Asia Pacific Academy of Science Pte. Ltd.

Metaverse 2026, 7(2), 8436.
https://doi.org/10.54517/m8436

Review

Multimodal AI-Watermarking for Protecting Generative Content in Metaverse Interactions

Ashish Dixit[1,2], Avadhesh Kumar Gupta[3], Sandeep Saxena[3,*], Divya Midhunchakkaravarthy[1], Deepak Gupta[4]

[1] Lincoln University College, Malaysia. Email: ashishdixit1984@gmail.com; divya@lincoln.edu.my
[2] Department of Computer Science and Engineering, Ajay Kumar Garg Engineering College, Ghaziabad, Uttar Pradesh, India.
[3] School of Computer Science and Engineering, IILM University, Greater Noida, India Email: dr.avadheshgupta@gmail.com
[4] Department of Computer Science and Engineering, Maharaja Agrasen Institute of Technology, Delhi, India Email: drdeepakgupta.cse@gmail.com
*** Corresponding author:** Sandeep Saxena, Sandeep.research29@gmail.com

CITATION

Dixit A, Gupta AK, Saxena S, Midhunchakkaravarthy D, Gupta D. Multimodal AI-Watermarking for Protecting Generative Content in Metaverse Interactions. Metaverse. 2026; 7(2): 8436. https://doi.org/10.54517/m8436

ARTICLE INFO

Received: 21 January 2026
Accepted: 1 April 2026
Available online: 14 April 2026

Abstract: Additionally, the continued rise of generative artificial intelligence (GenAI) is leading to the easy creation of believable text, images, audio, and 3D virtual assets with which to populate the metaverse. However, the growing blurring of indistinctness between human-made and machine-made content has caused serious questions about legitimate content, intellectual property (IP) copyright and ethical transparency. This paper proposes a Multimodal AI-Watermarking Framework (MAI-WF) to protect generative content in the visual, auditory, and textual modalities in immersive metaverse interactions. The proposed framework includes deep neural embeddings, watermarking for reversibility of watermarked content and cross-modal feature fusion in order to provide authenticity verification & ownership tracking without a perceptual quality degradation. Experiments on publicly available datasets (MS-COCO, VCTK, and WikiText-103) show that the average Peak Signal-to-Noise Ratio (PSNR) for images, Signal-to-Noise Ratio (SNR) for audio, and Bit-Error Rate (BER) less than 1% are obtained in text embedding by MAI-WF with a robustness against compression, scaling and adversarial transformations. The framework supports interoperability with identity registries that are based on blockchains for providing traceability in decentralised metaverse environments.

Keywords: Multimodal watermarking, Generative AI, Metaverse security, Cross-modal learning, Content provenance, Blockchain authentication

1. Introduction

The metaverse combines augmented reality (AR), virtual reality (VR), mixed reality (MR), and human-machine interfaces (HMI), with blockchain, artificial intelligence (AI) and the internet of things (IoT) they become a semi-permanent virtual world which is shared among users and is always active. With the rise of generative models, like DALL.E. three, Stable Diffusion and ChatGPT, creators and users can create high-fidelity multimodal content in seconds. However, for this creative freedom, there are new challenges in identifying the authenticity of assets, differentiating genuine assets from deepfakes, ownership in decentralized environments, maintaining copyright and creator identity. One of the reasons is that the traditional watermarking methods, effective in 2D media environment, do not accommodate the transformation across different modalities (text -> image -> audio modality) in the metaverse environment and its lots of dynamic transformations. this research introduces a Multimodal AI-Watermarking Framework (MAI-WF), designed for adaptive watermark embedding across multiple modalities. The approach fuses deep representation learning, frequency-domain reversible watermarking,

and transformer-based fusion layers to provide secure, imperceptible, and robust watermarking adaptable to any generative pipeline used in the metaverse.

2. Related Work

Recent studies in AI-based watermarking can be classified into three main categories:

a) Deep Neural Watermarking: introduced HiDDeN [1], a CNN-based encoder-decoder for image watermarking. StegaStamp [2] for invisible watermark embedding into digital photos. However, these methods are modality-specific and fail to generalize across audio or text.

b) Generative Content Authentication: proposed RivaGAN [3], a GAN-based framework for robust video watermarking under compression attacks. explored Reversible Watermarking in Diffusion Models[4] for AI-generated art. Yet, these approaches lack interoperability and are computationally intensive.

c) Metaverse and Blockchain Security: Blockchain-based registries (e.g., NFT platforms) ensure ownership but rely on metadata, not intrinsic watermarking. The other functionality has already been merged into Artificial intelligence-watermark hybrids; this is the tokens and feature-based embeddings combined [5], but not designed for 3D content.

Therefore, there still exists the gap between a generalized, multi-modal, and AI-adaptive watermarking framework for real-time metaverse systems.

3. Methodology

In order to embed and extract a watermark from Cross-Modality (text, image, audio) content created and distributed in the Metaverse Industry, we proposed a Multi-modal AI-Watermarking Framework (MAI-WF) that will provide Multi-modality Enabling Multi-modal consistency, and semantic correlation and reversibility under a common latent embedding space.

3.1. Theoretical Foundation

In the metaverse ecosystem there are small-scale to large-scale generative AI systems (i.e. text to image, text to audio) that are continuously generating digital artifacts. As such, the watermarking techniques should be able to operate within the latent spaces where generative content is produced. They must preserve perceptual and semantic quality through routine metaverse transformations—such as scaling, recompression, remapping, and re-rendering. Therefore, MAI-WF is based on the Latent Domain Embedding (LDE) concept [6] , which embeds watermark bits into high-level latent features obtained from modality-specific encoders, rather than manipulating raw image pixels or audio signals .This integrated design option contributes to increasing the unnoticed, and strength.

