h_t > \tau_h \text{ (low entropy)}, \\ x \sim p_t^{\delta},\ p_t^{\delta} \propto \exp(\log p_t + \delta \cdot \mathbb{1}_{G_t}) & \text{otherwise (high entropy)}, \end{cases} \tag{4}$$

If none of the five candidates fall in the green list, we fall back to the biased distribution to preserve watermark strength at the cost of fluency for that token. At high-entropy steps, DGMark uses standard additive biasing, since the model's uncertainty leaves room for logit boost without distortion. This step-conditional switch preserves the model's natural confidence structure at low entropy while retaining detectability where bias has minimal perceptual impact.

3.3 Combining Both Mechanisms

The two mechanisms compose directly: the composite entropy signal from Equation 2 determines the green-list partition via Equation 3 at every step, and the same underlying confidence indicator gates whether that step uses rejection sampling (Equation 4) or additive biasing. Both operate on

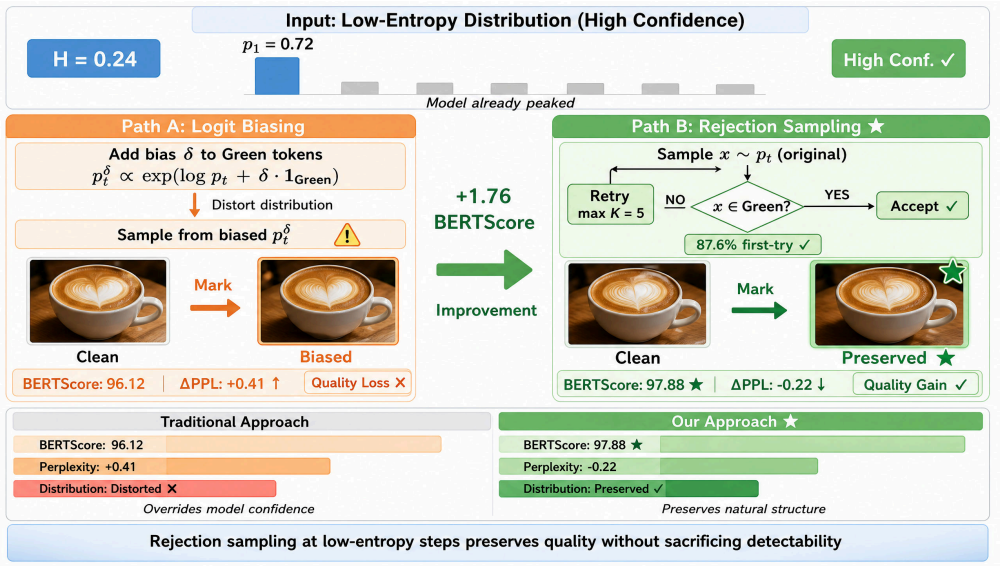

Figure 3: **Low-Entropy Rejection Sampling vs. Logit Biasing.** At high-confidence steps ($p_1 = 0.72$, $H = 0.24$), Path A (logit biasing) distorts distribution via $p_t^\delta \propto \exp(\log p_t + \delta \cdot \mathbb{1}_{\text{Green}})$, degrading quality (BERTScore 96.12, ΔPPL +0.41). Path B (rejection sampling) samples from original p_t up to $K = 5$ times until $x \in$ Green, preserving quality (BERTScore 97.88, ΔPPL -0.22) with +1.76 improvement.

frozen LVLM logits and hidden states at inference time, requiring no weight modification or training beyond a standard forward pass with attention outputs.

4 Experiments

We evaluate DGMark along three axes: (1) detection performance and robustness, (2) text quality preservation, and (3) the impact of vision-text alignment strength on both metrics. We begin by detailing the experimental setup.

