StableGuard: Towards Unified Copyright Protection and Tamper Localization in Latent Diffusion Models

Haoxin Yang[1] **Bangzhen Liu**[1] **Xuemiao Xu**[1†] **Cheng Xu**[2†] **Yuyang Yu**[1]
Zikai Huang[1] **Yi Wang**[3] **Shengfeng He**[2]
[1]South China University of Technology [2]Singapore Management University
[3]Dongguan University of Technology
harxis@outlook.com xuemx@scut.edu.cn
{liubz.scut, cschengxu, yyyoung0611, zikaihuang0428}@gmail.com
yiwang@dgut.edu.cn shengfenghe@smu.edu.sg
https://github.com/Harxis/StableGuard

Motivation

- Traditional forensic methods are **limited to watermark extraction or tampering detection**. Some combine copyright protection with localization, but they operate **post-hoc**, adding computational cost and degrading image quality.
- **Diffusion-naive methods** embed watermarks fall short of advanced forensic needs such as **localizing manipulations**.
- Our key insight is that the **holistically distributed watermarks** are naturally robust against localized manipulations. They can serve as **reliable signals for tampering localization** via missing-feature detection.

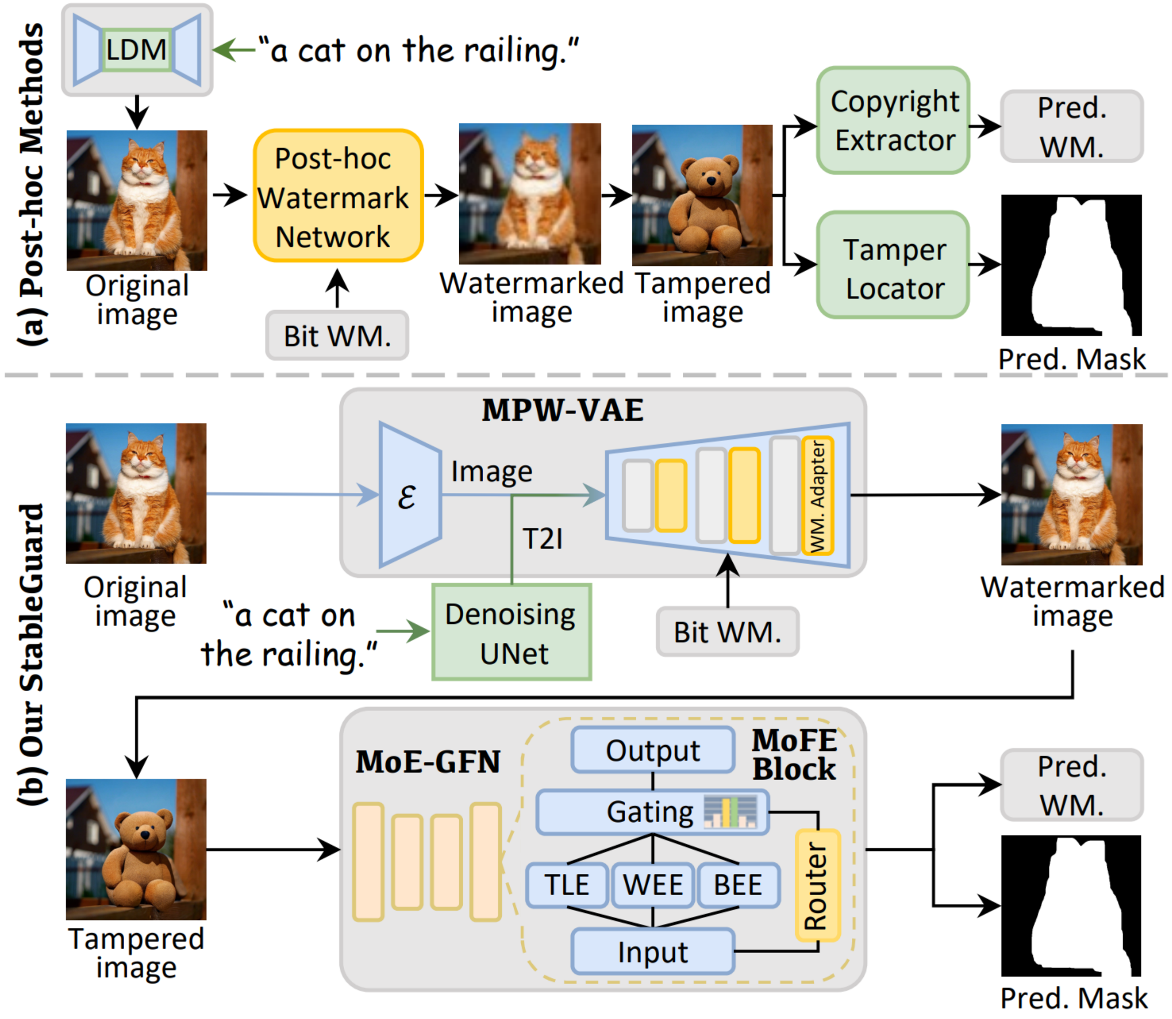


Figure 1. The main difference between StableGuard and other methods.

