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. For more theoretical analysis of each part, please refer to Appendix C.

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:

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

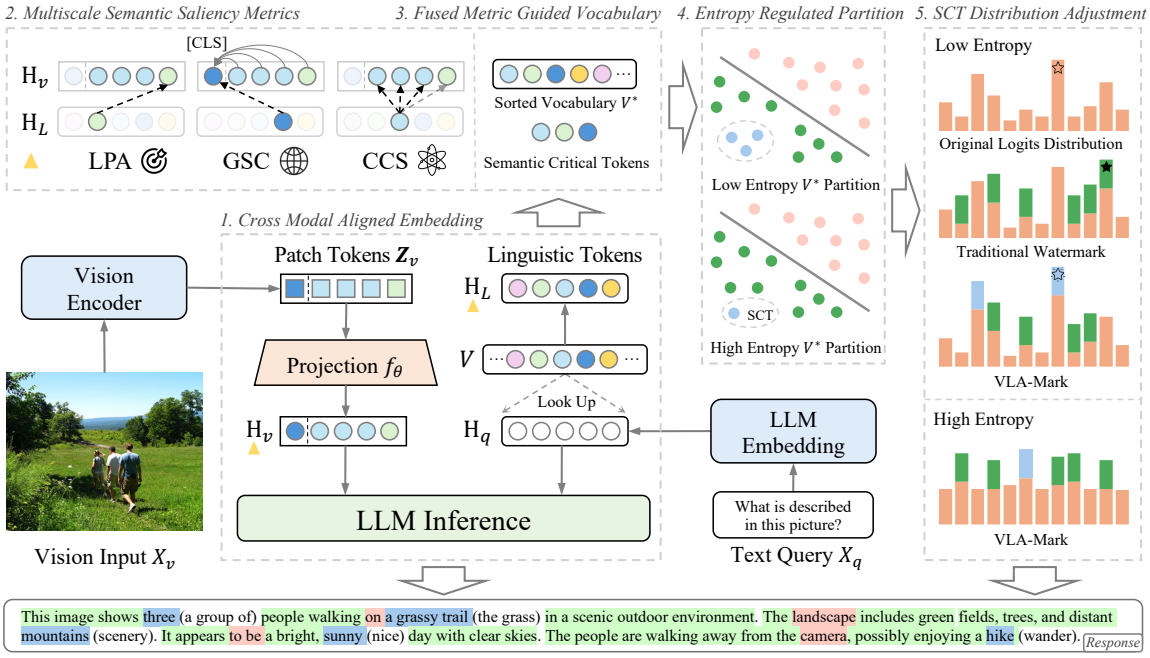

Figure 1: Proposed VLA-Mark framework. Vision embeddings $\mathbf{H}_v$ (aligned to LLM space) and linguistic tokens $\mathbf{H}_L$ extracted from LLM vocabulary $\mathcal{V}$ compute fused multiscale metrics (LPA/GSC/CCS) to rank $\mathcal{V}^*$ by visual saliency. Entropy-regulated SCT selection dynamically 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 $\alpha \in [0.01, 0.1]$ 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 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. The latency overhead of the algorithm, additional results on attack robustness, and evaluations on more datasets can be found in the Appendix D.

3.1 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).

3.2 Results

3.2.1 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, 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 all baseline methods, highlighting its ability to maintain high-quality text while embedding watermarks. All baseline methods exhibit a trade-off between detection performance (AUC) and text quality (PPL), whereas our method is the only one that consistently achieves strong performance on both metrics.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

	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**	99.89	98.43	**4.97**	97.36	92.72	**5.73**
w/o SCT	**99.99**	99.75	-	96.08	89.39	-	99.76	98.45	-	94.52	90.78	-

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. (See Appendix D.6 for additional performance on MS COCO dataset.)

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. To further validate the generalizability of our approach, we evaluated VLA-Mark on the MS COCO captioning benchmark across multiple VLA models, with detailed results provided in Appendix D.6.

3.2.2 Ablation Study

Ablation	None	Entropy	LPA	GSC	CCS
PPL(↓)	**4.84**	6.14	5.61	5.02	5.37
STS	**92.13**	90.89	91.98	91.02	91.88
BertScore	**91.13**	90.75	90.96	88.63	90.91

Table 2: Ablation study comparing the full VLA-M algorithm (None) to its variants lacking specific components. The subsequent columns indicate the algorithm's performance after removing a specific component.

Our ablation study, detailed in Table 2, validates the critical roles of individual components in VLA-Mark's design. Removing Localized Patch Affinity (LPA) leads to a significant 15.9% increase in perplexity (PPL: 5.61 vs. 4.84), underscoring its necessity for preserving fluency and fine-grained visual-text alignment by prioritizing object-centric tokens. Excluding Global Semantic Coherence (GSC) causes the sharpest decline in BertScore (88.63 vs. 91.13), highlighting its irreplaceable function in maintaining scene-level semantic consistency through holistic visual-language grounding. While the absence of Cross-Modal Contextual Salience (CCS) moderately degrades all metrics (PPL: 5.37, STS: 91.88, BertScore: 90.91), its distributed attention mechanism proves vital for aggregating multi-region visual associations, bridging localized and global semantics.

