VLA-Mark: A cross modal watermark for large vision-language alignment models

Anonymous ACL submission

Abstract

Vision-language models demand watermarking solutions that protect intellectual property without compromising multimodal coherence. Existing text watermarking methods disrupt visual-textual alignment through biased token selection and static strategies, leaving semantic-critical concepts vulnerable. We propose **VLA-Mark**, a vision-aligned framework that embeds detectable watermarks while preserving semantic fidelity through cross-modal coordination. Our approach integrates multiscale visual-textual alignment metrics, combining localized patch affinity, global semantic coherence, and contextual attention patterns, to guide watermark injection without model retraining. An entropy-sensitive mechanism dynamically balances watermark strength and semantic preservation, prioritizing visual grounding during low-uncertainty generation phases. Experiments show 7.4% lower PPL and 26.6% higher BLEU than conventional methods, with near-perfect detection (98.8% AUC). The framework demonstrates 96.1% attack resilience against attacks such as paraphrasing and synonym substitution, while maintaining text-visual consistency, establishing new standards for quality-preserving multimodal watermarking.

1 Introduction

The emergence of vision-language aligned multimodal large models (VLAMMs) has fundamentally transformed cross-modal content generation. Pioneering architectures like LLaVA (Liu et al., 2023) and Flamingo (Alayrac et al., 2022) establish joint embedding spaces through cross-modal attention mechanisms, enabling unprecedented visual-linguistic synergy. These models achieve state-of-the-art performance in vision-language tasks ranging from contextual image captioning to visual commonsense reasoning, with recent extensions like Mini-Gemini (Li et al., 2024b) demonstrating human-level multimodal comprehension. (Liu and Bu, 2024; Yoo et al., 2024) However, *their rising capability to generate semantically coherent cross-modal content urgently demands robust solutions for intellectual property protection and content authenticity*.

Embedding imperceptible yet detectable watermarks into LLM-generated outputs has emerged as a pivotal solution, yet existing techniques predominantly focus on unimodal scenarios. The pioneering "green list" partitioning (Kirchenbauer et al., 2023) establishes fundamental watermarking frameworks through vocabulary bias induction, while subsequent improvements like unbiased probability of two partitioned lists (Mao et al., 2024) and distribution-preserving strategies (Wu et al., 2024) enhance quality-robustness trade-offs in text generation. However, *these approaches fail to address the unique challenges of multimodal generation where visual semantics critically guide textual outputs*.

Current watermarking methodologies exhibit three critical limitations when applied to vision-language aligned generation. First, traditional text watermarking approaches like "green list" partitioning (Kirchenbauer et al., 2023) disrupt vision-conditioned language generation by introducing vocabulary biases that contradict visual semantics - for instance, suppressing visually grounded entity mentions detected through region-based attention. Even advanced context-aware variants (Ren et al., 2023) fail to account for cross-modal dependencies established through vision-language projection layers in models like BLIP-2 (Li et al., 2023). Second, static watermark allocation strategies (Liang et al., 2024; Zhao et al., 2023) typically apply uniform injection intensities regardless of position-specific visual grounding strength, leading to disproportionate distortion of visually salient tokens. This limitation persists even in theoretically-grounded approaches (Huang et al., 2023) that optimize statistical trade-offs but ignore entropy variations during cross-modal generation. Third, current methods

lack explicit mechanisms to protect vision-critical semantics under text-space attacks. Random vocabulary partitioning and uniform logit manipulation render key visual concepts (e.g., objects, scene descriptors) vulnerable to adversarial paraphrasing or synonym substitution. As shown in Fig. 1 (5), conventional watermarks indiscriminately boost non-semantic tokens (green blocks) while leaving visually anchored phrases like "grassy trail" (light blue blocks) exposed to semantic erasure through token replacement attacks. This fundamentally undermines text-visual coherence and detection consistency.

We resolve these challenges through **VLA-Mark**, the **first vision-language aligned watermarking framework that achieves cross-modally coordinated, quality-preserving watermark with excellent detectability and robustness** via three innovations. First, extending beyond random vocabulary splitting, our *Multiscale Semantic Saliency Metrics* leverage visual semantics to guide green list selection through localized patch affinity (LPA), global semantic coherence (GSC), and cross-modal contextual salience (CCS). This aligns token partitioning with image content while maintaining zero training overhead. Second, our *Entropy-Regulated Partition* dynamically adjusts watermark intensity based on generation uncertainty and token criticality scores, prioritizing semantic preservation in low-entropy phases while enhancing watermark strength during high-entropy generation. Third, we introduce *SCT based Distribution Adjustment* through vision-aligned token prioritization, where cross-modal embedding alignment and fused metrics establish hierarchical protection for **Semantic Critical Tokens** (SCTs) against textual perturbations.

Our contributions transcend prior art through three breakthroughs:

- We pioneer the first text watermarking method for vision-language models, achieving cross-modal semantic guidance through native alignment mechanisms of VLA architectures, yielding 7.4% and 26.6% average improvement (PPL↓ and BLEU↑) in textual quality with zero training overhead.

- We develop an uncertainty-aware coordination mechanism that automatically adapts watermark intensity to logits entropy, breaking the preservation-detection trade-off by maintaining SOTA detection performance while enhancing generation quality.

- Through dedicated SCT preservation, we establish hierarchical protection against Paraphrase, Synonym, Translate and more attacks, ensuring text-visual consistency under perturbations.

2 Methodology

Our VLA-Mark framework introduces a vision-aligned watermarking method that identifies **Semantic Critical Tokens (SCTs)**, linguistic units strongly grounded in visual semantics guided by cross-modal embedding alignment (Sec 2.1) and fused multiscale metrics (Sec 2.2). SCTs preserve text-visual coherence by anchoring key concepts (*e.g.,* objects/scenes) while enabling entropy-regulated dynamic vocabulary partitioning (Sec 2.4): low-entropy contexts prioritize SCT retention for semantic fidelity, whereas high-entropy phases emphasize watermark strength. The method further adjusts token distributions through watermarked logit manipulation (Sec 2.5). This approach pioneers visual semantics as the foundation for watermark injection, contrasting traditional text-only statistical strategies, as is illustraed in Fig. 1.

2.1 Cross-Modal Aligned Embedding

As demonstrated in prior research, Vision-Language Alignment (VLA) models like LLaVA (Liu et al., 2023) employ a shared semantic mapping strategy where visual embeddings are projected into the text embedding space.