Solution

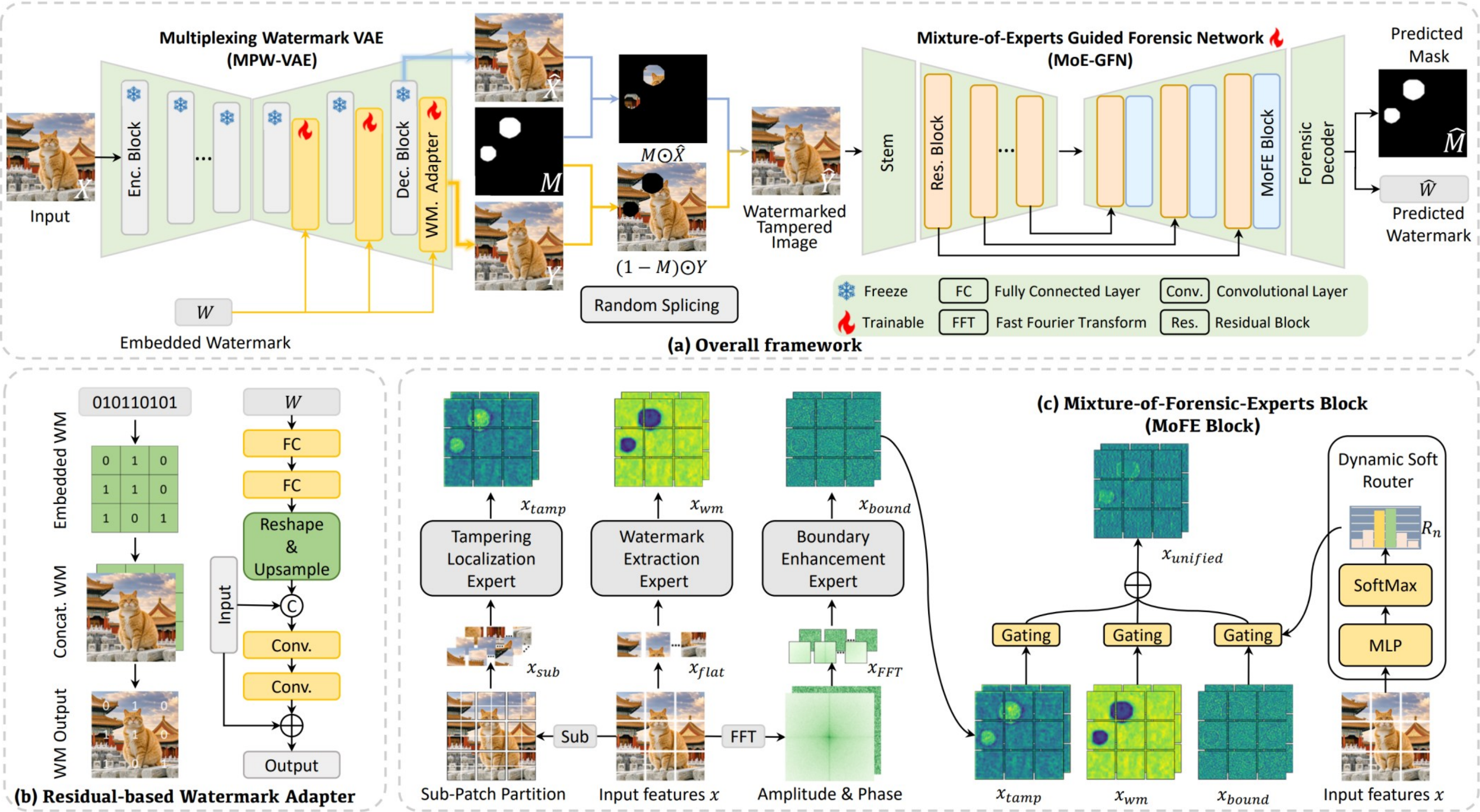


Figure 2. The overview of StableGuard.

Contribution

- We present StableGuard, a **unified proactive forensics framework** for LDMs that integrates **copyright protection and tampering localization** into image generation.
- Our **self-supervised** approach embeds an imperceptible bit watermark using a **multiplexing watermark VAE**, enabling precise protection and localization **without labeled data**.