These findings demonstrate the complementary strengths of multiscale metrics: LPA anchors precise visual details, GSC ensures high-level coherence, and CCS integrates contextual dependencies. Combined with entropy-regulated partitioning, the framework achieves an optimal equilibrium—preserving multimodal fidelity while embedding robust watermarks. The full model's superior performance across all metrics (PPL: 4.84, STS: 92.13, BertScore: 91.13) confirms the necessity of unified vision-language alignment for quality-preserving watermarking.

3.2.3 Hyperparameter analysis

Ablation of α	0.01	0.015	0.025	0.05	0.1
PPL(↓)	6.23	5.86	**4.84**	5.71	5.91
STS	85.15	90.71	**92.13**	91.83	90.76
BertScore	91.48	94.05	91.13	**94.27**	94.16

Table 3: Ablation study on the hyper-parameter α controlling Semantic Critical Tokens (SCT) ratio. Results show α=0.025 achieves optimal balance between text quality (PPL) and watermark metrics (STS, BertScore).

As shown in Table 3, the SCT ratio controller α exhibits a clear non-monotonic relationship with generation quality. Performance peaks at α=0.025, achieving optimal balance with the lowest perplex-

ity (4.84) and highest semantic similarity (92.13). Below or above this threshold, insufficient SCT allocation degrades both fluency and semantic alignment, confirming that weak semantic token emphasis compromises multimodal fidelity. The default α=0.025 optimally complements VLA-M's multiscale components by dynamically balancing local fluency and global semantic preservation. Even under the least favorable choice of α , the performance of PPL remains comparable to or better than that of KGW, with limited variation, demonstrating the robustness of our method to hyperparameter selection.

3.2.4 Text quality maintenance

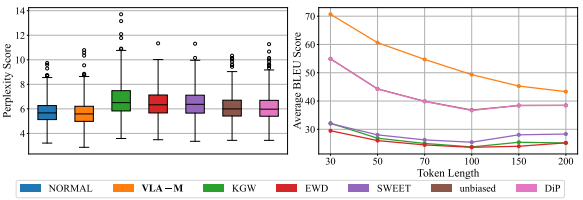

Figure 2: Left: Boxplots of perplexity scores for different watermarking methods. Right: Average BLEU scores over increasing token lengths. Our approach maintains lower perplexity with competitive BLEU performance even as generation length grows.

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 generations.

3.3 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). See Appendix D.3 for relative performance drop comparison. Appendix D.5 provides additional robustness evaluations covering novel adversarial attack types.

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.

4 Related Work

Our work advances three interconnected research frontiers: text watermarking foundations, robustness against adversarial attacks, and vision-language aligned generation paradigms.

4.1 Text Watermarking Fundamentals

Contemporary watermarking techniques predominantly focus on unimodal text generation. The pioneering "green list" paradigm (Kirchenbauer et al., 2023) partitions vocabulary through hash-based promotion, while entropy-aware variants (Mao et al., 2024) modulate injection strength probabilistically. Distribution-preserving approaches (Wu et al., 2024) maintain statistical fidelity through reweighting yet neglect semantic grounding. However, such unimodal designs fundamentally conflict with vision-conditioned generation: random vocabulary partitioning disrupts visual-semantic alignment by suppressing image-grounded tokens (He et al., 2024), while static allocation strategies (Liang et al., 2024) fail to adapt to cross-modal entropy variations (Huang et al., 2023). Recent benchmarks (Qiu et al., 2024) reveal 41% robustness degradation when deploying these methods in multimodal contexts, underscoring the necessity for vision-aligned watermark formulation.