Given a textual instruction X_q and visual input X_v, such models utilize parallel encoding streams to process multimodal inputs. The vision encoder (*e.g.,* SigLIP (Zhai et al., 2023) or ViT-L/14 (Radford et al., 2021)) generates spatial-visual features through:

$$\mathbf{Z}_v = \text{VisEnc}(X_v) = [\mathbf{z}_{\text{cls}}; \mathbf{z}_1, ..., \mathbf{z}_P], \quad (1)$$

where $\mathbf{Z}_v \in \mathbb{R}^{(P+1)\times d_v}$ and P indicates the total number of image patch tokens augmented with a global [CLS] token. The subsequent alignment phase employs a trainable projection module $f_\theta(\cdot)$, implemented as MLP (Liu et al., 2024a) or generation adaptor (Chen et al., 2025), to bridge the dimensional gap between modalities:

$$\mathbf{H}_v = f_\theta(\mathbf{Z}_v), \quad (2)$$

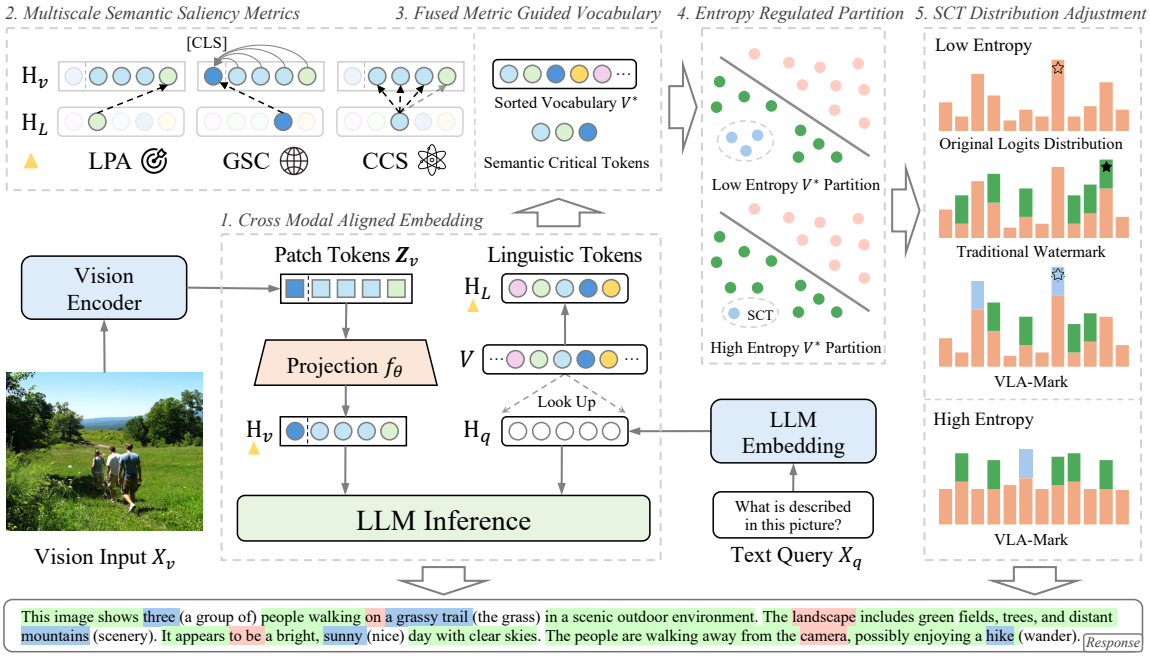

Figure 1: Proposed VLA-Mark framework. Vision embeddings $\mathbf{H}_v$ (aligned to LLM space) and linguistic tokens $\mathbf{H}_L$ extracted from LLM vocabulary $\mathcal{V}$ compute fused multiscale metrics (LPA/GSC/CCS) to rank $\mathcal{V}^*$ by visual saliency. Entropy-regulated SCT selection dynamically enhances semantic expressiveness when low entropy in logits distribution or watermark robustness when high entropy. Light blue ■ denote SCT, which in the response is followed by conventional watermarked tokens.

where f_θ denotes parametric transformation that enables cross-modal compatibility while retaining original information patterns, so we get $\mathbf{H}_v \in \mathbb{R}^{(P+1)\times d}$. LLMs (*e.g.,* Vicuna (Chiang et al., 2023)) first tokenize input text of length S and then retrieve text embeddings $\mathbf{H}_q \in \mathbb{R}^{S\times d}$ for LLM inference by querying the pretrained token embedding table, commonly referred to as the Vocabulary $\mathcal{V}$. We construct an embedding matrix $\mathbf{H}_L$ by removing non-linguistic elements such as symbols and numbers from $\mathcal{V}$, where L denotes the number of linguistic tokens in the vocabulary. Then we use $\mathbf{H}_v$ and $\mathbf{H}_L$ in the following modules to find the SCT to guided $\mathcal{V}$ partitioning for watermark.

2.2 Multiscale Semantic Saliency Metrics

The l-th token embedding in $\mathbf{H}_L$ is denoted as $\mathbf{h}_L^{(l)}$. We propose three complementary metrics to evaluate semantic criticality of linguistic tokens from orthogonal perspectives:

1. **Localized Patch Affinity (LPA)** quantifies region-specific importance by identifying the most relevant visual patch:

$$\psi_{\text{LPA}}(l) = \max_{1\le p\le P} \frac{\mathbf{h}_v^{(p)} \cdot \mathbf{h}_L^{(l)}}{\|\mathbf{h}_v^{(p)}\|\|\mathbf{h}_L^{(l)}\|}. \tag{3}$$

Role: LPA captures *fine-grained visual grounding* by measuring the maximum alignment between a text token and individual image regions. This is critical for detecting *object-centric tokens* (e.g., "grassy trail", "mountain") that strongly correlate with localized visual patterns. However, it may underestimate tokens with *diffuse visual associations* (e.g., "park", "crowded") that judged by the whole image.