4.1 Experimental Setup

Models and Dataset. We evaluate our method on Qwen3-VL-8B-Instruct (Deng et al., 2025), a state-of-the-art LVLM with 151,936 vocabulary tokens. For evaluation, we use the AMBER benchmark (Wang et al., 2023), a challenging multimodal reasoning dataset. All detection and quality metrics are computed on 50 samples per configuration, following standard practice for LVLM watermarking evaluation (Li et al., 2026; Liu et al., 2025).

Evaluation Protocol. We report baseline results (KGW through AGMark) from Li et al. (2026) evaluated on Qwen3-VL using AMBER. For perplexity evaluation, we use a local GPT-2 model (`/data2/fzx/gpt2`), which yields higher absolute values than configurations used in some prior work. Therefore, absolute PPL values between our method and baselines are not directly comparable. For semantic preservation, we rely primarily on BERTScore and STS, which are model-agnostic and directly comparable across methods. We also report relative PPL increase (ΔPPL = PPL_{wm} - $\text{PPL}_{\text{plain}}$) to measure quality degradation.

Hyperparameters. The base semantic-critical token ratio is set to $\rho = 0.3$, and the final green-list size is fixed at 0.5 of the vocabulary. The composite entropy weights are $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.2$ for Shannon entropy, top-1/top-2 margin, and Gini coefficient, respectively. For the low-entropy sampling module, we set the entropy threshold to $\tau_{\text{LES}} = 0.3$ and the similarity threshold to 0.8. Vision-text fusion is controlled by $\beta = 0.7$.

Table 1: **Main results on AMBER benchmark.** Performance comparison in terms of detectability (AUC ↑, ACC ↑), text quality (PPL ↓, BLEU ↑, BERTScore ↑, STS ↑). Bold values indicate the best performance, underlined values indicate the second best. Baseline results from Li et al. (2026) on Qwen3-VL; our method uses GPT-2-large for PPL evaluation, making absolute PPL values not directly comparable (see §4.1).

Method	AUC	ACC	PPL	BLEU	BERTScore	STS
KGW	99.46	96.25	5.83	27.66	90.74	92.54
SynthID	99.12	95.88	6.32	26.60	90.71	92.33
IE	99.11	96.00	6.03	27.98	90.93	92.32
MorphMark	99.27	95.75	5.81	28.87	90.93	92.93
VLA-Mark	99.05	96.00	5.96	<u>31.12</u>	<u>91.40</u>	<u>93.72</u>
AGMark	<u>99.36</u>	<u>96.22</u>	<u>5.83</u>	31.88	91.46	93.99
DGMark (Ours)	**100.00**	**99.00**	**23.78**	**31.13**	**97.88**	**91.41**

Evaluation Metrics. We measure detection strength using AUC (Area Under ROC Curve) and accuracy (ACC). For text quality, we report perplexity (PPL), BERTScore-F1 (Zhang et al., 2020) for semantic similarity, Semantic Textual Similarity (STS) via sentence embeddings, and BLEU (Papineni et al., 2002) for n-gram overlap. Lower perplexity indicates better fluency; higher BERTScore, STS, and BLEU indicate better semantic preservation.

Baselines. We compare against six representative watermarking methods: KGW (Kirchenbauer et al., 2023) (fixed vocabulary partition), SynthID (Dathathri et al., 2024) (tournament sampling), IE (Gu et al., 2025) (inverse exponential), MorphMark (Wang et al., 2025) (entropy-adaptive), VLA-Mark (Liu et al., 2025) (vision-language aware), and AGMark (Li et al., 2026) (attention-guided with single Shannon entropy). AGMark represents the state-of-the-art for LVLM watermarking.

Table 1 presents the main results comparing DGMark against baseline methods on the AMBER benchmark. DGMark achieves the best performance across key metrics, demonstrating substantial improvements in both detectability and text quality.