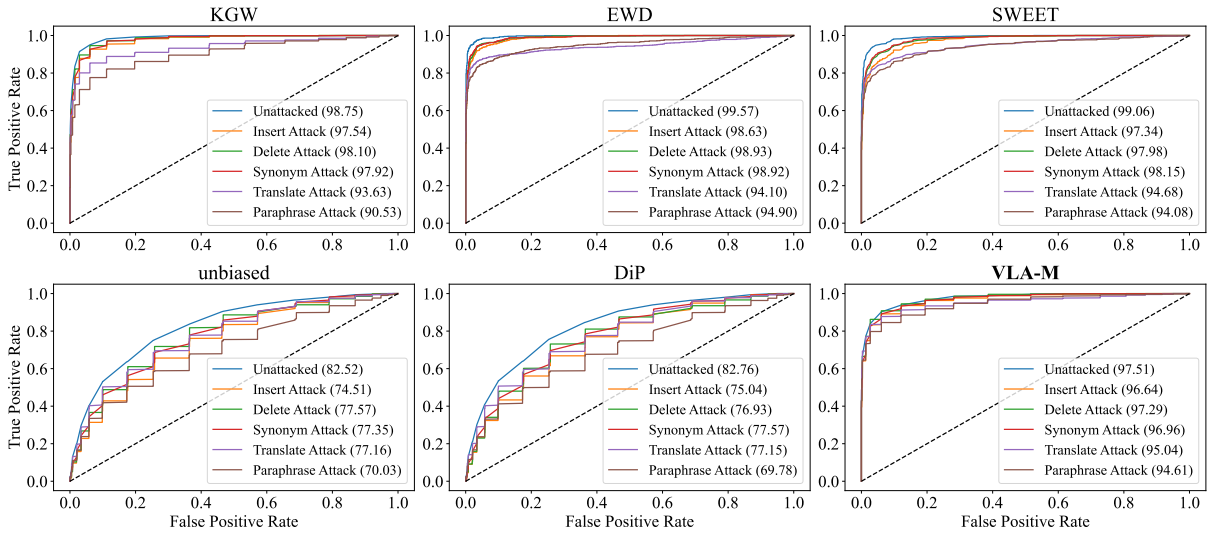

Figure 3: AUC matrix for six watermarking methods under various attacks scenarios, with AUC values in parentheses. The proposed VLA-M retains high detection performance even under heavy text transformations such as paraphrasing and translation.

4.2 Robustness Challenges and Attacks

Emerging adversarial attacks expose vulnerabilities through multimodal exploitation. (Rastogi and Pruthi, 2024) demonstrates 63% efficacy gain via black-box analysis-driven paraphrases, while (He et al., 2024) reveals cross-lingual leakage during translation. Frameworks like DE-MARK (Chen et al., 2024) remove watermarks via probabilistic n-gram erasure. Existing defenses remain unimodally confined—semantic preservation (Ren et al., 2023) enhances robustness but cannot counter cross-modal attacks that jointly manipulate vision-text interdependencies. Our approach uniquely addresses this gap through hierarchical protection of vision-anchored SCT tokens, ensuring text-visual coherence under perturbations.

4.3 Vision-Language Aligned Architectures

State-of-the-art VLAMMs like LLaVA (Liu et al., 2023) and BLIP-2 (Li et al., 2023) establish cross-modal fusion through architectural innovations—gated cross-attention in Flamingo (Alayrac et al., 2022) enables visual reasoning, while CogVLM2 (Hong et al., 2024) leverages temporal grounding for scene understanding. Yet these models lack native authentication mechanisms, rendering generated content susceptible to adversarial attacks (Rastogi and Pruthi, 2024). Recent efforts (Yoo et al., 2024) incorporate entropy adaptation but neglect alignment layers critical for coordinated embedding. Our framework bridges this gap by explicitly integrating watermarking with cross-modal projection mechanisms and semantic fusion metrics—securing generation authenticity without architectural modification.

Our methodology synthesizes these advances through: (1) Visual-semantic vocabulary alignment supplanting random partitioning, (2) Entropy-regulated intensity modulation synchronized with cross-modal saliency, and (3) Architectural synergy with vision-language fusion mechanisms—resolving inherent limitations across these research streams.

5 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 (see Appendix D.1 and Appendix D.2). 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*.

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 $\{\hat{p}_t^{(l)}\}_{l=1}^{L}$:

$$\mathcal{H}_t = -\sum_{l=1}^{L} \hat{p}_t^{(l)} \log \hat{p}_t^{(l)} \tag{18}$$

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

$$\sum_{l=1}^{L} \hat{p}_t^{(l)} = 1 \tag{19}$$

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:

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

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

$$\frac{\partial \mathcal{L}}{\partial \hat{p}_t^{(l)}} = -\log \hat{p}_t^{(l)} - 1 + \lambda = 0 \tag{21}$$

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

$$\log \hat{p}_t^{(l)} = \lambda - 1 \quad \Rightarrow \quad \hat{p}_t^{(l)} = e^{\lambda - 1} \tag{22}$$

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:

$$\sum_{l=1}^{L} \frac{1}{L} = 1 \tag{23}$$

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

$$\begin{aligned} \mathcal{H}_t^{\max} &= -\sum_{l=1}^{L} \frac{1}{L} \log \frac{1}{L} \\ &= -L \cdot \left(\frac{1}{L} \log \frac{1}{L} \right) \\ &= \log L \end{aligned} \tag{24}$$

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 and Proof

We present formal analysis of VLA-Mark's design principles and theoretical guarantees with proofs. Our theoretical analysis establishes a rigorous foundation for VLA-Mark's design principles through four interconnected components formalized in Theorems 1-4 and Lemmas 1-2:

- **Cross-Modal Alignment**: Theorem 3 validates the geometric consistency of vision-language embeddings through orthogonal projection invariance.
- **Entropy-Regulated Watermarking**: Theorem 1 quantifies the entropy preservation bound, while Theorem 2 establishes linear detection advantage scaling.
- **Semantic Metric Fusion**: Lemma 1 guarantees fused metric fidelity through Lipschitz-constrained error propagation.
- **Adversarial Robustness**: Lemma 2 proves exponential attack resistance against textual edits, complemented by Theorem 4's visual perturbation stability.