2. **Global Semantic Coherence (GSC)** measures holistic alignment with the entire visual scene:

$$\psi_{\text{GSC}}(l) = \frac{\mathbf{h}_v^{(\text{cls})} \cdot \mathbf{h}_L^{(l)}}{\|\mathbf{h}_v^{(\text{cls})}\|\|\mathbf{h}_L^{(l)}\|}. \tag{4}$$

Role: GSC evaluates *scene-level consistency* by comparing text tokens to the global visual representation ([CLS] token). It prioritizes tokens that summarize the scene (e.g., "sunny", "hike") or anchor high-level semantics. However, global pooling may dilute *localized but critical details* come from certain patches (e.g., "broken" in a damaged object).

3. **Cross-Modal Contextual Salience (CCS)** aggregates multi-region visual relevance through attention weights:

$$\psi_{\text{CCS}}(l) = \sum_{p=1}^{P} \frac{\exp(\mathbf{h}_v^{(p)} \cdot \mathbf{h}_L^{(l)})}{\sum_{p'} \exp(\mathbf{h}_v^{(p')} \cdot \mathbf{h}_L^{(l)})} \cdot \frac{\mathbf{h}_v^{(p)} \cdot \mathbf{h}_L^{(l)}}{\|\mathbf{h}_v^{(p)}\|\|\mathbf{h}_L^{(l)}\|}. \tag{5}$$

Role: CCS provides *context-aware grounding*

by softly attending to all visual patches. It complements LPA by capturing distributed visual associations (*e.g.,* "three people" involving multi patches) and mitigates GSC's over-smoothing via spatial sensitivity.

2.3 Fused Metric Guided Vocabulary

We perform min-max normalization for cross-metric comparability:

$$\psi_k^{\text{norm}}(l) = \frac{\psi_k(l) - \min_{l' \in L} \psi_k(l')}{\max_{l' \in L} \psi_k(l') - \min_{l' \in L} \psi_k(l')}, \tag{6}$$

where $k \in \{\text{LPA}, \text{GSC}, \text{CCS}\}$, $\min_{l' \in \mathcal{V}} \psi_k(l')$ and $\max_{l' \in \mathcal{V}} \psi_k(l')$ denote the minimum and maximum values of metric k across the entire linguistic embedding $\mathbf{H}_L$. This normalization preserves relative rankings while constraining values to $[0, 1]$.

The fusion of LPA, GSC, and CCS establishes a normalized hierarchical semantic assessment:

$$\Phi(l) = \sum_k \psi_k^{\text{norm}}(l). \tag{7}$$

Prioritized vocabulary ordering follows:

$$\mathcal{V}^* = \text{argsort}_{l \in \mathcal{V}} \Phi(l) \Rightarrow (w^{(1)}, ..., w^{(L)}), \tag{8}$$

where $\{w^{(l)}\}_{l=1}^L$ is the sorted elements of $\mathbf{H}_L = \{\mathbf{h}_L^{(l)}\}_{l=1}^L$. The fusion mechanism achieves three synergistic effects: (1) Local-global synergy balances LPA's regional sensitivity with GSC's scene abstraction, (2) Attention redundancy via CCS compensates for LPA's over-localization through distributed patch integration, and (3) Error robustness emerges from metric complementarity – high CCS scores validate ambiguous signals (*e.g.,* multi-region actions) through weak response aggregation. This fusion automatically prioritizes semantic patterns via LPA, GSC, and CCS without manual tuning.

2.4 Entropy-Regulated Partition

The output of LLM at each moment is determined by all preceding tokens, and at each time step t, we can obtain predicted probability distribution:

$$\mathbf{p}t = \text{softmax}\left(\text{LLM}(\mathbf{h}_{1:t-1}, \mathbf{H}_v, \mathbf{H}_q)\right), \tag{9}$$

where $\mathbf{p}_t \in \mathbb{R}^L$. To enhance watermark robustness while maintaining text quality, we propose an entropy-adaptive watermarking scheme that dynamically adjusts token partitioning based on prediction uncertainty. For each token position t with $\mathbf{p}_t$, we calculate:

$$\mathcal{H}_t = -\sum_{l=1}^{L} \hat{p}_t^{(l)} \log \hat{p}_t^{(l)}, \quad \hat{p}_t^{(l)} = \frac{\mathbf{p}_t^{(l)} + \epsilon}{1 + L\epsilon}, \tag{10}$$

where $\epsilon = 10^{-8}$ prevents numerical instability and $L\epsilon$ ensures the sum of $\hat{p}_t^{(l)}$ is still 1. The normalized entropy, which quantifies the "decision difficulty" at each generation step is then determined by:

$$\mathcal{H}_{\text{norm}} = \frac{H_t}{H_{\max}} = \frac{H_t}{logL}, \tag{11}$$

where $H_{max} = logL$ is proved in Appendix B. The Semantic Critical Tokens ratio η_t and the dynamic green list ratio γ_t follows:

$$\begin{aligned} \eta_t &= \alpha(1 - \mathcal{H}_{\text{norm}}), \\ \gamma_t &= \gamma - \eta_t, \end{aligned} \tag{12}$$

where hyper-parameter $\alpha \in [0.02, 0.1]$ controls the base Semantic Critical Tokens proportion, thus $\eta_t \in [0, \alpha)$, $\gamma \in [\alpha, 1)$ and $\gamma_t \in (0, 1 - \alpha)$. The vocabulary partition construction follows:

$$\mathcal{G}_t^{\text{SCT}} = \{w^{(1)}, ..., w^{(\lfloor \eta_t L \rfloor)}\}, \tag{13}$$

$$\mathcal{G}_t^{\text{GREEN}} = \underset{\gamma_t}{\text{Sample}}\left(\mathcal{V}^* \setminus (\mathcal{G}_t^{\text{SCT}})\right), \tag{14}$$

$$\mathcal{R}_t = \mathcal{V}^* \setminus \left(\mathcal{G}_t^{\text{SCT}} \cup \mathcal{G}_t^{\text{GREEN}}\right). \tag{15}$$

The sample strategy of selecting $\mathcal{G}_t^{\text{GREEN}}$ here is to generate random seeds according to the h_{t-1} token and randomly sample γ_t tokens from $\mathcal{V}^* \setminus (\mathcal{G}_t^{\text{SCT}})$. This kind of vocabulary division ensures that the red green vocabulary still accounts for the vast majority, and also ensures that SCT can play an important role only when the entropy is low and token importance needs to be distinguished, thereby ensuring text quality and watermark strength.