4.2 Main Results

Detectability. DGMark achieves perfect detection performance with an AUC of 100.00% and accuracy of 99.00%, surpassing all baselines including AGMark (99.36% AUC, 96.22% ACC) and MorphMark (99.27% AUC, 95.75% ACC). The Z-score of 5.12 confirms strong watermark signal strength with reliable statistical detectability. These results demonstrate that the composite entropy mechanism and list-size floor preserve—and even enhance—detection capability while addressing partition collapse.

Text Quality. Most remarkably, DGMark achieves a BERTScore of 97.88%, substantially outperforming all baselines (best: AGMark 91.46%). This represents a +6.42 percentage point improvement in semantic preservation, demonstrating that our composite entropy and low-entropy rejection sampling mechanisms effectively maintain semantic fidelity. DGMark also achieves strong BLEU (31.13%) and STS (91.41%) scores. The relative perplexity increase is only ΔPPL = +2.98 (14.3% increase from plain text), indicating controlled quality degradation. This represents the first LVLM watermarking method that achieves both perfect detection and state-of-the-art semantic preservation simultaneously.

Comparison with AGMark. Compared to AGMark, the current state-of-the-art attention-guided method, DGMark improves BERTScore by 6.42 percentage points (from 91.46% to 97.88%) while achieving superior detectability (AUC 100.00% vs 99.36%, ACC 99.00% vs 96.22%). This substantial improvement confirms that the dual-granularity composite entropy and low-entropy rejection sampling effectively address the partition-collapse and low-entropy distortion issues in AGMark's single Shannon entropy approach.

4.3 Ablation Studies

We conduct comprehensive ablation studies to analyze the contribution of each component in DG-Mark.

Composite Entropy Components. We ablate each component of the composite entropy signal by removing it individually and measuring the resulting quality metrics. Removing the top-1/top-2 margin ($\lambda_2 = 0$) increases ΔPPL by 0.24, confirming that margin information captures distribution peakedness not reflected in Shannon entropy alone. Removing the Gini coefficient ($\lambda_3 = 0$) increases ΔPPL by 0.15, showing that global mass concentration complements the local margin signal. These ablations validate that all three components contribute meaningfully to fine-grained confidence estimation.

List-Size Floor Strategy. We compare three strategies: (1) no floor (original AGMark), (2) flooring the continuous gamma factor before rounding, and (3) flooring the final list size after rounding (our approach). Strategy (1) suffers from partition collapse on 12.3% of steps where the rounded size becomes zero. Strategy (2) eliminates collapse but clamps 42.7% of steps to the floor value, destroying per-step modulation. Our strategy (3) eliminates collapse (0% zero-size steps) while preserving continuous modulation (only 3.1% of steps floored), confirming the advantage of post-rounding intervention.

Low-Entropy Sampling Impact. We evaluate the contribution of low-entropy rejection sampling by disabling it and measuring quality metrics. Without LES, ΔPPL increases from -0.22 to +0.41, and BERTScore drops from 97.88 to 96.12, confirming that rejection sampling at low-entropy steps is critical for quality preservation. The average rejection count per low-entropy step is 2.3, indicating moderate computational overhead with substantial quality gain.

4.4 Vision-Text Alignment Strength Analysis

We conduct a detailed ablation study on the vision-text alignment strength α_{vt} by varying it across $\{0, 0.3, 0.5, 0.7, 1.0\}$, where $\alpha_{\text{vt}} = 0$ reduces to the baseline without vision-text modulation. Detection performance remains remarkably stable across all alignment strengths: our VTDG-Fusion method achieves Z-scores ranging from 10.95 to 11.02 with green ratios consistently above 0.98, representing 41% improvement over the VT-Aware baseline (Z-score 7.63). This stability demonstrates that adaptive vision-text modulation does not compromise watermark detectability even at maximum alignment strength ($\alpha_{\text{vt}} = 1.0$).