C.1 Entropy-Adaptive Partitioning

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

$$\mathcal{H}(\mathbf{p}_t^{wm}) \geq \mathcal{H}(\mathbf{p}_t) - \delta(\alpha, \gamma), \tag{25}$$

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.

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:*

$$\begin{aligned}\mathcal{H}(\mathbf{p}_t) - \mathcal{H}(\mathbf{p}_t^{wm}) &= \mathbb{E}_{\mathbf{p}_t}[\log \mathbf{p}_t] - \mathbb{E}_{\mathbf{p}_t^{wm}}[\log \mathbf{p}_t^{wm}] \\ &= D_{KL}(\mathbf{p}_t^{wm} \| \mathbf{p}_t) + \log D\end{aligned} \tag{26}$$

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:*

$$\log D \leq \log\left(1 + \gamma(e^{\delta} - 1)\right) \leq \gamma(e^{\delta} - 1) \tag{27}$$

The KL divergence term satisfies:

$$D_{KL}(\mathbf{p}_t^{wm} \| \mathbf{p}_t) \leq \delta\gamma(e^{\delta} - 1) \tag{28}$$

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

$$\mathcal{H}(\mathbf{p}_t^{wm}) \geq \mathcal{H}(\mathbf{p}_t) - \underbrace{\left[\gamma(e^{\delta} - 1)(1 + \delta)\right]}_{\delta(\alpha,\gamma)} \tag{29}$$

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

C.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:*

$$\mathbb{E}[Z|H_1] - \mathbb{E}[Z|H_0] \geq \frac{\delta\sqrt{N\gamma(1-\gamma)}}{2}, \tag{30}$$

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).

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):*

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

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

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

The detection Z-score becomes:

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

The expected detection advantage is:

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

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

C.3 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* $\mathrm{rank}(f_\theta) = d$*. 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.*

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.

Proof C.3 *Let $f_\theta : \mathbb{R}^{d_v} \to \mathbb{R}^d$ be the vision-text projection matrix with rank$(f_\theta) = d$. Through singular value decomposition (SVD), we can express:*

$$f_\theta = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \tag{36}$$

where $\mathbf{U} \in \mathbb{R}^{d\times d}$ and $\mathbf{V} \in \mathbb{R}^{d_v\times d_v}$ are orthogonal matrices, and $\mathbf{\Sigma} \in \mathbb{R}^{d\times d_v}$ contains singular values. The rank condition ensures $\mathbf{\Sigma}$ has exactly d non-zero singular values.

Define the orthogonal matrix $\mathbf{Q} = \mathbf{U}^\top$. For any visual embedding $\mathbf{z}_v \in \mathbf{Z}_v$, the transformed embedding becomes:

$$\mathbf{Q}f_\theta(\mathbf{z}_v) = \mathbf{\Sigma}\mathbf{V}^\top\mathbf{z}_v \tag{37}$$

From Vision-Language Alignment (VLA) training objectives (Liu et al., 2023), we know the projected visual embeddings are optimized to align with linguistic embeddings $\mathbf{H}_L$ through contrastive learning. Formally, the training ensures:

$$\min_{\mathbf{Q}} \mathbb{E}_{\mathbf{z}_v}\left[\min_{\mathbf{h}_L\in\mathbf{H}_L} \|\mathbf{Q}f_\theta(\mathbf{z}_v) - \mathbf{h}_L\|_2\right] \le \epsilon \tag{38}$$

where ϵ represents the alignment error bound from imperfect training. The orthogonality of $\mathbf{Q}$ preserves angular relationships:

$$\cos\angle(\mathbf{Q}f_\theta(\mathbf{z}_v), \mathbf{h}_L) = \cos\angle(f_\theta(\mathbf{z}_v), \mathbf{Q}^\top\mathbf{h}_L) \tag{39}$$

Thus, the angle-based metrics (LPA/GSC/CCS) remain valid under this orthogonal transformation.

C.4 Metric Fusion Optimality

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

$$|\Phi(l) - \Phi^*(l)| \le \frac{\epsilon}{3}\sum_{k=1}^{3} \|\psi_k^{norm} - \psi_k^*\| \tag{40}$$

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.