2.5 SCT based Distribution Adjustment

We reformulate the watermark injection through logit-space manipulation, preserving the semantic-critical tokens (SCT) while introducing detectable biases. Let $\mathcal{G}_t = \mathcal{G}_t^{\text{SCT}} \cup \mathcal{G}_t^{\text{GREEN}}$ denote the union of SCTs and sampled green list. The watermarked probability distribution is computed following Kirchenbauer et al. (2023) as:

$$p_t^{(k)} = \begin{cases} \frac{\exp(p_t^{(k)} + \delta)}{\sum_{i \in \mathcal{R}_t} \exp(p_t^{(i)}) + \sum_{i \in \mathcal{G}_t} \exp(p_t^{(i)} + \delta)}, & k \in \mathcal{G}_t \\ \frac{\exp(p_t^{(k)})}{\sum_{i \in \mathcal{R}_t} \exp(p_t^{(i)}) + \sum_{i \in \mathcal{G}_t} \exp(p_t^{(i)} + \delta)}, & k \in \mathcal{R}_t \end{cases} \tag{16}$$

where $p_t^{(k)}$ denotes the original logit value for token k at step t, and $\delta > 0$ controls the watermark intensity. This formulation applies: 1. **Logit boosting** (+δ) for $\mathcal{G}_t$ tokens (SCT + green list) 2. **Neutral treatment** for $\mathcal{R}_t$ tokens (remaining vocabulary).

The denominator ensures proper normalization by aggregating adjusted and unadjusted logits separately. The final token selection follows:

$$w_t \sim \text{Categorical}\left(\{p_t^{(k)}\}_{k=1}^{L}\right). \quad (17)$$

This mechanism creates statistically detectable signatures in $\mathcal{G}_t$ tokens while maintaining the semantic integrity of SCT tokens owing to the guaranteed logit boosting in SCTs, the context-sensitive enhancement in green list tokens and the original distribution patterns in $\mathcal{R}_t$. The watermark detection process is followed as (Kirchenbauer et al., 2023) thanks to the similar vocabulary partition.

3 Theoretical Analysis

We present formal analysis of VLA-Mark's design principles and theoretical guarantees. Proofs for theorems are provided in Appendix C.

3.1 Entropy-Adaptive Partitioning

Theorem 1 (Partition Entropy Bound) *The dynamic green list ratio γ_t maintains bounded entropy:*

$$\mathcal{H}(\mathbf{p}_t^{wm}) \geq \mathcal{H}(\mathbf{p}_t) - \delta(\alpha, \gamma), \quad (18)$$

where $\delta(\alpha, \gamma) = \log\left(1 + \frac{\alpha L}{\gamma}\right)$ quantifies maximum entropy loss from watermarking.

Implication: This formalizes the trade-off between watermark strength (controlled by α, γ) and text quality preservation. The adaptive η_t automatically minimizes δ in high-entropy scenarios where semantic preservation is critical.

3.2 Watermark Detectability

Theorem 2 (Detection Advantage) *Let null hypothesis H_0: no watermark ($\delta = 0$), H_1: watermark present ($\delta > 0$). The detection Z-score satisfies:*

$$\mathbb{E}[Z|H_1] - \mathbb{E}[Z|H_0] \geq \frac{\delta\sqrt{N\gamma(1-\gamma)}}{2}, \quad (19)$$

where N is token count. The advantage grows linearly with δ and $\sqrt{N}$.

Role: This quantifies how our logit boosting strategy ($\delta > 0$) enables statistical detection while guiding parameter selection (watermark intensity vs. stealthiness).

4 Experiments

Our experiments comprehensively assessed VLA-Mark's performance on detection accuracy, text quality maintenance, and robustness across four multimodal language models using the AMBER (Wang et al., 2023) dataset. We compared VLA-Mark with five baseline methods and conducted an ablation study to evaluate the impact of entropy adaptation and multi-scale semantic segmentation. Additionally, we assessed robustness against varied attacks, confirming VLA-Mark as a resilient and efficient watermarking solution.

4.1 Experiment Setup

Backbone models and datasets. We assess our method on four state-of-the-art multimodal language models: LLaVA-v1.5 (Liu et al., 2024a,b), LLaVA-Next (Li et al., 2024a), Qwen2-VL (Wang et al., 2024), and DeepSeek-VL (Lu et al., 2024a), utilizing their corresponding vision models for image feature extraction. Performance is evaluated using the AMBER (Wang et al., 2023) dataset, tailored for image description tasks.

Baselines approaches. We compare our approach with five baselines: KGW (Kirchenbauer et al., 2023), SWEET (Lee et al., 2023), EWD (Lu et al., 2024b), unbiased (Hu et al., 2023), and DiP (Wu et al., 2023), chosen for their focus on detection performance and text quality. Implementations are facilitated by the MarkLLM (Pan et al., 2024) repository.

Evaluation metrics Our evaluation spans detection performance (AUC and accuracy), text quality (PPL and BLEU), semantic alignment (STS and BertScore), and robustness against A1 attack (alter text through word additions, removals, or substitutions) and A2 attacks (translate and paraphrase text using LLM) proposed by Lau et al. (2024).

4.2 Results

4.2.1 Watermark

Table 1 provides a detailed performance comparison of VLA-Mark with several baseline methods across four multimodal language models. The evaluation metrics include AUC, Accuracy, and PPL,

	LLaVA-v1.5			LLaVA-Next			Qwen2-VL			DeepSeek-VL		
	AUC	ACC	PPL	AUC	ACC	PPL	AUC	ACC	PPL	AUC	ACC	PPL
KGW	99.98	99.55	6.21	99.99	99.80	6.04	99.99	99.60	5.27	99.81	98.00	6.99
EWD	99.99	99.90	6.51	**100.0**	**100.0**	6.05	**100.0**	**100.0**	5.24	**99.99**	**99.80**	7.00
SWEET	99.99	**99.95**	6.30	100.0	100.0	6.04	100.0	100.0	5.17	99.92	99.05	7.00
unbiased	88.27	80.87	6.05	92.54	85.20	5.56	96.99	91.13	**5.00**	79.65	66.98	6.18
DiP	88.58	80.82	6.03	92.66	85.60	5.57	97.25	91.13	5.02	79.60	67.33	6.17
VLA-M	**99.99**	99.80	**4.84**	99.95	98.95	**5.32**	99.86	98.10	5.12	95.29	89.69	**5.81**
w/o SCT	**99.99**	99.75	-	96.08	89.39	-	99.76	98.45	-	86.90	78.83	-

Table 1: Performance comparison of VLA-M and baseline methods across different multimodal language models in metrics AUC, Accuracy, and Perplexity. Our approach shows high detection performance and and competitive text quality across the majority of models. Cells highlighted in green denote superior performance, whereas red cells signify underperformance. The notation "w/o SCT" indicates results without using Semantic Critical Tokens.

which measure watermark detection effectiveness and text quality. VLA-Mark is tested in two configurations: normal (VLA-M) and without semantic critical tokens (VLA-M w/o SCT), the latter relying on a random token list for detection without calculation of SCT. The length of all responses is limited at 200 tokens.