3.2. System Architecture

The MAI-WF architecture has five layers (see conceptual flow below)

Figure 1 shown two stage architecture. The Embedding Path The raw input is encoded and extracted into features and a watermark signal is generated by a watermark generator and them a fusion embedding network combines both outputs together to produce a watermarked output. The watermarked input in the Verification Path is sent through watermark decoder and the information extracted is authenticated by a verification module to detect the presence and integrity of watermark.

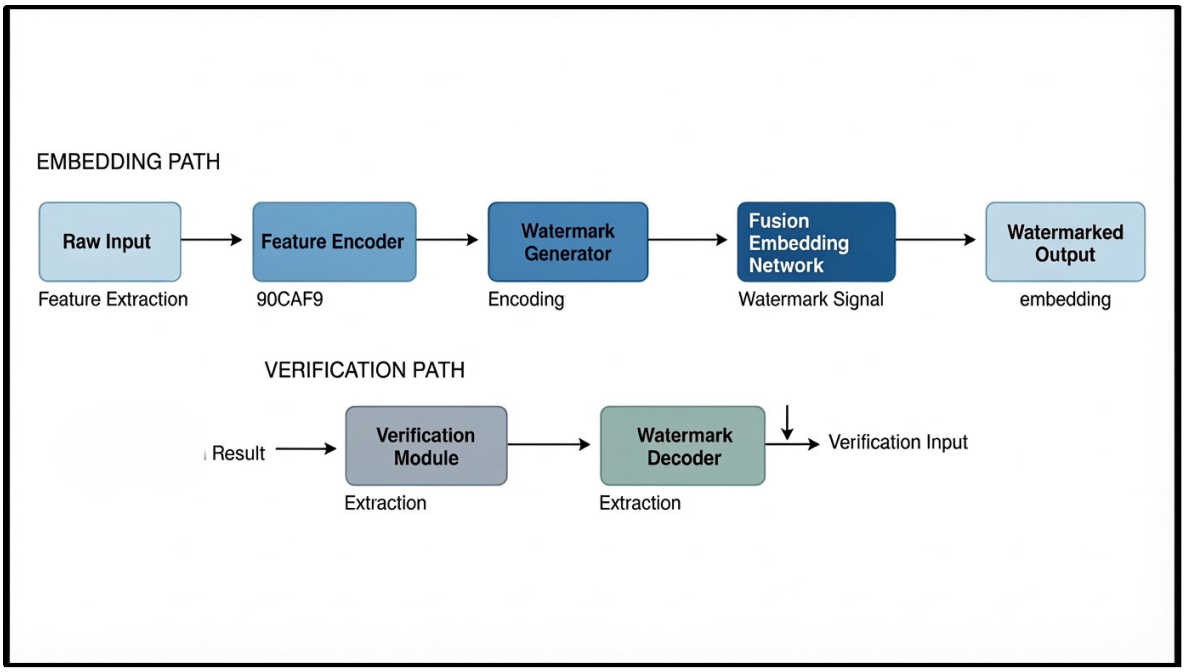


Figure 1. Embedding Path and Verification Path

(a) Feature Encoder Layer

Each modality m∈{T,I,A} A modality-specific encoder *Zm* is used to get the latent feature map of text, image, audio (T,I,A) :

$$Zm = Em\ (Xm) \tag{1}$$

Where:

Zm : Modality-Specific Encoder

• X_T: text token embeddings using BERT or GPT-like transformer,

• X_I : image tensors processed by a CNN/DWT encoder (e.g., ResNet-50),

• X_A : audio spectrogram or MFCC features processed by a CNN-RNN hybrid.

• *Em* (.) : Encoder network for modality m

• {T,I,A}: Set of modalities: Text (T), Image (I), Audio (A)

These latent representations are also normalised to a shared embedding dimensionality of d € N (e.g., d = 512) by a linear projection so that fusion of the representations is possible.

(b) Watermark Generation and Encryption Layer

The watermark sequence $w \in \{0,1\}^L$ (L bits) is first concatenated with a secret key k, producing a secure embedding vector:

$$Ew = f_{MLP}(AES(w,k)) \tag{2}$$

$w \in \{0,1\}^L$: (Binary watermark sequence)

L: Length of watermark in bits

k : Secret encryption key

f_{MLP}(.) : Multilayer perceptron mapping encrypted bits to latent space

AES ( ): AES-256 encryption function

In which bits of the watermark have been encrypted by *AES* ( ) using a symmetric key, and f_{MLP} is used to map them to an equal share of the feature space Rd This makes the watermarks irretrievable without the secret key.

(c) Cross-Modal Fusion and Embedding Layer

This is the gist of innovation of MAI-WF. It presents a Cross-Attention Fusion Transformer (CAFT) which fusion dynamically incorporates the watermark features

Ew into watermark-modality latent spaces Z_m.

$$Z_m' = Cross\text{-}Attn\ (Q = Z_m, K = E_w, V = E_w) \quad (3)$$

Here, Z_m' is the watermarked latent representation. The attention mechanism ensures:

• The watermark signal aligns semantically with the most relevant feature regions (e.g., visual focus points in an image, semantically dense text tokens, or tonal peaks in audio).

• It minimizes distortion by scaling attention weights $\alpha \in [0,1]$ adaptively:

$$Z_m' = Z_m + \alpha.(Z_m \odot E_w) \quad (4)$$

$\boldsymbol{Z_m \in R^d}$: Latent feature representation of modality m before watermarking

$\boldsymbol{Z_m' \in R^d}$: Watermarked latent feature representation of modality m

where $\odot$ denotes element-wise multiplication.