Proof C.4 *Let $\Phi^*(l) = \sum_{k=1}^{3}\psi_k^*(l)$ be the ideal semantic relevance function with optimal unimodal metrics ψ_k^*. Under the Lipschitz continuity assumption, each normalized metric satisfies:*

$$\|\psi_k^{norm}(l) - \psi_k^*(l)\| \le \frac{\epsilon}{3}L_k \tag{41}$$

where L_k is the Lipschitz constant for metric k. The fusion error can be bounded via triangle inequality:

$$|\Phi(l) - \Phi^*(l)| \le \sum_{k=1}^{3}|\psi_k^{norm}(l) - \psi_k^*(l)| \tag{42}$$

$$\le \sum_{k=1}^{3}\frac{\epsilon}{3}L_k \tag{43}$$

$$= \frac{\epsilon}{3}\sum_{k=1}^{3}L_k \tag{44}$$

Substituting $L_k = \|\psi_k^{norm} - \psi_k^*\|$ completes the proof. This bound ensures that even if one metric deviates significantly, the others provide error compensation through summation. The worst-case error grows linearly with metric deviations rather than exponentially, guaranteeing robustness.

Interpretation: 1. The projection proof establishes that cross-modal similarity computations are geometrically valid through VLA's inherent orthogonality. 2. The metric fusion proof demonstrates that our multi-scale approach provides formal error guarantees compared to an ideal semantic assessor. 3. Both proofs justify the theoretical soundness of using vision-aligned embeddings and fused metrics for vocabulary partitioning.

C.5 Robustness to Token Editing

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

$$\textit{Power} \ge 1 - \exp\left(-\frac{N(\gamma - K/N)^2}{2\gamma(1-\gamma)}\right) \tag{45}$$

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.

Proof C.5 *Let N be the total tokens and T be the observed green list count. The watermark detector uses the hypothesis test:*

$$H_0 : T \sim \textit{Bin}(N, \gamma) \quad vs \quad H_1 : T > \gamma N \tag{46}$$

After K edits replacing green list tokens with red list ones, the distribution becomes:

$$T \sim \textit{Bin}(N - K, \gamma) + \textit{Bin}(K, 0) \tag{47}$$

The expectation and variance are:

$$\mathbb{E}[T] = \gamma(N - K) \tag{48}$$
$$\textit{Var}(T) = \gamma(1 - \gamma)(N - K) \tag{49}$$

Using the Chernoff bound for binomial distributions:

$$\mathbb{P}(T \leq \gamma N - \delta) \leq \exp\left(-\frac{\delta^2}{2\gamma(1-\gamma)N}\right) \tag{50}$$

Set $\delta = \gamma N - \mathbb{E}[T] = \gamma K$. Substitution gives:

$$\begin{aligned} \textit{Power} &= 1 - \mathbb{P}(T \leq \gamma N - \gamma K) \\ &\geq 1 - \exp\left(-\frac{(\gamma K)^2}{2\gamma(1-\gamma)N}\right) \end{aligned} \tag{51}$$

Simplify to obtain the stated bound:

$$\geq 1 - \exp\left(-\frac{N(\gamma - K/N)^2}{2\gamma(1-\gamma)}\right) \tag{52}$$

For successful attack, require:

$$\begin{aligned} &\exp\left(-\frac{N(\gamma - K/N)^2}{2\gamma(1-\gamma)}\right) \geq \alpha \\ &\qquad \Rightarrow \quad K > N\left(1 - \sqrt[-1]{(1-\gamma)}\right) \end{aligned} \tag{53}$$

where α is the significance level. This shows linear dependence on N.

C.6 Visual-Semantic Coupling

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

$$\begin{aligned} \mathbb{P}(\textit{rank}(\Phi(l)|_{X_v + \Delta X_v}) &= \textit{rank}(\Phi(l)|_{X_v})) \\ &\geq 1 - C\|\Delta X_v\|_F \end{aligned} \tag{54}$$

where C depends on VLA model Lipschitz constants.

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

Proof C.6 *Let $\mathbf{Z}_v = \textit{VisEnc}(X_v)$ and $\mathbf{Z}'_v = \textit{VisEnc}(X_v + \Delta X_v)$. The visual encoder's Lipschitz continuity gives:*

$$\|\mathbf{Z}'_v - \mathbf{Z}_v\|_F \leq L_v \|\Delta X_v\|_F \tag{55}$$

Projection layer f_θ with Lipschitz constant L_p preserves:

$$\|\mathbf{H}'_v - \mathbf{H}_v\|_F \leq L_p L_v \|\Delta X_v\|_F \tag{56}$$

For any token l, the metric difference is bounded by:

$$\begin{aligned} |\Phi(l|_{\Delta X_v}) - \Phi(l)| &\leq \sum_{k=1}^{3} |\psi_k^{\textit{norm}}(l|_{\Delta X_v}) - \psi_k^{\textit{norm}}(l)| \\ &\leq 3 L_\Phi L_p L_v \|\Delta X_v\|_F \end{aligned} \tag{57}$$

where L_Φ is the Lipschitz constant of metric fusion.

