AGMark: Attention-Guided Dynamic Watermarking for Large Vision-Language Models

Yue Li, Xin Yi, Dongsheng Shi, Yongyi Cui, Gerard de Melo, Linlin Wang

KDD 2026 Jeju, Korea August 9–13, 2026

Introduction

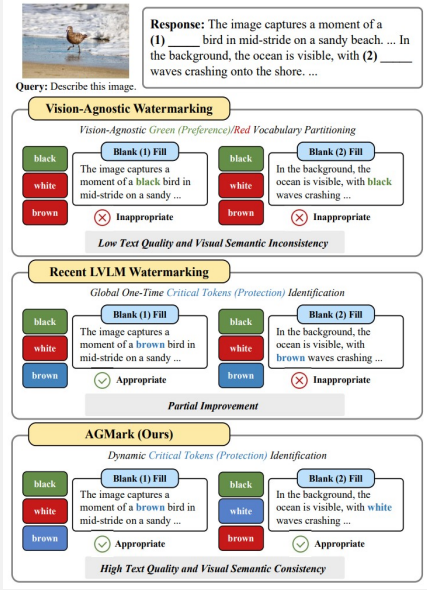

Fig. 1: Paradigm comparison of our AGMark, visionagnostic watermarking methods, and recent LVLM watermarking methods.

Background

- ***IP Protection and Provenance:*** LVLMs excel in multimodal tasks, urgently demanding watermarking for IP protection and content traceability.
- ***Limitations of Watermarking Techniques:*** Existing text-only LLM watermarks introduce vision-irrelevant vocabulary biases when applied to LVLMs.

Motivation

- ***Static Weights Fail Dynamic Decoding***: One-shot vision-critical weight computation misses dynamic visual dependence shifts during decoding, causing later-stage semantic misalignment.
- ***Ignored Weight Density Hurts Quality***: Overlooking weight distribution density preserves low-quality long-tail tokens within concentrated distributions, inducing hallucinations.

Contributions

- ***Dynamic Watermarking for LVLMs:*** AGMark is a vision-language aligned dynamic watermarking framework. It extracts semantic-critical weights via attention guidance and adaptively partitions vocabulary with uncertainty-aware evidence calibration.
- ***Superior Overall Performance:*** Across mainstream LVLMs, AGMark maintains high detectability and strong robustness against attacks, while significantly improving perplexity, BLEU, BertScore, and reducing hallucination rates (CHAIR).

Contacts

GitHub
arXiv

Method

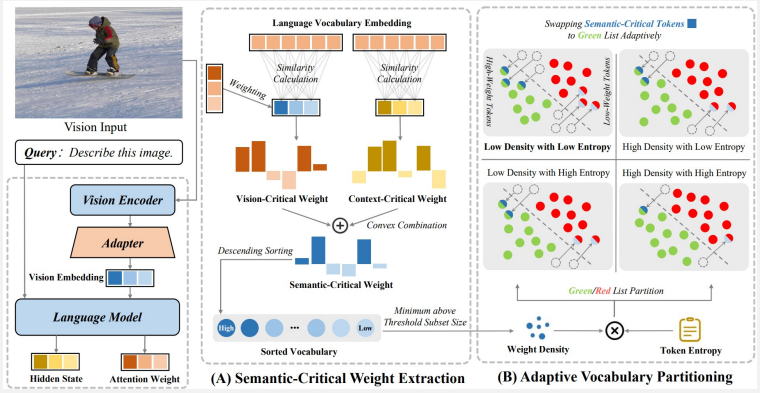

Fig. 2: Overview of our AGMark framework

AGMark consists of two components:

- ***Semantic-Critical Weight Extraction***: dynamically extracting weights via similarity calculation to fuse vision-critical weights and context-critical weights.
- ***Adaptive Vocabulary Partitioning***: leveraging token entropy and density of extracted weights to adaptively swap high evidence tokens into the green list, protecting visual fidelity.