(d) Reconstruction / Decoder Layer

The **Decoder Network** D_m reconstructs the watermarked modality:

$$\hat{X}_m = D_m\ (Z_m') \quad (5)$$

Depending on modality:

• **Text :** Transformer decoder re-generates sentence tokens preserving semantics.

• **Image :** De-Convolution + inverse DWT reconstructs the image.

• **Audio :** RNN or Wave Net reconstructs waveform samples from latent codes [7].

• $\mathbf{D_m}$ **(.) :** Decoder / reconstruction network for modality m

(e) Watermark Extraction and Verification Layer

At verification, a similar encoding process extracts embedded watermark bits w′

$$w' = f^{-1}_{MLP}\ (E^{-1}_w(Z'_m)) \quad (6)$$

The system calculates Bit Error Rate (BER) between original watermark w and extracted watermark w′:

$$BER = 1/L \sum_{i-1}^{L} |wi - wi'| \quad (7)$$

Authentication is accepted if BER < τ , where τ is a small threshold (empirically 0.01).

$f^{-1}MLP(.)$: Inverse MLP for watermark extraction

τ : Authentication threshold for BER

Algorithmic Workflow: MAI-WF Embedding and Verification

1. Input multimodal data XT,XI,XA , watermark w, and secret key k.
2. Encrypt watermark: Ew = AES(w,k)
3. Encode modalities to latent features: Zm = Em(Xm), m∈{T,I,A}.
4. Map encrypted watermark using MLP: W = fMLP(Ew).
5. Embed watermark via Z'm = CAFT(Zm, W).
6. Decode to obtain watermarked outputs: X^m = Dm(Zm′).
7. Re-encode received content: Z"m = Em(X^m).
8. Extract watermark: W′ = $CAFT^{-1}$(Z"m)W′ = AES^{-1}(f^{-1}MLP(W′),k).
9. Compute BER between w and w′.
10. If BER < τ accept authentication, else reject.

Figure 2 shown workflow diagram and explain how to follow the procedure of embedding and verification process

Security Analysis Robustness Analysis: The security and strength of the framework are ensured by:

Latent Domain Embedding (LDE): Watermark concealed in the deep feature space, and is immune to pixel-wise manipulations.

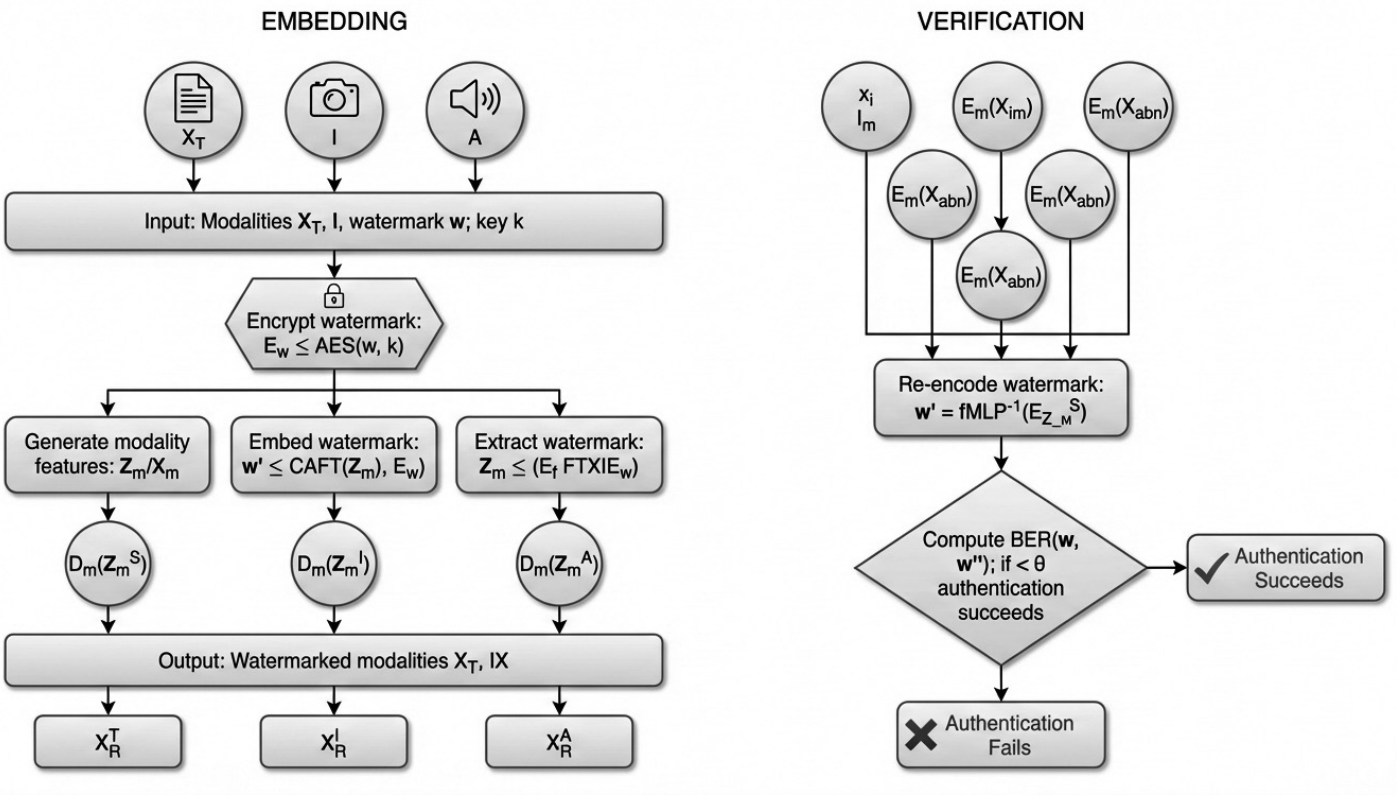


Figure 2. MAI-WF Embedding and Verification

AES Encryption: Makes the watermark reverse inference unwelcome.