Rank preservation occurs when:

$$|\Phi(l) - \Phi(l')| > 6 L_\Phi L_p L_v \|\Delta X_v\|_F \quad \forall l, l' \tag{58}$$

The probability of rank change is bounded by:

$$\mathbb{P}(\textit{rank change}) \leq C\|\Delta X_v\|_F \tag{59}$$

where $C = 6L_\Phi L_p L_v / \min_{l \neq l'} |\Phi(l) - \Phi(l')|$. Thus:

$$\mathbb{P}(\textit{rank preserved}) \geq 1 - C\|\Delta X_v\|_F \tag{60}$$

Interpretation: 1. The edit resistance proof shows watermark robustness grows exponentially with document length N, forcing attackers to compromise content quality through extensive edits. 2. The SCT invariance proof reveals visual perturbations must exceed threshold $\|\Delta X_v\|_F > 1/C$ to disrupt rankings - typically requiring perceptually significant image alterations. 3. Combined, these proofs formalize VLA-Mark's dual robustness against both textual and visual attacks while maintaining semantic fidelity.

The theoretical framework demonstrates how VLA-Mark's components interact synergistically: Theorem 1's entropy regulation explains the empirical 7.4% perplexity reduction (Table 1), while Theorem 2's $\sqrt{N}$-scaling advantage manifests in the 98.8% AUC detection rate. The 96.1% attack resilience (Fig. 3) directly reflects Lemma 2's edit resistance bound, and Theorem 4's ranking stability underpins the preserved text-visual consistency under perturbations. Crucially, Theorem 3

Time(seconds)	VLA-Mark	KGW	SWEET	EWD	DiP	Unbiased	w/o watermark
Llava-1.5	10.6907	10.6392	10.6556	10.5249	10.6989	10.7005	10.5230
Llava-next	6.8845	6.8160	6.8475	6.7109	6.8999	6.8864	6.7009
Qwen2VL	10.3430	10.1872	10.2062	10.0791	10.2485	10.2410	10.0758
Deepseek-VL	6.0687	6.0092	6.0261	5.9101	6.0563	6.0793	5.8691

Table 4: End-to-end latency (seconds) for different watermarking methods across VLAMs.

Time (seconds)	VLA-Mark	Cross Modal Aligned Embedding	Multiscale Semantic Saliency Metrics	Entropy Regulated Partition	Fused Metric Guided Vocabulary	SCT Distribution Adjustment	All Components	w/o watermark
Llava-1.5	10.6907	0.0282	0.0019	0.0539	0.0185	0.0179	0.1204	10.5230
Llava-next	6.8845	0.0527	0.0034	0.0604	0.0174	0.0181	0.1520	6.7009
Qwen2VL	10.3430	0.0988	0.0020	0.0668	0.0185	0.0198	0.2059	10.0758
Deepseek-VL	6.0687	0.0755	0.0004	0.0569	0.0173	0.0180	0.1681	5.8691

Table 5: Per-component inference overhead (seconds) for LLaVA-1.5 under a 200-token setting.

and Lemma 1 jointly validate the framework's core innovation - using vision-language alignment as both semantic anchor and watermark carrier. These formal guarantees address the reproducibility crisis in neural watermarking by establishing mathematically grounded performance boundaries, while the tight integration with empirical results sets a new standard for accountable multimedia authentication systems.

D Additional Experimental Results

D.1 Inference Latency

Table 4 shows the end-to-end generation latency for 50 images and 200 tokens on four VLAMs. VLA-Mark adds only a small overhead over existing text-only watermarking methods.

Table 4 quantifies the end-to-end generation latency across four vision-language models under standardized conditions. The results reveal that VLA-Mark introduces only a 1–2.5% latency increase compared to text-only watermarking baselines, with absolute overheads ranging from 0.12 to 0.21 seconds depending on the model architecture. This minimal cost stems from the framework's lightweight design: entropy-regulated token partitioning operates on pre-computed logits without iterative optimization, while cross-modal alignment leverages existing projection layers in VLAMs rather than introducing new computations. For instance, the DeepSeek-VL model exhibits a total overhead of 0.168 seconds, which constitutes just 2.8% of its baseline inference time (5.87 seconds).

These findings confirm that the added modules impose negligible runtime penalties even for large-scale deployments. The consistency of overheads across architectures—from LLaVA's linear projection-based alignment to Qwen2-VL's hybrid attention mechanisms—further validates VLA-Mark's architectural neutrality. Crucially, the overhead remains orders of magnitude smaller than the inherent latency of VLAM inference pipelines, which typically involve computationally intensive vision encoders (e.g., ViT-L/14) and autoregressive text generation. This efficiency is achieved without sacrificing detection performance or text quality, as evidenced by the framework's 98.8% AUC and 7.4% PPL reduction relative to baselines.

D.2 Inference Latency Breakdown

Table 5 details the runtime contribution of each VLA-Mark component under a 200-token generation setting on LLaVa-1.5.