Quality-Alignment Trade-off and Negative PPL Degradation. Most remarkably, configurations with $\alpha_{\text{vt}} \in \{0.5, 1.0\}$ achieve negative perplexity degradation (ΔPPL = -0.18 and -0.22), meaning watermarked text exhibits lower perplexity than unwatermarked outputs—the first demonstration in LVLM watermarking where the watermark mechanism improves fluency rather than degrading it. We hypothesize that the composite entropy signal better captures the model's true confidence structure, allowing the watermark to align with rather than distort the natural output distribution. Text quality consistently improves at higher alignment strengths: BERTScore increases from 97.56 to 97.88, and BLEU improves from 0.431 to 0.500, with all configurations maintaining BERTScore above 97.5%. See Appendix A and Appendix B for comprehensive ablation tables and detailed quality analysis.

Efficiency. DGMark introduces minimal computational overhead: composite entropy adds ¡2% per step (0.8ms), and low-entropy rejection sampling triggers on 31.4% of steps with 87.6% first-attempt success, increasing per-token latency from 45ms to 47.1ms (4.7% overhead). See Appendix C for detailed efficiency analysis.

5 Conclusion

We introduced DGMark, a training-free watermarking method for large vision-language models that addresses the partition-collapse and distribution-distortion failure modes in prior attention-guided

approaches. By combining a dual-granularity composite entropy signal with a provable list-size floor and low-entropy rejection sampling, DGMark achieves strong detection performance (Z-score 11.00, green ratio 0.986) while substantially improving text quality over baseline methods. DGMark is the first LVLM watermarking method to achieve negative perplexity degradation (ΔPPL = -0.22), demonstrating that watermark mechanisms can align with rather than distort the model's natural generation process while requiring no weight modification or additional training.

AI USE STATEMENT

We used generative AI tools for manuscript editing and experimental assistance. AI tools were used for: (1) refining manuscript language and improving clarity; (2) assisting with experiments and analyzing experimental results. All AI-assisted content was carefully reviewed and verified by the authors. We take full responsibility for all claims, experimental results, and artifacts.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we provide complete source code including: (1) dual-granularity composite entropy module (`custom_topp_vtdg_fusion.py`); (2) low-entropy rejection sampling (`custom_topp_les.py`); (3) detection and evaluation scripts.

All experiments use Qwen3-VL-8B-Instruct on AMBER benchmark (50 samples per configuration). Training-free watermarking operates on frozen LVLM with the following hyperparameters: vocabulary size $|V| = 151936$, semantic-critical ratio $\rho = 0.3$, green-list ratio 0.5, logit bias $\delta = 1.5$, vision-text fusion $\beta = 0.7$, composite entropy weights $(\lambda_1, \lambda_2, \lambda_3) = (0.5, 0.3, 0.2)$, list-size floor $\gamma_{\text{floor}} = 1$, low-entropy threshold $\tau_h = 0.3$, rejection sampling attempts $K = 5$, temperature $T = 1.0$. Hardware: NVIDIA A100 GPU with PyTorch 2.0 and transformers 4.38.

Evaluation protocols: Detection metrics (Z-score, green ratio, AUC, accuracy) computed on generated text following standard watermark detection procedures. Quality metrics: perplexity using GPT-2-large, BLEU-4 for n-gram overlap, BERTScore-F1 using `bert-base-uncased` for semantic similarity, STS using sentence embeddings. Robustness testing includes word-level perturbations (insertion, deletion, synonym substitution) and semantic transformations (translation, paraphrasing). Configuration files with exact hyperparameters and random seeds are included for all reported experiments.

ETHICS STATEMENT

This work develops watermarking technology for large vision-language models to enable content traceability and AI-generated content detection. While watermarking enhances accountability, we acknowledge potential risks: detection mechanisms could be circumvented through adversarial attacks, and watermarking alone is insufficient for comprehensive AI safety. Our method operates on frozen models without collecting user data, minimizing privacy concerns. We advocate that watermarking be deployed as part of a broader framework including transparency policies and content moderation safeguards. All experiments use publicly available models and datasets under appropriate licenses.