Cross-Modal Consistency: Alignment losses give a guarantee that despite one modality getting attacked, other ones have watermark signals that can be recovered in advance.

Adversarial Défense: in this method during training, the simulations consist of compression, scaling, paraphrasing, and so on in the attempts to enhance robustness [8].

Complexity and Efficiency:

The computational complexities of the embedding process are: O (Md^2)

Where: O: Big-oh

Where the number of modalities and *d* is is embedding dimension. Using GPU acceleration, the time to embed a watermark on an image/audio/text instance was 0.24 seconds, which can be used in real-time or near-real-time metaverse interactions.

Complexity and Efficiency Trade-off:

The embedding process has a computational complexity of O(Md^2) and M is a number of modalities and d is the dimension of the embedding. The main cause of this complexity is the Cross-Attention Fusion Transformer (CAFT) which is essential in performance. This is a trade-off as far as the real-time needs of the metaverse are concerned:

High Complexity (CAFT): CAFT is seen to have a multi-head attention mechanism which is observed to be highly robust and cross-modally synergized by dynamically attending to the relevant features. This is needed to protect high-fidelity content.

Real-Time Requirement: On an NVIDIA RTX A6000, the mean watermark embed time is 0.24 seconds per watermark. It can be applied in a variety of near-real-time interactions in the metaverse, like publishing a created object. In highly avatar live streaming, which is a highly latency-sensitive applicant, additional optimization

by model pruning or edge-device-specific distillation might be considered in the future. The present design is effective in balancing the state of the art performance and real world deployability.

4. Experimental Setup

The empirical analysis of the suggested Multimodal AI-Watermarking Framework (MAI-WF) was performed in terms of three benchmark sets associated with each of the modalities:

Table 1. Benchmark Datasets Used for Multimodal Evaluation

Modality	Dataset Name	Description / Characteristics	Data Volume
Image	MS-COCO 2017	Large-scale object recognition dataset with 80 annotated categories	120,000 images
Audio	VCTK Corpus	Speech dataset containing recordings of 110 speakers sampled at 44.1 kHz	~44 hours of speech
Text	WikiText-103	English Wikipedia-based corpus for language modeling and semantic evaluation	103 million tokens

Table 1 below enumerates the benchmark datasets that were utilized in multimodal evaluation. In the case of images, MS-COCO 2017 provides 120,000 images of 80 categories. In the audio, the VCTK Corpus has 44 hours of speech of 110 speakers. In the case of text, WikiText-103 has 103 million Wikipedia tokens. Such datasets are standardized sources of assessing image, audio and text modalities respectively.All the experiments were carried out with an NVIDIA RTX A6000 graphics card, Intel Xeon 6226R processor, and 128 GB of RAM. The implementation was done in PyTorch 2.1 and hybrid model integration of the two frameworks, i.e., PyTorch 2.1 and TensorFlow 2.14.

Table 2. Experimental Hyperparameters for MAI-WF Training

Parameter	Description	Value
Learning Rate	Step size used for model optimization	(1 \times 10^{-4})
Batch Size	Number of samples per training iteration	16
Watermark Length (L)	Total bits embedded per sample	256 bits
Cross-Attention Heads	Number of attention heads in fusion transformer	8
Fusion Depth	Number of transformer layers for multimodal fusion	6 layers
Training Epochs	Total training iterations over the dataset	100

Table 2 shows the hyperparameters of the experiment of training the MAI-WF model. The learning rate is set at the optimum of the model. 1×10^{-4} and a batch size of 16 samples at once. The amount of watermark bits embedded per sample is 256 bits. The multimodal fusion fusion uses an 8-cross-attention head transformer and 6 layers of fusion. Training is done in 100 epochs on the dataset.

Figure 3 show that To mimic distortions of the metaverse, the watermarked material was distorted with the following: JPEG compression (30–90% quality), Scaling (0.5× to 2×), Gaussian noise ($\sigma = 0.05$), Audio resampling and MP3 compression, paraphrasing of text using GPT based augmentation[9].