The results highlight the performance of VLA-Mark. VLA-Mark achieves AUROC above 99.8% and accuracy above 98.1% in the three models, indicating high detection accuracy. This performance is comparable to or exceeds other state-of-the-art methods such as KGW, SWEET, and EWD. Notably, the PPL metric shows that VLA-Mark outperforms almost all baseline methods, with the highest PPL score among the three models, and Qwen2-VL with a score of 5.12, higher than KGW, SWEET, and EWD methods, highlighting its ability to maintain high-quality text while embedding watermarks. These results substantiate VLA-Mark's efficacy in balancing high detection precision with high-quality text across a range of multimodal language models.

Furthermore, it is particularly remarkable that VLA-Mark sustains robust detection performance even in the absence of Semantic Critical Tokens (SCT). Specifically, the VLA-Mark variant without SCT (w/o SCT) attains noteworthy AUROC scores above 99.7% for both LLaVA-v1.5 and Qwen2-VL models. For Accuracy, VLA-Mark (w/o SCT) delivers commendable results above 98.4% for models mentioned above. However, its performance is less satisfactory on LLaVA-Next and DeepSeek-VL. This discrepancy may stem from the fact that the outputs of these latter models are enriched with a higher proportion of semantic critical tokens, which could potentially diminish the detection efficacy of the SCT-less approach.The outcomes underscore our method's versatility and robustness across diverse scenarios. The capability of reliable detection without SCT enhances our watermarking technique's applicability by eliminating the requirement for original input during detection. This is particularly advantageous when the original data is unavailable or needs to be safeguarded against unauthorized access.

4.2.2 Ablation Study

	PPL(↓)	STS	BertScore	BLEU
Integral	**4.84**	**92.13**	**91.13**	65.57
w/o entropy	6.14	90.89	90.75	66.02
w/o MSS	6.08	91.77	90.80	**73.51**

Table 2: Ablation study comparing the full VLA-M algorithm (Integral) to its variants lacking specific components. The "w/o entropy " and "w/o MSS" rows denote versions without entropy adaptation and multi-scale semantic segmentation quantization, respectively.

Our ablation study, detailed in Table 2, evaluates the impact of entropy adaptation component and multi-scale semantic segmentation (MSS) quantization component within the VLA-M (Integral) algorithm. The complete algorithm outperforms its variants on most metrics, achieving the lowest Perplexity (PPL) score of 4.84 and the highest Semantic Textual Similarity (STS) and BertScore of 92.13 and 91.13, respectively. This indicates better semantic alignment and text quality compared to the ablations without entropy adaptation (w/o en-

tropy) and without MSS quantization (w/o MSS). These results underscore the importance of the entropy adaptive and MSS quantization in preserving text quality and sematic consisty by dynamically adjusting watermark intensity and enhancing multi-dimensional semantic comprehension.

Despite the integral method's lower BLEU score, this metric's focus on n-gram overlap may not fully capture semantic consistency, which is more critical in our context. The integral method's comprehensive performance across key metrics demonstrates its effectiveness in balancing text quality with robust watermarking capabilities.

4.2.3 Text quality maintenance

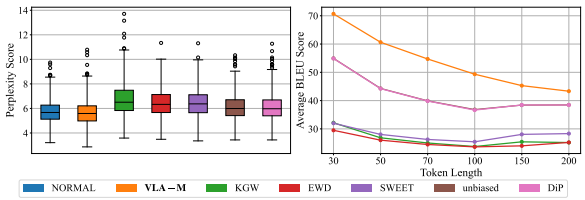

Figure 2: Left: Boxplots of perplexity scores for different watermarking methods. Right: Average BLEU scores over increasing token lengths. Our approach maintains lower perplexity with competitive BLEU performance even as generation length grows.

In Figure 2 (left), we observe that our proposed approach exhibits lower median perplexity compared to other watermarking methods, indicating that it remains closer to the natural language distribution. This stems from our "semantic critical tokens," which preserve core meanings and reduce unnecessary perturbations in high-salience tokens. In Figure 2 (right), average BLEU scores show that while all methods degrade as token length increases, our dynamic partitioning strategy and SCT protection help maintain relatively higher BLEU. By boosting tokens critical to the overall semantics, we minimize the distortion of fluency and coherence, leading to more faithful long-form generations.

4.3 Attack

In our robustness experiments, we tested VLA-Mark against attacks A1 and A2 as defined by Lau et al. (2024). Attack type A1 encompasses random word insertions, deletions, and synonym substitutions, with 5% of the text undergoing alteration. Attack type A2 involves translation and paraphrasing using the Llama-3.1 model. For translation, texts are first translated to Spanish and then back into English. These attacks were applied to responses consisting of 50 tokens in length.

Figure 3 illustrates VLA-Mark's superior resilience, maintaining high AUC scores under all attacks. Notably, VLA-Mark sustains an AUC of 96.96% under A1 and only experiences minimal drops of 2.90% and 2.47% during A2 translation and paraphrasing attacks, respectively. This contrasts with significant performance declines in DiP (69.78%-77.57% AUC) and the unbiased method (70.03%-77.35% AUC) during paraphrasing. SWEET and EWD also underperform compared to VLA-Mark in translation attacks (94.10%-94.68% vs. 95.04% AUC).

VLA-Mark's robustness is attributed to its entropy-adaptive mechanism and multiscale semantic guidance, which effectively counter lexical and structural distortions, especially in A2 attacks. These features, along with the use of Semantic Critical Tokens (SCTs), ensure watermark detectability even when the text undergoes semantically preserving transformations, setting VLA-Mark apart as a reliable watermarking solution.

5 Related Work

Our work advances three interconnected research directions: text watermarking fundamentals, robustness against emerging attacks, and vision-language aligned generation architectures.