Result

Model	Method	AMBER AUC	AMBER ACC	AMBER PPL	AMBER BLEU	AMBER BertS	AMBER STS	MS-COCO AUC	MS-COCO ACC	MS-COCO PPL	MS-COCO BLEU	MS-COCO BertS	MS-COCO STS
Llava-Next	KGW	**99.99**	**99.75**	5.29	22.88	90.00	89.43	**99.93**	98.50	5.81	24.72	89.35	87.75
	SynthID	99.91	98.50	5.29	25.93	90.28	89.79	99.88	98.50	5.33	26.37	90.28	89.87
	IE	99.58	97.38	5.56	22.37	90.11	89.64	99.78	**98.88**	5.58	20.52	89.74	88.39
	MorphMark	99.67	97.38	4.98	26.92	90.44	90.07	99.75	97.75	5.07	25.62	90.24	89.53
	VLA-Mark	99.13	96.38	5.07	25.98	90.34	90.12	99.68	98.13	5.22	20.66	89.53	87.63
	AGMark (Ours)	99.91	99.50	**4.80**	**31.88**	**91.14**	**91.85**	99.91	**98.88**	**4.95**	**29.86**	**90.76**	**90.51**
Qwen3-VL	KGW	**99.46**	**96.25**	5.83	27.66	90.74	92.54	99.74	98.00	5.88	27.28	90.73	92.34
	SynthID	99.12	95.88	6.32	26.60	90.71	92.33	99.77	98.00	6.70	24.80	90.41	91.42
	IE	99.11	96.00	6.03	27.98	90.93	92.32	99.78	**98.25**	6.26	24.87	90.39	91.65
	MorphMark	99.27	95.75	**5.81**	28.87	90.93	92.93	99.64	97.25	5.87	28.21	90.81	92.43
	VLA-Mark	99.05	96.00	5.96	31.12	91.40	93.72	99.69	97.32	6.04	31.10	91.22	92.91
	AGMark (Ours)	99.36	96.22	5.83	**31.88**	**91.46**	**93.99**	**99.82**	97.88	**5.85**	**31.95**	**91.45**	**93.81**
InternVL-3.5	KGW	99.57	**97.63**	4.64	25.99	90.16	90.67	99.65	97.63	4.62	27.78	90.25	90.59
	SynthID	**99.69**	96.96	4.65	27.27	90.21	90.89	99.49	97.25	4.60	28.38	90.34	90.11
	IE	99.19	97.00	4.52	20.82	88.75	87.12	99.59	97.38	4.54	23.50	89.48	88.33
	MorphMark	99.56	97.38	4.56	27.20	90.31	90.81	**99.68**	97.63	4.56	27.57	90.32	90.56
	VLA-Mark	99.23	97.00	4.72	25.00	90.17	91.03	99.58	96.75	4.55	27.86	90.42	90.44
	AGMark (Ours)	99.57	97.50	**4.50**	**27.45**	**90.40**	**91.12**	99.65	**97.85**	**4.53**	**30.88**	**91.08**	**91.74**

Fig. 3: Performance comparison in terms of metrics AUC , Accuracy (ACC) , Perplexity (PPL) ↓, BLEU , BertScore (BertS) , and STS .

AGMark enhances generation quality without compromising detectability.

- **Preserving Generation Quality:** AGMark achieves the best performance in the vast majority of settings.
- **Maintaining Strong Detectability**: AGMark consistently ranks among the top two methods in terms of AUC across the vast majority of settings.

We further analyze the **impact of generation length**. By dynamically preserving semantically critical tokens, AGMark minimizes fluency and coherence distortions, enabling more accurate generation in later stages.

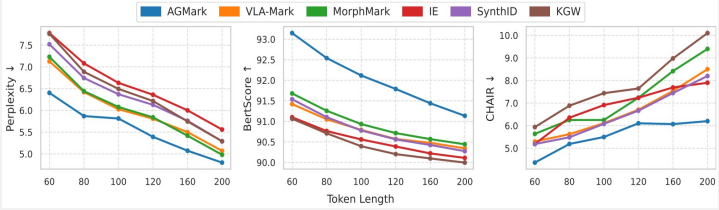

Fig. 4: Line charts depict length-wise performance retention across response token lengths ranging from 60 to 200 tokens, evaluated using: Perplexity (Left), BertScore (Middle), and CHAIR (Right).