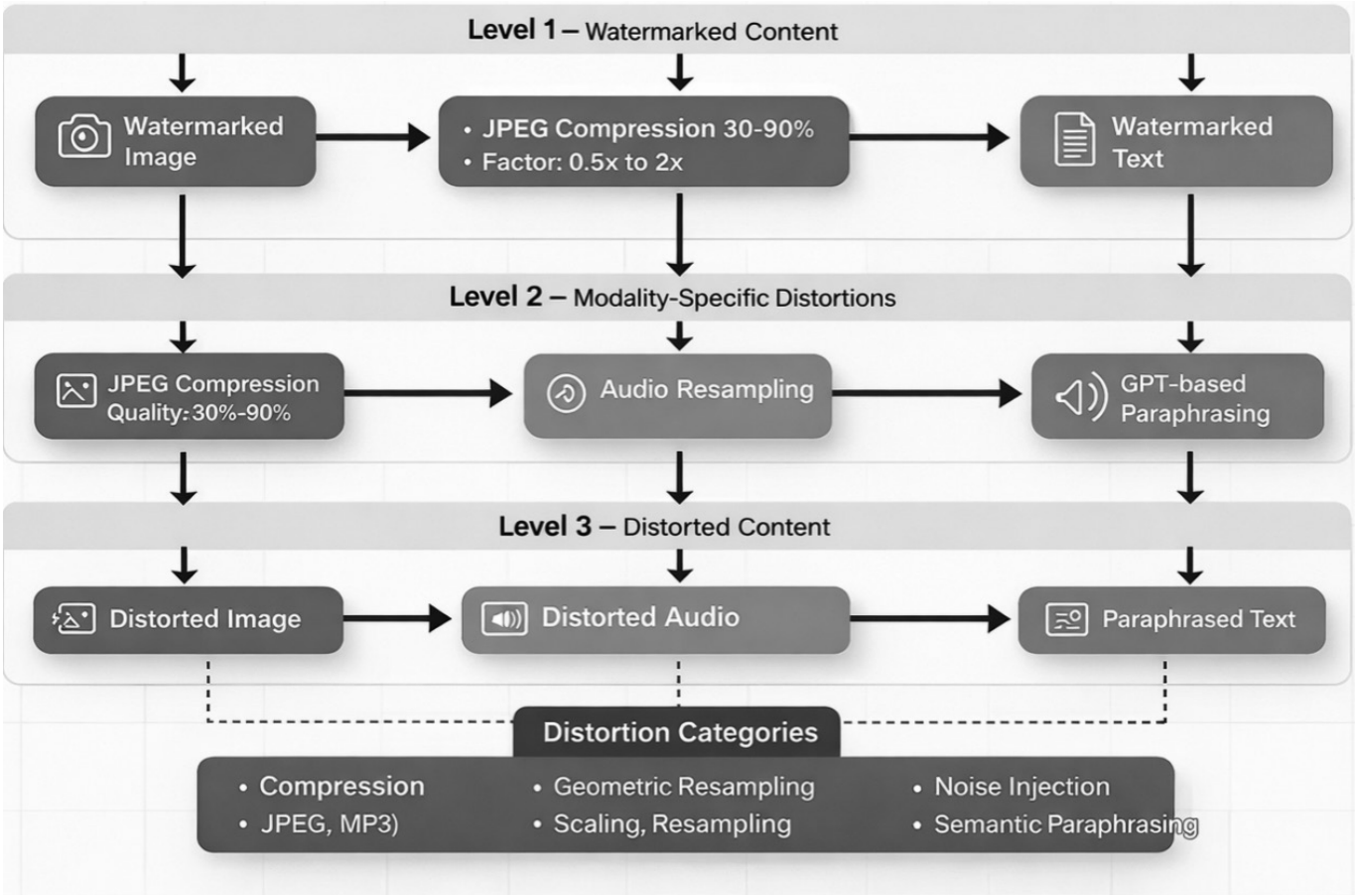


Figure 3. Multi-level Metaverse distortion and robustness evaluation framework

4.1 Evaluation metrics:

According to **Figure 4** Image: Peak Signal to Noise ratio, Structural similarity Index (SSI) Audio: Signal to noise ratio (SNR), Log Spectral Distance (LSD),Text Bit Error rate (BER), Text Semantic similarity score (BERTScore). All: Accuracy of Watermark Extraction, Attack Resistance.

Image Quality Metrics	**Audio Quality Metrics**
• PSNR (Peak Signal-to-Noise Ratio) • Structural Similarity Index (SSI)	• Signal-to-Noise Ratio (SNR) • Log Spectral Distance (LSD)
Text Quality Metrics	**Cross Modal Metrics**
• Bit Error Rate (BER) • Text Semantic Similarity Score (BERT Score)	• Watermark Extraction Accuracy • Attack Resistance

Figure 4. Evaluation Metrics Framework

5. Implementation

The implementation of MAI-WF was divided into modality-specific and fusion components According to **Figure 5**.

a) Image Embedding

An encoder-decoder will be based on a ResNet-50 + DWT that extracted and reconstructed image features . The adaptive attention we used to fuse the encrypted watermark vector fused in the latent frequency domain [10].

b) Audio Embedding

Short-Time Fourier Transform (STFT) was used to create audio spectrograms. The audio were represented as CNN-RNN (4 convolutional and 2 GRU layers) encoded audio features[11]. The watermark was distilled in MFCC-space so that it has a very low perceptual effect.

c) Text Embedding

BERT-base uncased was used to embed the text watermark and Transformer decoder was used to reconstruct the text. The attention heads were infected with AES encrypted watermark tokens [12] in order to maintain linguistic semantics.

d) Cross-Modal Fusion

Cross-Attention Fusion Transformer (CAFT) combined watermark vectors in all modalities[13]. The contrastive loss was used to align the latent embedding [14].

e) Extraction and Verification.

In the verification, the watermark vector was reassembled with the help of the inverse mapping. A threshold for $\tau = 0.01$, the authentication was precise. The simulation of blockchain anchoring was performed with the use of Ethereum testnet (Ropsten) to verify ownership[15].