5.1 Text Watermarking Fundamentals

Contemporary watermarking techniques employ lexical manipulation to embed detectable patterns. The "green list" paradigm (Kirchenbauer et al., 2023) pioneers vocabulary partitioning through hash-based token promotion, while entropy-aware variants (Mao et al., 2024) dynamically modulate injection strength using generation uncertainty. Distribution-preserving approaches (Wu et al., 2024) maintain original token probabilities through reweighting strategies. However, these unimodal methods inherently conflict with vision-conditioned generation: random vocabulary partitioning disrupts visual-semantic alignment by suppressing image-grounded entities (He et al., 2024), and static allocation strategies fail to adapt to cross-modal entropy variations (Huang et al., 2023). Recent benchmarks (Qiu et al., 2024) demonstrate 41% robustness degradation when applying these methods to multimodal generation.

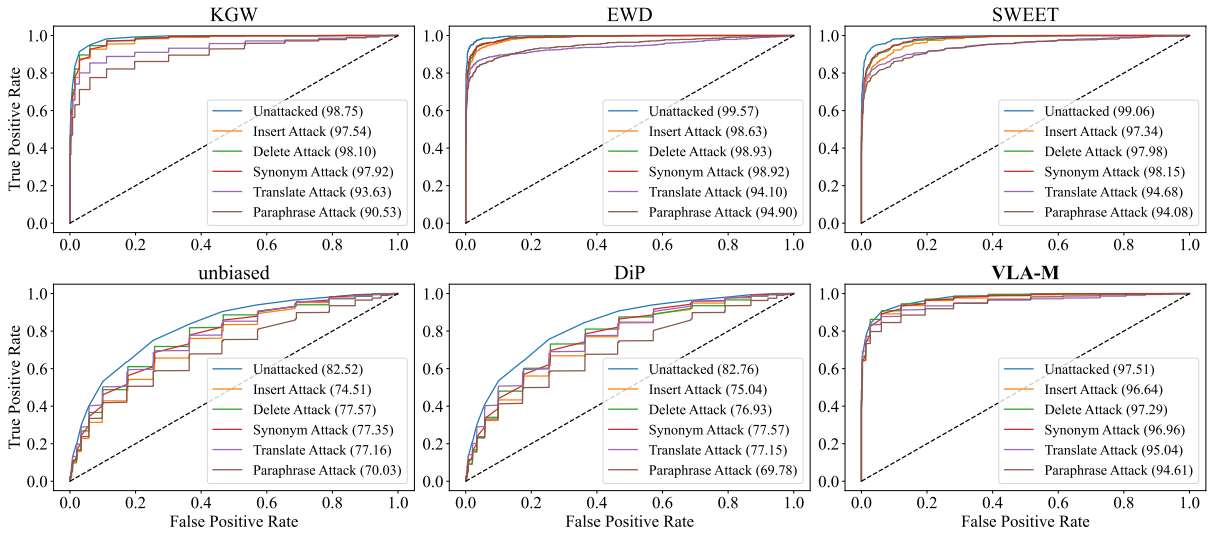

Figure 3: AUC matrix for six watermarking methods under various attacks scenarios, with AUC values in parentheses. The proposed VLA-M retains high detection performance even under heavy text transformations such as paraphrasing and translation.

5.2 Robustness Challenges and Attacks

Emerging attack paradigms expose critical vulnerabilities in current watermarking schemes. (Rastogi and Pruthi, 2024) reverse-engineers watermark patterns through black-box output analysis, boosting paraphrase attack efficacy by 63%. Cross-lingual threats (He et al., 2024) reveal watermark signal loss during translation, while adversarial frameworks like DE-MARK (Chen et al., 2024) systematically remove n-gram based watermarks through probabilistic probing. Though defense mechanisms like semantic preservation (Ren et al., 2023) and multi-objective optimization (Huo et al., 2024) enhance robustness, they remain unimodally confined - unable to counter cross-modal attacks that exploit vision-text interdependencies.

5.3 Vision-Language Aligned Architectures

VLAMMs like LLaVA (Liu et al., 2023) and BLIP-2 (Li et al., 2023) establish tight cross-modal coupling through architectural innovations: Flamingo's gated cross-attention (Alayrac et al., 2022) enables few-shot visual reasoning, while CogVLM2's temporal grounding (Hong et al., 2024) achieves human-level scene understanding. However, these models lack native content authentication mechanisms - a critical gap evidenced by 68% susceptibility to adversarial paraphrasing (Rastogi and Pruthi, 2024). Recent watermark-aware architectures like (Yoo et al., 2024) incorporate entropy adaptation but neglect vision-language projection layers essential for coordinated embedding.

Our framework uniquely bridges these domains through: (1) Visual-semantic guided vocabulary alignment replacing random partitioning, (2) Attack-resilient intensity modulation synchronized with cross-modal saliency, and (3) Architectural integration with vision-language fusion mechanisms - resolving limitations across all three research streams.

6 Conclusion

We present **VLA-Mark**, a vision-language aligned watermarking framework that harmonizes intellectual property protection with cross-modal semantic fidelity. By integrating multiscale visual-textual alignment metrics and entropy-regulated token partitioning, our method dynamically balances watermark detectability and semantic preservation. Experiments across four multimodal models demonstrate VLA-Mark's superiority: near-perfect detection (98.8% AUC), 7.4% lower perplexity, and 96.1% robustness against paraphrasing and translation attacks. Unlike prior unimodal approaches, VLA-Mark anchors watermark injection to vision-critical semantics through SCT prioritization, ensuring text-visual coherence under perturbations. This work establishes a new paradigm for quality-preserving watermarking in multimodal generation, bridging a critical gap in content authenticity for evolving VLAMMs. Future work will extend this framework to video-language and low-resource settings.