- We introduce a tampering-agnostic **mixture-of-experts forensic network** that combines **holistic, subtle, and boundary features** for reliable watermark retrieval and accurate tampering detection under diverse attacks.
- Extensive experiments show that StableGuard **surpasses state-of-the-art methods** in accuracy, robustness, and visual fidelity.

Comparison

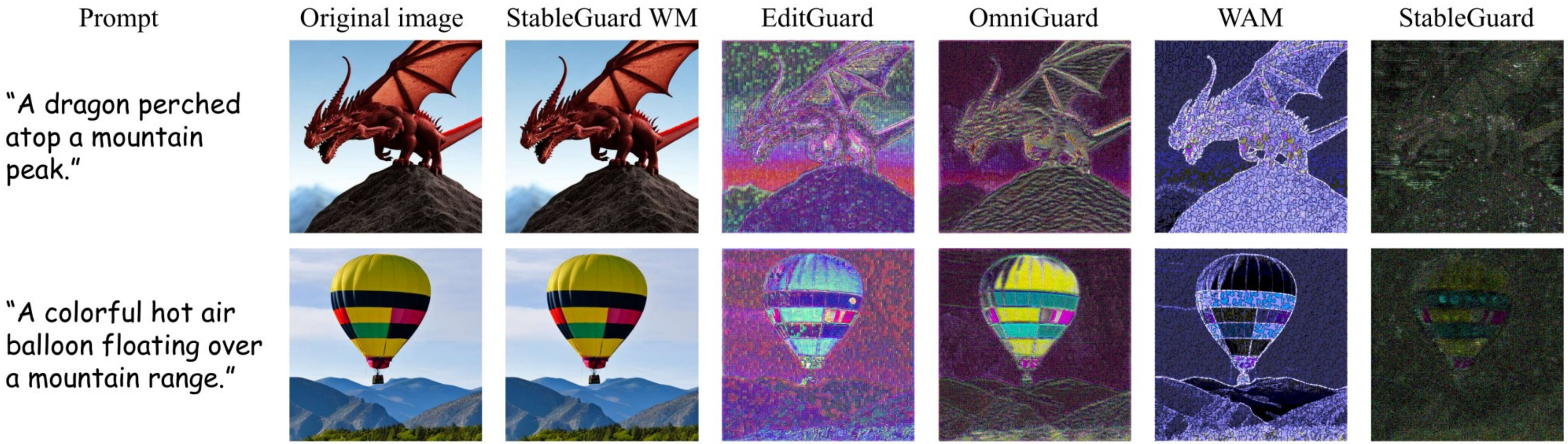
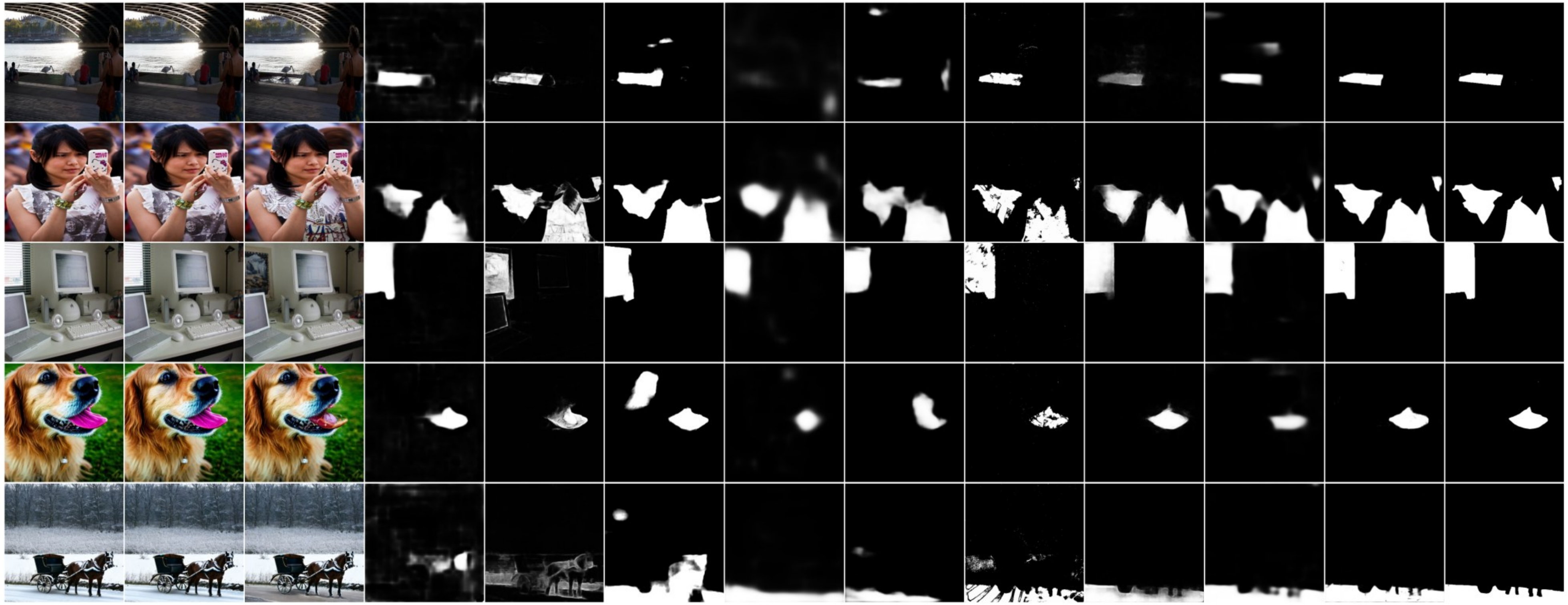