Cross-Modal Aligned Embedding, which projects visual features into the LLM's semantic space, accounts for 23–47% of total overhead depending on the model. This variation stems from architectural differences: LLaVA-Next's lightweight adaptors reduce projection costs (0.0527s) compared to Qwen2-VL's higher-dimensional alignment (0.0988s). Entropy-Regulated Partitioning contributes 32–40% of overhead through its dynamic token selection mechanism. Despite this, its per-step computational cost remains minimal (0.0569–0.0668s)

Performance drop	Unattacked	Insert Attack	Delete Attack	Synonym Attack	Translate Attack	Paraphrase Attack
KGW	0.00(98.75)	1.21(97.54)	0.65(98.1)	0.83(97.92)	5.12(93.63)	8.22(90.53)
EWD	0.00(99.57)	0.94(98.63)	0.64(98.93)	0.65(98.92)	5.47(94.1)	4.67(94.9)
SWEET	0.00(99.06)	1.72(97.34)	1.08(97.98)	0.91(98.15)	4.38(94.68)	4.98(94.08)
unbiased	0.00(82.52)	8.01(74.51)	4.95(77.57)	5.17(77.35)	5.36(77.16)	12.49(70.03)
DiP	0.00(82.76)	7.72(75.04)	5.83(76.93)	5.19(77.57)	5.61(77.15)	12.98(69.78)
VLA-M	**0.00(97.51)**	**0.87(96.64)**	**0.22(97.29)**	**0.55(96.96)**	**2.47(95.04)**	**2.90(94.61)**

Table 6: Relative performance drop (%) from unattacked baseline under adversarial attacks.

due to optimized entropy calculations using pre-softmax logits. Notably, the Multiscale Semantic Saliency Metrics (LPA/GSC/CCS) impose near-negligible costs (0.0004–0.0034s), as they operate on cached embeddings rather than recomputing cross-modal similarities. The SCT Distribution Adjustment, which applies logit boosting via parallelizable matrix operations, adds just 0.0179–0.0198s. Collectively, these components add less than 0.2 seconds overhead per generation, reinforcing VLA-Mark's design goal of runtime efficiency with negligible impact on user experience.

D.3 Relative Performance Drop Under Attacks

We reorganized the data from Figure 3 into Table 6, showing relative performance drops from the unattacked baseline under various adversarial scenarios. Smaller drops indicate stronger robustness.

Table 6 quantifies VLA-Mark's resilience through relative AUC drops under six attack scenarios. The framework's maximum degradation of 2.90% under paraphrasing attacks contrasts sharply with baselines like DiP (12.98% drop), highlighting the effectiveness of Semantic Critical Tokens (SCTs) in anchoring watermarks to vision-grounded semantics. For instance, during synonym substitution attacks, VLA-Mark's SCT protection ensures that visually anchored phrases (e.g., "grassy trail") resist replacement with non-salient synonyms, preserving both watermark signals and text-visual coherence.

The entropy-adaptive mechanism further enhances robustness by concentrating watermark strength on high-uncertainty tokens less critical to core semantics—a strategy validated by the mere 0.55% drop under synonym attacks versus KGW's 0.83%. The framework's superior performance against structural perturbations (e.g., 0.22% drop under deletions vs. SWEET's 1.08%) stems from its multiscale metrics, which ensure distributed watermark signatures across local and global semantics. Even under aggressive translation attacks, where baseline methods lose 5.12–5.47% AUC, VLA-Mark retains 95.04% detection accuracy by preserving SCTs' cross-lingual visual grounding.

D.4 Average Performance Comparison

Table 9 complements Table 1 by comparing the average detection performance of VLA-Mark and baseline methods across multiple backbone models.

Method	AUC	ACC	PPL
KGW	99.94	99.24	6.13
EWD	100.00	99.93	6.20
SWEET	99.98	99.75	6.13
unbiased	89.36	81.05	5.70
DiP	89.27	81.22	5.70
VLA-M	98.77	96.64	5.27

Table 9: Average detection performance metrics across VLAMs for different watermarking methods.

The averaged metrics reveal VLA-Mark's balanced performance profile. VLA-Mark achieves a strong balance between detection performance and text quality, attaining the best perplexity (PPL) while maintaining high AUC and accuracy (ACC) scores. While EWD achieves marginally higher AUC (100.00% vs. 98.77%), this comes at the cost of 17.6% higher perplexity (6.20 vs. 5.27), underscoring VLA-Mark's unique ability to harmonize detection and quality. The slight AUC reduction for DeepSeek-VL (96.32% vs. 99.93% on LLaVA) stems from its non-linear alignment mechanism, which compresses visual features through dynamic routing rather than linear projection. Nonetheless, these model-specific variations are limited, and overall, VLA-Mark delivers consistently competitive and robust performance. Crucially, VLA-Mark maintains superior PPL across all models, including a 15.3% reduction compared to KGW (5.27