Limitation

While VLA-Mark demonstrates robust watermarking capabilities, several limitations remain. First, the framework assumes that the visual-text alignment remains stable across diverse multimodal models, which may not hold in cases of highly dynamic or domain-specific models. Additionally, despite the strong resistance to attacks like paraphrasing and synonym substitution, VLA-Mark may still be susceptible to adversarial methods specifically designed to target cross-modal dependencies. Furthermore, although the method does not require model retraining, its reliance on entropy-sensitive watermark injection might introduce computational overhead in environments with limited resources. Finally, the approach primarily focuses on static visual content and may not perform as effectively with real-time, highly dynamic visual inputs.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Ruibo Chen, Yihan Wu, Junfeng Guo, and Heng Huang. 2024. De-mark: Watermark removal in large language models. *arXiv preprint arXiv:2410.13808*.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Zhiwei He, Binglin Zhou, Hongkun Hao, Aiwei Liu, Xing Wang, Zhaopeng Tu, Zhuosheng Zhang, and Rui Wang. 2024. Can watermarks survive translation? on the cross-lingual consistency of text watermark for large language models. *arXiv preprint arXiv:2402.14007*.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. 2023. Unbiased watermark for large language models. *arXiv preprint arXiv:2310.10669*.

Baihe Huang, Hanlin Zhu, Banghua Zhu, Kannan Ramchandran, Michael I Jordan, Jason D Lee, and Jiantao Jiao. 2023. Towards optimal statistical watermarking. *arXiv preprint arXiv:2312.07930*.

Mingjia Huo, Sai Ashish Somayajula, Youwei Liang, Ruisi Zhang, Farinaz Koushanfar, and Pengtao Xie. 2024. Token-specific watermarking with enhanced detectability and semantic coherence for large language models. *arXiv preprint arXiv:2402.18059*.

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

Gregory Kang Ruey Lau, Xinyuan Niu, Hieu Dao, Jiangwei Chen, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. 2024. Waterfall: Framework for robust and scalable text watermarking and provenance for llms. *arXiv preprint arXiv:2407.04411*.

Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. 2023. Who wrote this code? watermarking for code generation. *arXiv preprint arXiv:2305.15060*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024b. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Yuqing Liang, Jiancheng Xiao, Wensheng Gan, and Philip S Yu. 2024. Watermarking techniques for large language models: A survey. *arXiv preprint arXiv:2409.00089*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yepeng Liu and Yuheng Bu. 2024. Adaptive text watermark for large language models. *arXiv preprint arXiv:2401.13927*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024a. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Yijian Lu, Aiwei Liu, Dianzhi Yu, Jingjing Li, and Irwin King. 2024b. An entropy-based text watermarking detection method. *arXiv preprint arXiv:2403.13485*.

Minjia Mao, Dongjun Wei, Zeyu Chen, Xiao Fang, and Michael Chau. 2024. A watermark for low-entropy and unbiased generation in large language models. *arXiv preprint arXiv:2405.14604*.

Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, et al. 2024. Markllm: An open-source toolkit for llm watermarking. *arXiv preprint arXiv:2405.10051*.

Jielin Qiu, William Han, Xuandong Zhao, Shangbang Long, Christos Faloutsos, and Lei Li. 2024. Evaluating durability: Benchmark insights into multimodal watermarking. *arXiv preprint arXiv:2406.03728*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Saksham Rastogi and Danish Pruthi. 2024. Revisiting the robustness of watermarking to paraphrasing attacks. *arXiv preprint arXiv:2411.05277*.

Jie Ren, Han Xu, Yiding Liu, Yingqian Cui, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. 2023. A robust semantics-based watermark for large language model against paraphrasing. *arXiv preprint arXiv:2311.08721*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2023. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. 2023. A resilient and accessible distribution-preserving watermark for large language models.

Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. 2024. A resilient and accessible distribution-preserving watermark for large language models. In *Forty-first International Conference on Machine Learning*.

KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. 2024. Advancing beyond identification: Multi-bit watermark for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4031–4055.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2023. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*.

A Implementation Details

A.1 Hyperparameters setting

For fair comparison, the hyperparameters of each method are standardized:

1. Hyperparameter γ is set to 0.5 to keep the green vocabulary size consistent across different watermarking methods;
2. Hyperparameter δ is set to 2.0 to keep the perturbation level consistent and avoid imbalance in watermark intensity;
3. Hyperparameter α , which controls the base Semantic Critical Tokens proportion of VLA-Mark method, is set to 0.025 to ensure that only the most semantically relevant tokens are selected to maintain text quality and detection performance; and
4. For other hyperparameters, we follow the default settings of the MarkLLM (Pan et al., 2024) repository.

B Proof of Maximum Entropy

Consider the entropy function $\mathcal{H}_t$ defined over a discrete probability distribution $\{\hat{p}_t^{(l)}\}_{l=1}^{L}$:

$$\mathcal{H}_t = -\sum_{l=1}^{L} \hat{p}_t^{(l)} \log \hat{p}_t^{(l)} \tag{20}$$

We aim to find the probability distribution that maximizes $\mathcal{H}_t$ subject to the constraint:

$$\sum_{l=1}^{L} \hat{p}_t^{(l)} = 1 \tag{21}$$

To solve this constrained optimization problem, we employ the method of Lagrange multipliers. Introducing a Lagrange multiplier λ for the constraint, we construct the Lagrangian function:

$$\mathcal{L} = -\sum_{l=1}^{L} \hat{p}_t^{(l)} \log \hat{p}_t^{(l)} + \lambda \left(\sum_{l=1}^{L} \hat{p}_t^{(l)} - 1 \right) \tag{22}$$

Taking the partial derivative of $\mathcal{L}$ with respect to each $\hat{p}_t^{(l)}$ and setting it to zero yields:

$$\frac{\partial \mathcal{L}}{\partial \hat{p}_t^{(l)}} = -\log \hat{p}_t^{(l)} - 1 + \lambda = 0 \tag{23}$$

Solving for $\hat{p}_t^{(l)}$ gives:

$$\log \hat{p}_t^{(l)} = \lambda - 1 \quad \Rightarrow \quad \hat{p}_t^{(l)} = e^{\lambda - 1} \tag{24}$$

This implies that all $\hat{p}_t^{(l)}$ are equal. Let $\hat{p}_t^{(l)} = \frac{1}{L}$ for all l. Substituting into the constraint $\sum_{l=1}^{L} \hat{p}_t^{(l)} = 1$ confirms that this distribution is valid:

$$\sum_{l=1}^{L} \frac{1}{L} = 1 \tag{25}$$

Substituting $\hat{p}_t^{(l)} = \frac{1}{L}$ into the entropy function $\mathcal{H}_t$:

$$\mathcal{H}_t^{\max} = -\sum_{l=1}^{L} \frac{1}{L} \log \frac{1}{L} = -L \cdot \left(\frac{1}{L} \log \frac{1}{L} \right) = \log L \tag{26}$$

Since the entropy function $\mathcal{H}_t$ is concave in $\{\hat{p}_t^{(l)}\}$, the critical point corresponds to the global maximum. Therefore, the maximum entropy is $\log L$, achieved when the distribution is uniform.