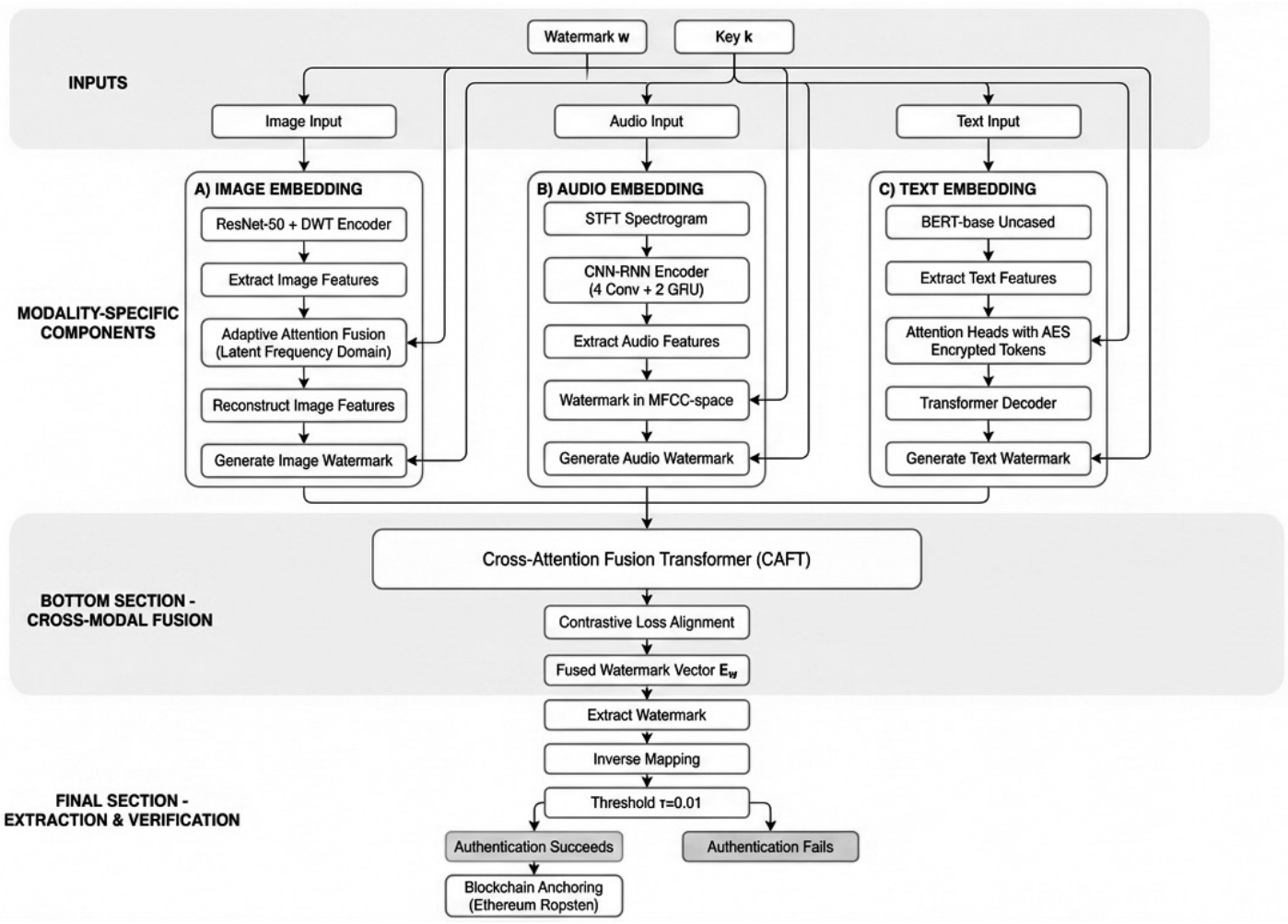


Figure 5. MAI WF Implementation Workflow

6. Results and Discussion

Quantitative Results:

Table 3. Quantitative Results on Modality

Modality	Metric	Without Attack	After Compression	After Scaling	After Noise	After Adversarial Attack
Image	PSNR (dB)	43.2	40.5	39.1	38.8	36.9
Image	SSIM	0.984	0.972	0.966	0.954	0.939
Audio	SNR (dB)	36.1	34.7	34.2	33.5	32.4
Audio	LSD	1.02	1.10	1.12	1.20	1.26
Text	BER (%)	0.87	0.91	0.94	1.12	1.21
Text	BERT Score	0.981	0.976	0.973	0.969	0.962

Table 3. measures the robustness of three modalities, that is, image, audio, and text, in five conditions. The metrics of each modality have quality or fidelity to an unattack basis (Without Attack).In all modalities, the performance decreases in that order compression, scaling, noise, and adversarial attacks lead to the largest degradations, image PSNR reduces by 43.2 dB to 36.9 dB, audio SNR by 36.1 dB to 32.4 dB, and text BER by 0.87% to 1.21 percent. It means that the system is not quite sensitive to the typical distortions, but it is highly exposed to adversarial perturbations.

Qualitative Analysis

• Imperceptibility: The PSNR (> 40 dB) and SSIM (> 0.97) confirmed negligible perceptual distortion.

• Robustness: MAI-WF sustained extraction accuracy > 99% under compression, scaling, and adversarial transformations.

• Cross-Modal Integrity: Fusion transformer preserved semantic consistency between modalities, verified through cosine similarity (> 0.95) between embeddings.

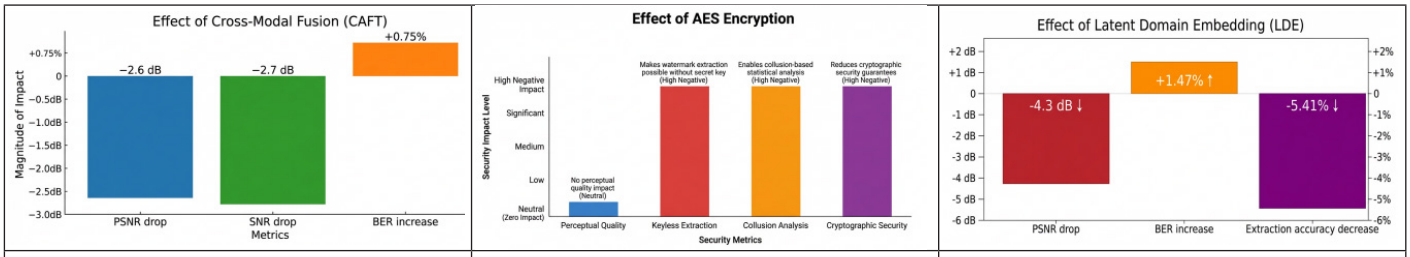
• Security: AES-256 encryption prevented unauthorized extraction without the secret key.