Figure 3. Comparison between Stable Diffusion VAE and a variant using our MPW-VAE alongside other watermarking methods.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m)

Figure 4. The visualization of tampering localization results on the AIGC tampering dataset.

Method	SD Inpainting [1]			SD XL [2]			Kandinsky [57]			ControlNet [4]			LAMA [58]		
	F1↑	AUC↑	IoU↑	F1↑	AUC↑	IoU↑	F1↑	AUC↑	IoU↑	F1↑	AUC↑	IoU↑	F1↑	AUC↑	IoU↑
MVSS-Net [15]	0.862	0.934	0.791	0.848	0.929	0.775	0.848	0.928	0.775	0.856	0.930	0.782	0.860	0.934	0.789
IML-ViT [16]	0.907	0.923	0.879	0.904	0.921	0.876	0.906	0.923	0.877	0.898	0.883	0.840	0.898	0.894	0.880
PSCC-Net [17]	0.898	0.976	0.829	0.899	0.977	0.830	0.894	0.975	0.825	0.899	0.977	0.830	0.898	0.976	0.829
ObjectFormer [18]	0.476	0.722	0.398	0.479	0.719	0.398	0.472	0.718	0.398	0.467	0.724	0.390	0.503	0.738	0.425
HDF-Net [19]	0.556	0.762	0.468	0.763	0.470	0.560	0.544	0.759	0.457	0.551	0.764	0.463	0.565	0.767	0.476
EditGuard [21]	0.937	0.977	0.911	0.938	0.976	0.913	0.935	0.966	0.913	0.939	0.969	0.907	0.939	0.977	0.917
OmniGuard [22]	0.853	0.964	0.810	0.867	0.973	0.824	0.868	0.966	0.830	0.858	0.965	0.815	0.864	0.969	0.823
WAM [23]	0.924	0.977	0.868	0.918	0.976	0.862	0.921	0.976	0.865	0.917	0.977	0.860	0.922	0.967	0.864
Ours	**0.980**	**0.993**	**0.962**	**0.981**	**0.991**	**0.961**	**0.980**	**0.992**	**0.960**	**0.981**	**0.993**	**0.963**	**0.979**	**0.993**	**0.961**