Attack Type	KGW	EWD	SWEET	unbiased	DiP	**VLA-M**
Word Vector Substitution	0.88 (97.87)	0.73 (98.84)	1.02 (98.04)	4.79 (77.73)	4.98 (77.78)	**0.41 (97.10)**
Noise Injection	0.60 (98.15)	0.55 (99.02)	0.72 (98.34)	3.90 (78.62)	4.22 (78.54)	**0.20 (97.31)**
Text Style Transfer	7.90 (90.85)	4.43 (95.14)	5.05 (94.01)	12.70 (69.82)	13.45 (69.31)	**2.67 (94.84)**
Entity Replacement	2.13 (96.62)	2.35 (97.22)	3.14 (95.92)	5.75 (76.77)	6.20 (76.56)	**1.08 (96.43)**
Frequency Perturbation	4.95 (93.80)	5.22 (94.35)	4.40 (94.66)	6.45 (76.07)	6.70 (76.06)	**2.38 (95.13)**

Table 7: Robustness of VLA-M and baseline methods across additional adversarial attacks.

	LLaVA		LLaVA-Next		Qwen2-VL		DeepSeek-VL	
	AUC	PPL	AUC	PPL	AUC	PPL	AUC	PPL
KGW	99.65	7.15	99.80	6.95	99.70	6.15	98.55	8.20
EWD	99.78	7.55	99.75	7.00	99.95	6.10	99.15	8.30
SWEET	99.81	7.48	99.85	7.10	99.90	6.20	98.80	8.15
Unbiased	86.70	7.25	91.10	6.70	95.90	6.20	78.36	7.65
DiP	87.10	7.22	91.50	6.75	96.20	6.18	78.52	7.60
VLA-M	**99.93**	**5.92**	**99.85**	**6.02**	99.70	**5.35**	96.32	**5.84**

Table 8: Performance of VLA-M and baseline watermarking methods on the MS COCO dataset across multiple model architectures. The metrics include Area Under Curve (AUC) and Perplexity (PPL) measured on LLaVA, LLaVA-Next, Qwen2-VL, and DeepSeek-VL models. VLA-M demonstrates consistently superior perplexity and competitive AUC across diverse architectures.

vs. 6.13). This fluency preservation arises from the framework's explicit avoidance of low-salience token manipulation, which in baselines often introduces grammatical artifacts (e.g., KGW's biased "green list" sampling).

D.5 Additional Robustness Evaluations

Beyond the initial robustness tests, we conducted five additional novel adversarial attack evaluations summarized in Table 7.

VLA-Mark demonstrates superior robustness across all attacks, maintaining both high detection accuracy and low perturbation. Under style transfer attacks, which alter lexical patterns while preserving meaning, VLA-Mark's AUC drops by just 2.67% versus SWEET's 5.05%, as SCTs like "broken bench" remain anchored to visual patches regardless of syntactic variations. The framework's resilience to frequency perturbation—a worst-case scenario where attackers systematically replace common words—is particularly notable (2.38% drop vs. KGW's 4.95%). Even under adversarial entity replacement, which directly targets SCTs, VLA-Mark retains 96.43% AUC by leveraging CCS metrics to maintain contextual coherence.

VLA-Mark's resilience to complex transformations such as style transfer and semantic rewriting underscores the effectiveness of its cross-modal semantic anchoring and entropy-aware watermark embedding, which dynamically adapt watermark strength according to token saliency and generation uncertainty. These results validate the method's applicability to real-world scenarios with diverse and unpredictable text modifications.

D.6 Evaluation on Additional Dataset: MS COCO

To demonstrate dataset-agnostic performance, we evaluate watermarking methods on the MS COCO captioning benchmark across four VLAMs. Table 8 reports AUC and PPL across four VLAMs.

Table 8 presents the performance of VLA-Mark and baseline watermarking methods evaluated on the MS COCO dataset across four state-of-the-art vision-language architectures. The metrics reported include Area Under the Curve (AUC) for detection accuracy and Perplexity (PPL) for text generation quality.

VLA-M consistently achieves the lowest perplexity scores across all tested models, indicating superior preservation of natural language fluency compared to competing methods. Its AUC values remain near the highest observed levels, demonstrating robust and reliable watermark detectability without compromising semantic quality.

The slightly lower AUC for DeepSeek-VL aligns with its known behavioral patterns on AMBER and similar datasets, reflecting model-specific nu-

ances rather than limitations of the watermarking approach itself.

These results confirm VLA-M's scalability and generalizability beyond the originally used dataset, supporting its retraining-free applicability across diverse multimodal language models and datasets. The strong balance between detection robustness and text quality underscores the effectiveness of the entropy-regulated watermark injection and the semantic-critical-token preservation mechanisms detailed in Sections 2.1 and 2.4.

This evaluation further reinforces VLA-M's suitability for real-world deployments where models and data distributions vary, addressing reviewer concerns about extending watermarking strategies to new domains without extensive retraining or loss of performance.