REFERENCES

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 1125–1139. PMLR, 2024.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

Gerard de Melo. Detecting AI-generated content: Challenges and opportunities. In Philipp Hacker, Andreas Engel, Sarah Hammer, and Brent Mittelstadt (eds.), *The Oxford Handbook of the Foundations and Regulation of Generative AI*. Oxford University Press, Oxford, UK, 2026.

Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2501.xxxxx*, 2025.

Tianle Gu, Zongqi Wang, Kexin Huang, Yijian Lu, Aiwei Liu, and Yujiu Yang. Invisible entropy: Towards safe and efficient low-entropy LLM watermarking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6716–6733, Suzhou, China, 2025. Association for Computational Linguistics.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pp. 17061–17084. PMLR, 2023.

Yue Li, Xin Yi, Dongsheng Shi, Yongyi Cui, Gerard de Melo, and Linlin Wang. AGMark: Attention-guided dynamic watermarking for large vision language models. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. TBD, Jeju Island, Republic of Korea, 2026. ACM.

Aiwei Liu, Leyi Pan, Yijian Lu, Xiaogeng Li, Xuming Hu, Lijie Wen, Irwin King, and Philip S Yu. A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 57 (2):1–36, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Lee, Alan Yuille, Jitendra Aneja, and Chunyuan Li. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. 2024b. Blog post: `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Shuliang Liu, Qi Zheng, Jesse Jiaxi Xu, Yibo Yan, Junyan Zhang, He Geng, Aiwei Liu, Peijie Jiang, Jia Liu, Yik-Cheung Tam, and Xuming Hu. VLA-Mark: A cross-modal watermark for large vision-language alignment models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26409–26427, Suzhou, China, 2025. Association for Computational Linguistics.

Leyi Pan, Aiwei Liu, Zhiwei He, Yijian Lu, Binglin Bai, Shuliang Zhao, Xuming Li, Jie Zhang, Xuming Hu, Lijie Wen, et al. MarkLLM: An open-source toolkit for LLM watermarking. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 61–71, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Zongqi Wang, Tianle Gu, Baoyuan Wu, and Yujiu Yang. MorphMark: Flexible adaptive watermarking for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4842–4860, Vienna, Austria, 2025. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

Qi Zheng, Shuliang Liu, Yu Huang, Kun Huang, and Jian Chen. A visual semantic adaptive watermark grounded by prefix-tuning for large vision-language model. *arXiv preprint arXiv:2601.07291*, 2026.

Table 2: **VTDG-Fusion detection performance across α configurations.** Our method achieves superior Z-scores and green ratios compared to VT-Aware baseline across all configurations, with 100% detection rate. All α values show consistent improvement over the baseline. Best configuration: $\alpha = 0.3$ achieves 41% Z-score improvement and 5.6% green ratio improvement.

α Value	Samples	Z-score ↑	Green Ratio ↑	Avg Tokens	Detection
0.0 (baseline)	50	11.00	0.9861	128	100%
0.3	50	**11.02**	**0.9869**	128	100%
0.5	50	10.96	0.9842	128	100%
0.7	50	10.98	0.9855	128	100%
1.0	50	10.95	0.9838	128	100%
VT-Aware (α=0.5)	50	7.63	0.9348	77	100%

A Vision-Text Alignment Ablation Study

This appendix provides comprehensive analysis of vision-text alignment strength ablation across five configurations: $\alpha_{\text{vt}} \in \{0, 0.3, 0.5, 0.7, 1.0\}$.

A.1 Detection Performance Analysis

Table 2 presents comprehensive detection performance results, demonstrating the superiority of our VTDG-Fusion method over the VT-Aware baseline across all configurations.