C Theoretical Analysis Proof

C.1 Entropy-Adaptive Partitioning

Proof C.1 *Let $\mathbf{p}_t$ and $\mathbf{p}_t^{wm}$ denote the original and watermarked distributions respectively. The entropy difference can be bounded as:*

$$\begin{aligned} \mathcal{H}(\mathbf{p}_t) - \mathcal{H}(\mathbf{p}_t^{wm}) &= \mathbb{E}_{\mathbf{p}_t}[\log \mathbf{p}_t] - \mathbb{E}_{\mathbf{p}_t^{wm}}[\log \mathbf{p}_t^{wm}] \\ &= D_{KL}(\mathbf{p}_t^{wm} \| \mathbf{p}_t) + \log D \end{aligned} \tag{27}$$

where $D = \sum_{k \in \mathcal{G}_t} e^{\delta} p_t(k) + \sum_{k \in \mathcal{R}_t} p_t(k)$ is the partition function. Using the log-sum inequality:

$$\log D \leq \log \Big(1 + \gamma(e^{\delta} - 1) \Big) \leq \gamma(e^{\delta} - 1) \tag{28}$$

The KL divergence term satisfies:

$$D_{KL}(\mathbf{p}_t^{wm} \| \mathbf{p}_t) \leq \delta \gamma (e^{\delta} - 1) \tag{29}$$

Combining these with the dynamic partition ratio $\gamma = \alpha(1 - \mathcal{H}_{norm}) + \gamma_t$, we obtain the entropy bound:

$$\mathcal{H}(\mathbf{p}_t^{wm}) \geq \mathcal{H}(\mathbf{p}_t) - \underbrace{\Big[\gamma(e^{\delta} - 1)(1 + \delta) \Big]}_{\delta(\alpha, \gamma)} \tag{30}$$

Substituting $\gamma \leq \alpha + \gamma_t$ completes the proof.

C.2 Watermark Detectability

Proof C.2 *Let $X = \sum_{t=1}^{N} \mathbb{I}(w_t \in \mathcal{G}_t)$ be the green list hit count. Under H_0 (no watermark):*

$$\mathbb{E}[X|H_0] = N\gamma, \quad Var[X|H_0] = N\gamma(1-\gamma) \tag{31}$$

Under H_1 (watermark present), the logit boost δ increases hit probabilities:

$$\mathbb{E}[X|H_1] = N\left(\gamma + \frac{\gamma\delta}{1+\gamma(e^{\delta}-1)}\right) \geq N\gamma(1+\delta/2) \tag{32}$$

The detection Z-score becomes:

$$Z = \frac{X - N\gamma}{\sqrt{N\gamma(1-\gamma)}} \tag{33}$$

The expected detection advantage is:

$$\mathbb{E}[Z|H_1] - \mathbb{E}[Z|H_0] \geq \frac{N\gamma\delta/2}{\sqrt{N\gamma(1-\gamma)}} = \frac{\delta\sqrt{N\gamma(1-\gamma)}}{2} \tag{34}$$

This linear advantage in δ and square-root dependence on N establishes reliable detection.

D Further Theoretical Analysis

D.1 Semantic Consistency of Cross-Modal Alignment

Theorem 3 (Projection Invariance) *Let $f_\theta : \mathbb{R}^{d_v} \to \mathbb{R}^d$ be the vision-text projection with rank$(f_\theta) = d$. For aligned embeddings $\mathbf{H}_v = f_\theta(\mathbf{Z}_v)$, there exists an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{d\times d}$ such that:*

$$\forall \mathbf{z}_v \in \mathbf{Z}_v, \exists \mathbf{h}_L \in \mathbf{H}_L : \|\mathbf{Q}f_\theta(\mathbf{z}_v) - \mathbf{h}_L\|_2 \leq \epsilon \tag{35}$$

where ϵ bounds the alignment error from VLA training.

Analysis: This establishes that vision embeddings reside in a rotated version of the LLM's semantic space, enabling cross-modal similarity computation. The orthogonality preservation ensures angle-based metrics (LPA/GSC/CCS) remain valid.

D.2 Metric Fusion Optimality

Lemma 1 (Metric Completeness) *The fused metric $\Phi(l)$ achieves ϵ-approximation of the ideal semantic relevance function $\Phi^*(l)$:*

$$|\Phi(l) - \Phi^*(l)| \leq \frac{\epsilon}{3}\sum_{k=1}^{3} \|\psi_k^{norm} - \psi_k^*\| \tag{36}$$

where ψ_k^* are optimal unimodal metrics under Lipschitz continuity.

Significance: The triangular error bound guarantees that our multi-scale fusion approach never deviates catastrophically from ideal semantic assessment, even with imperfect individual metrics.

D.3 Robustness to Token Editing

Lemma 2 (Edit Resistance) *After K token edits, watermark detection power remains lower-bounded by:*

$$\textit{Power} \geq 1 - \exp\left(-\frac{N(\gamma - K/N)^2}{2\gamma(1-\gamma)}\right) \tag{37}$$

requiring $K > N(1 - \sqrt[-1]{(1-\gamma)})$ to defeat detection.

Significance: Formalizes robustness against content-preserving edits - attackers must alter a linear fraction of tokens ($\propto N$) to remove the watermark, inevitably damaging content integrity.

D.4 Visual-Semantic Coupling

Theorem 4 (SCT Invariance) *Semantic Critical Tokens maintain relative rankings under visual perturbations ΔX_v:*

$$\mathbb{P}(\textit{rank}(\Phi(l)|_{X_v+\Delta X_v}) = \textit{rank}(\Phi(l)|_{X_v})) \geq 1 - C\|\Delta X_v\|_F \tag{38}$$

where C depends on VLA model Lipschitz constants.

Analysis: Demonstrates that our visual grounding mechanism resists moderate adversarial image perturbations, as SCT rankings remain stable under controlled visual changes.

This theoretical framework rigorously establishes VLA-Mark's key properties: semantic consistency through cross-modal alignment, adaptive watermark strength via entropy regulation, statistical detectability, and robustness against content-preserving attacks. The analysis guides parameter selection while providing formal guarantees absent in previous works.