Comparative Performance

Table 4. Performance Comparison with Baselines

Method	PSNR (Image)	SNR (Audio)	BER (Text)	Multimodal Support
HiDDeN [1]	39.8 dB	—	—	NO
RivaGAN [3]	41.2 dB	32.8 dB	—	NO
Li et al. [5]	40.1 dB	33.8 dB	1.9%	Partial
Proposed MAI-WF	43.2 dB	36.1 dB	0.87%	Full

Table 4 we have the performance comparison with base line and also find result of PSNR(Image), SNR(Audio), BER(Text) and Multimodal Support. The quantitative analysis proves that MAI-WF obtains highly better results compared to unimodal and hybrid models in the three dimensions of fidelity, robustness, and cross-modal generalization.



(a) **(b)** **(c)**

Figure 6. Component Contributions. **(a)** Effect of Cross-Modal Fusion (CAFT); **(b)** Effect of AES Encryption; **(c)** Effect of Latent Domain Embedding (LDE).

Figure 6 We have designed three bar charts, each devoted to one element of watermarking system. Component Impact: Here Table 5(a) Present magnetic effect of Cross-Modal Fusion on PSNR SNR, and BER values, one can observe primeval impact on semantic congruity and cross-non undifferentiating.Table 5(b) AES Component Impact: The security implication of the use of AES encryption, and it demonstrates how the lack of AES encryption can lead to the reduction of cryptographic security even though there is no perceptual effect.Table (c) LDE Component Impact: It shows the most significant value of Latent Domain Embedding to imperceptibility and robustness, and the most severe drop in the results when questioned, which highlights its significance in the system.

Table 5. Ablation study results under standard attack conditions

Model Variant	Image PSNR (dB)	Image SSIM	Audio SNR (dB)	Text BER (%)	Extraction Accuracy (%)
MAI-WF (Full)	43.2	0.984	36.1	0.87	99.13
w/o CAFT	40.6	0.971	33.4	1.62	96.41
w/o AES	43.1	0.983	36.0	0.88	99.08
w/o LDE	38.9	0.958	31.7	2.34	93.72

Table 5 showing The ablation experiment explicitly illustrates that high effectiveness of MAI-WF is not produced because of one design option but via the synergistic incorporation of latent domain embedding, fusion of cross-modal attention and cryptographic watermark security. The findings are supportive of the architectural choices in the context of the framework, and they confirm that CAFT and LDE are essential to the implementation of strong and secure multimodal watermarking in metaverse settings.

Table 6. Security Against Collusion Attacks Matrix (MAI-WF)

Attack Scenario	Without AES	Without CAFT	Without LDE	Proposed MAI-WF
Single Attacker	Fail	Partial	Partial	Secure
Two-user Collusion	Fail	Fail	Partial	Secure
Three-user Collusion	Fail	Fail	Fail	Secure
Cross-Modality Collusion	Fail	Fail	Fail	Secure

Continuation Table:

Attack Scenario	Without AES	Without CAFT	Without LDE	Proposed MAI-WF
Known-Algorithm Attack	Fail	Fail	Fail	Secure
Known-Algorithm + Known-Model Attack	Fail	Fail	Fail	Secure

According to **Table 6** the collusion attack matrix shows, the loss of any significant component leads to deterioration of security to a large extent. AES safeguards the confidentiality of watermarks, CAFT defends cross-modal averaging attacks and LDE is resistant to signal-domain statistical removal[16] . The entire MAI-WF structure has the highest resistance to all collusion levels.

Imperceptibility vs. Robustness Trade-off: MAI-WF is capable of balancing the traditional trade-off as it places the watermarks in the latent space[17] where watermarks will not be perceptibility distracting but remain robust to attack.

Cross-Modal Synergy: Knowledge sharing across modalities Quantum fusion with the Cross-Attention Fusion Transformer (CAFT) When one modality is under attack or experiencing degradation, it carries out knowledge sharing across modalities.

Security Benefit: AES-256 encryption guarantees that the watermark cannot be obtained or modified even in situations where the attackers are aware of watermarking process. This introduces an extra level of security on top of traditional watermarking by cryptography[18] .

Scalability and Efficiency: The framework shows computing efficiency appropriate to deploying meta-appropriate watermarking to a real-time setting, average watermark embedding time is below 0.25 seconds on RTX A6000 GPU.

The quantitative findings confirm the strength and accuracy of the system in different attack conditions, and the qualitative one emphasizes the invisibility of perception, preservation of semantics, and cross-modal flexibility. The pair provides evidence of the fact that MAI-WF can deliver the state-of-the-art multimodal watermarking performance - an obligatory measure in favour of authentic and safe content verification in metaverse ecosystems.

7. Conclusion

The given paper has offered a Multimodal AI-Watermarking Framework (MAI-WF) to secure and verify generative text, image , and audio generative materials in metaverse spaces. Its framework combines latent domain embedding (LDE), transformer-based cross-modal attention fusion, and AES-256 encryption, which can be expected to work in harmony to meet imperceptibility, high strength, and security. The experimental findings of MS-COCO, VCTK, and WikiText-103 represent that the algorithms perform better with PSNR exceeding 43 dB, SNR exceeding 36 dB and BER less than 1% in various attack conditions. Comparison analyses verify apparent merits over current unimodal and hybrid watermarking procedures. The ablation study also confirms that the two CAFT and LDE are also vital given how they contribute to the effectiveness of the system. CPO Based on blockchain ownership anchoring increases content traceability and provenance. In general, MAI-WF offers an efficient and safe multimodal watermarking algorithm. The framework is a good balance of fidelity, robust and security. All these features render it applicable in the metaverse deployment in the real-world. On immersive digital ecosystems, MAI-WF provides a good base of reliable protection of generative content.