Table 1. Localization precision comparison on the AIGC tampering dataset.

Ablation

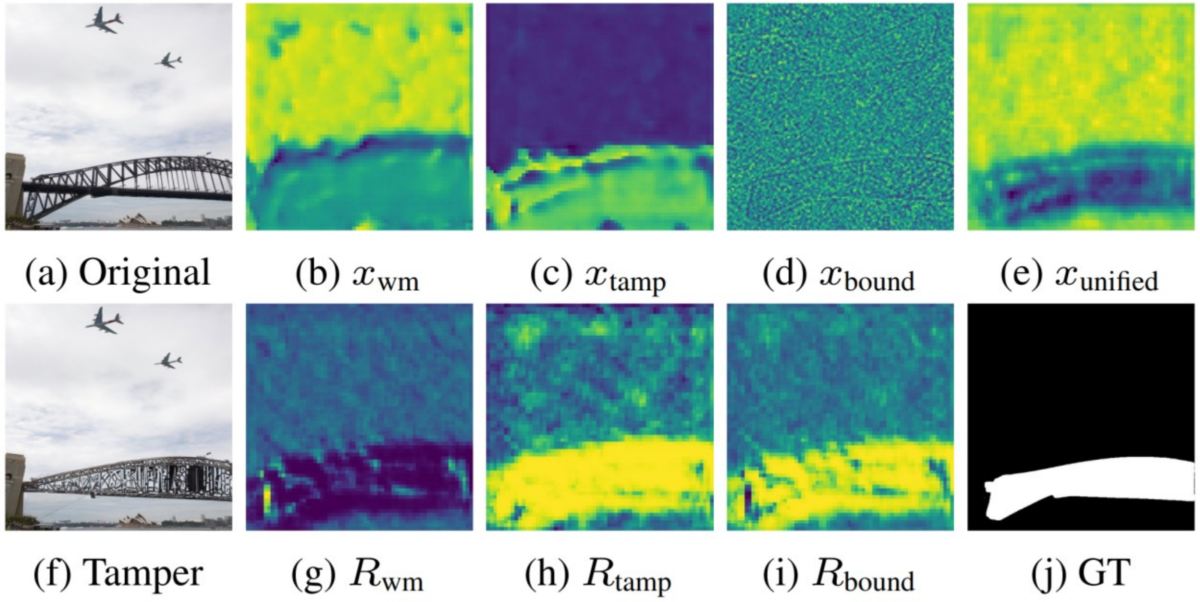


Figure 5. The visualizations of the distinct features extracted by each expert and their corresponding soft weights.

Method	F1↑	AUC↑	IoU↑	Bit Acc↑	Param↓	Flops↓
w/o MPW-VAE	0.811	0.796	0.774	99.13	52.02M	78.51G
w/o MoFE	0.931	0.920	0.905	95.12	38.11M	45.72G
w/o WEE	0.969	0.958	0.945	98.69	48.51M	70.21G
w/o TLE	0.952	0.940	0.930	98.90	48.51M	70.21G
w/o BEE	0.962	0.950	0.940	99.11	45.23M	62.71G
w/o DSR	0.966	0.955	0.948	98.97	51.26M	75.74G
w/o JOS	0.921	0.919	0.908	99.14	52.02M	78.51G
Enc	0.974	0.970	0.960	99.79	125.51M	96.15G
Enc & Dec	0.982	0.988	0.976	99.96	139.41M	104.93G
Dec †	0.980	0.992	0.961	99.98	52.02M	78.51G

Table 2. Quantitative ablation study on the AIGC tampering dataset.

Code

https://github.com/Harxis/StableGuard