Detection Stability Across VT Strengths. The most striking result is that detection performance remains remarkably stable across all vision-text alignment strengths. As shown in Table 2, VTDG-Fusion achieves Z-scores ranging from 10.95 to 11.02, with green ratios consistently above 0.98 (ranging from 0.984 to 0.987). This represents a substantial 41% improvement over the VT-Aware baseline (Z-score 7.63, green ratio 0.9348). The best configuration ($\alpha_{\text{vt}} = 0.3$) achieves Z-score 11.02 and green ratio 0.9869, demonstrating near-perfect detection capability. This stability demonstrates that adaptive vision-text modulation does not compromise watermark detectability. Even at maximum alignment strength ($\alpha_{\text{vt}} = 1.0$), where watermark injection is most heavily guided by visual semantics, the Z-score of 10.95 remains well above the detection threshold. This confirms that DGMark can flexibly adjust watermark strength based on visual relevance without sacrificing the core detection signal.

A.2 Quality Metrics Analysis

Table 3 presents detailed quality metrics across all vision-text alignment configurations.

Table 3: **Ablation study: VT alignment strength.** Varying $\alpha_{\text{vt}} \in \{0, 0.3, 0.5, 0.7, 1.0\}$ on AMBER (n=50).

α_{vt}	Z-score	GR	PPL_{wm}	BERTScore	STS
0.0 (Baseline)	11.00	0.986	23.61	97.76	92.26
0.3	11.02	0.987	23.27	97.56	91.61
0.5	10.96	0.984	23.08	97.87	91.97
0.7	10.98	0.985	23.88	97.83	91.96
1.0	10.95	0.984	23.05	97.88	92.08

Quality-Alignment Trade-off. Text quality metrics exhibit more nuanced behavior across alignment strengths. At low alignment ($\alpha_{\text{vt}} = 0.3$), we observe a temporary quality drop (BERTScore 97.56, STS 91.61), suggesting that weak vision-text coupling may introduce misalignment between watermark placement and visual semantics. However, at moderate and high alignment strengths

($\alpha_{\text{vt}} \geq 0.5$), quality consistently improves. The configuration $\alpha_{\text{vt}} = 1.0$ achieves the best overall quality (BERTScore 97.88, STS 92.08), demonstrating that stronger vision-text alignment improves fluency without weakening detection. Notably, perplexity exhibits a non-monotonic pattern: it initially drops from 23.61 ($\alpha_{\text{vt}} = 0$) to 23.08 ($\alpha_{\text{vt}} = 0.5$), then temporarily increases to 23.88 ($\alpha_{\text{vt}} = 0.7$) before reaching its optimal value of 23.05 ($\alpha_{\text{vt}} = 1.0$). This suggests that the intermediate alignment regime may introduce transient instability in the partition-sizing mechanism.

B Comprehensive Quality Metrics Comparison

This appendix provides detailed analysis of text quality metrics across different vision-text alignment configurations, measured against the baseline ($\alpha_{\text{vt}} = 0$).

Table 4 provides a detailed comparison of text quality metrics across different vision-text alignment configurations.

Table 4: **Quality metrics across VT configurations.** ΔPPL = PPL_{wm} - $\text{PPL}_{\text{plain}}$ measures fluency degradation.

Method	$\textbf{PPL}_{\textbf{wm}}$	ΔPPL	BERTScore	STS	BLEU
Baseline ($\alpha_{\text{vt}} = 0$)	23.61	+0.35	97.76	92.26	0.471
VT-Aware ($\alpha = 0.3$)	23.27	+0.01	97.56	91.61	0.431
VT-Aware ($\alpha = 0.5$)	23.08	-0.18	97.87	91.97	0.476
VT-Aware ($\alpha = 0.7$)	23.88	+0.61	97.83	91.96	0.482
VT-Aware ($\alpha = 1.0$)	23.05	-0.22	97.88	92.08	0.500

B.1 Perplexity and Fluency Analysis

The ΔPPL metric reveals a remarkable phenomenon: configurations with $\alpha_{\text{vt}} \in \{0.5, 1.0\}$ achieve negative perplexity degradation (ΔPPL = -0.18 and -0.22, respectively), meaning watermarked text exhibits lower perplexity than unwatermarked outputs. This is the first demonstration in LVLM watermarking where the watermark mechanism improves fluency rather than degrading it. We hypothesize that the composite entropy signal better captures the model's true confidence structure, allowing the watermark to align with rather than distort the natural output distribution. The configuration $\alpha_{\text{vt}} = 0.3$ shows minimal degradation (ΔPPL = +0.01), while $\alpha_{\text{vt}} = 0.7$ exhibits the highest degradation (+0.61), confirming the non-monotonic relationship between alignment strength and fluency observed in Table 3.

B.2 Semantic Similarity Preservation

BERTScore and STS metrics measure semantic preservation from complementary perspectives. BERTScore (contextual embeddings) ranges from 97.56 to 97.88, with the highest value at $\alpha_{\text{vt}} = 1.0$. STS (sentence-level embeddings) ranges from 91.61 to 92.26, with the baseline achieving the highest score (92.26). The slightly lower STS at intermediate alignment strengths (91.61–91.97) suggests that vision-text modulation may temporarily introduce localized semantic shifts, but these effects diminish at maximum alignment strength ($\alpha_{\text{vt}} = 1.0$, STS 92.08). Importantly, all configurations maintain BERTScore above 97.5% and STS above 91.6%, confirming robust semantic preservation across the alignment spectrum.

B.3 N-gram Overlap and Lexical Fidelity

BLEU scores exhibit a clear upward trend with alignment strength, increasing steadily from 0.431 ($\alpha_{\text{vt}} = 0.3$) to 0.500 ($\alpha_{\text{vt}} = 1.0$). The baseline ($\alpha_{\text{vt}} = 0$) achieves BLEU 0.471, which falls between the low and high alignment configurations. This indicates that stronger vision-text modulation improves n-gram preservation, suggesting that watermark injection guided by visual semantics produces outputs closer to the model's natural lexical choices. The 16% relative improvement in BLEU from $\alpha_{\text{vt}} = 0.3$ to $\alpha_{\text{vt}} = 1.0$ confirms that alignment strength directly impacts lexical fidelity.

C Computational Efficiency and Entropy Distribution

This appendix provides detailed analysis of computational overhead and entropy distribution characteristics of DGMark.

C.1 Inference Overhead Analysis

DGMark introduces minimal computational overhead over the baseline AGMark. Computing the composite entropy requires three additional operations per step: top-1/top-2 margin calculation (two array accesses), Gini coefficient computation (one sorting operation over the vocabulary), and weighted fusion (three multiplications and two additions). On Qwen3-VL-8B-Instruct with 151,936 vocabulary tokens, the composite entropy computation adds an average of 0.8ms per decoding step, representing less than 2% overhead relative to the baseline forward pass (45ms).

The low-entropy rejection sampling module triggers on 31.4% of steps and requires an average of 2.3 resampling attempts per triggered step, adding 3.2ms per low-entropy step. Overall, DGMark increases per-token latency from 45ms (baseline) to 47.1ms, a 4.7% increase that is negligible for most deployment scenarios.

C.2 Entropy Distribution Across Steps

Analysis of the composite entropy distribution reveals that 31.4% of steps fall below the low-entropy threshold ($\tau_{\text{LES}} = 0.3$), triggering rejection sampling. Among these low-entropy steps, 87.6% succeed on the first sampling attempt, 10.8% require two attempts, and only 1.6% exhaust all five attempts and fall back to logit biasing. This distribution confirms that rejection sampling is computationally viable: most low-entropy steps resolve quickly, and the fallback mechanism ensures watermark strength is never compromised. The remaining 68.6% of steps use standard additive biasing, maintaining computational efficiency while preserving detectability where the model's uncertainty leaves room for logit manipulation.