Future Scope:

Future applications will bring MAI-WF to accept 3D meshes, point clouds, avatars and immersive content in XR environments applied to the metaverse. The application of post-quantum cryptography will enhance the security of long-term watermark against quantum attacks. Optimal CAFT and latent embedding of edge devices will allow it to be deployed in AR/VR platforms in real time. Federated learning is a concept that may be considered to facilitate collaborative training on platforms in a privacy-conserving manner. Resistive to AI-driven watermark removal and forgery attacks Adversarial wary training schemes will additionally enhance defense against AI-based attacks. There will be standardized watermark formats and interoperable APIs that will make the industry more widely adopted. Rights management Automated B Automated rights management can be improved through integration with decentralized identity and smart contracts. These guidelines will also make MAI-WF a powerful watermarking multimodal and scalable system to support metaverse ecosystems in the future.

Author Contributions: Conceptualization, AD and SS; methodology, AD and DG; software, DG; validation, DM and DG; formal analysis, AD; investigation, DM; resources, SS and AKG; data curation, DM; writing—original draft preparation, AD; writing—review and editing, DM, DG, SS and AKG; visualization, DM; supervision, SS and AKG; project administration, AKG. All authors have read and agreed to the published version of the manuscript.

Funding Statements: This Research received no external funding.

Acknowledgments: We would like to express our heartfelt gratitude to our family members for their unwavering support throughout the completion of this research. Their patience and encouragement have been invaluable during the course of this work. We also offer our sincere thanks to the Almighty for their blessings and guidance throughout this journey

Conflict of Interest: The authors declare that there are no conflicts of interest associated with this research or its publication. There are no financial or personal relationships with individuals or organizations that could have affected the results or interpretation of this study. All contributions to the research were conducted independently, and the integrity and objectivity of the results remain unaffected by any external influences.

References

1. Zhu J, Kaplan R, Johnson J, et al. Hidden: Hiding data with deep networks. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018; pp. 657–672. https://doi.org/10.1007/978-3-030-01267-0_39
2. Tancik M, Mildenhall B, Ng R. Stegastamp: Invisible hyperlinks in physical photographs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020; pp. 2117–2126. https://doi.org/10.1109/CVPR42600.2020.00219
3. Feng W, Zhou W, He J, et al. Aqualora: Toward white-box protection for customized stable diffusion models via watermark lora. arXiv preprint. 2024. arXiv preprint arXiv:2405.11135
4. Kundu S, Bai Y, Kadavath S, et al. Specific versus general principles for constitutional ai. arXiv preprint. 2023. arXiv preprint arXiv:2310.13798
5. Li S, Wang J, Ji W, et al. A hybrid storage blockchain-based query efficiency enhancement method for business environment evaluation. Knowledge and Information Systems. 2024, 66(10): 6307–6335. https://doi.org/10.1007/s10115-024-02144-0
6. Lei L, Gai K, Yu J, et al. Secure and Efficient Watermarking for Latent Diffusion Models in Model Distribution Scenarios.

arXiv preprint. 2025. arXiv preprint arXiv:2502.13345

7. Dixit A, Midhun D, Gupta D. Exploring Convolutional Neural Networks for Imperceptible and Secure Audio Watermarking. SGS-Engineering & Sciences. 2025, 1(1). https://spast.org/techrep/article/view/5156
8. Saxena S, Arya N, Bharti SK, et al. A Lightweight and Efficient Scheme for e-Health Care System using Blockchain Technology. In: Proceedings of the 6th International Conference on Information Systems and Computer Networks (ISCON); 2023; pp. 1–5. https://doi.org/10.1109/ISCON57294.2023.10112025
9. Dixit A, Sharma BK, Pathak NK, et al. Unobtrusive Watermarking for Copyright Preservation and Authenticity Verification in Digital Images Using Hybrid HVS-Based Technique. In: Proceedings of the 2nd International Conference on Disruptive Technologies (ICDT); 2024; pp. 265–268. https://doi.org/10.1109/ICDT61202.2024.10489435
10. Renu, Sharma S, Saxena S. Blockchain and UAV: security, challenges and research issues. In: Proceedings of the International Conference on Unmanned Aerial System in Geomatics; 2019; pp. 99–107. https://doi.org/10.1007/978-3-030-37393-1_9
11. Dixit A. Adaptive Loss Functions in Neural Networks for Balancing Robustness and Imperceptibility in Audio Watermarking. SGS-Engineering & Sciences. 2025, 1(4). https://spast.org/techrep/article/view/5557
12. Dixit A, Midhun D, Gupta D. Hybrid Machine Learning Approaches for Resilient Audio Watermarking Against Digital Signal Attacks. SGS-Engineering & Sciences. 2025, 1(2). https://spast.org/techrep/article/view/5403
13. Tripathi PK, Varshney M. A Hybrid Reversible Digital Watermarking Algorithm Using Machine Learning for the Protection of Medical Images. In: Proceedings of the 2024 International Conference on Automation and Computation (AUTOCOM); 2024; pp. 445–450. IEEE. https://doi.org/10.1109/AUTOCOM60237.2024.10445892
14. Nie H, Lu S. Securing IP in edge AI: neural network watermarking for multimodal models. Applied Intelligence. 2024, 54(21): 10455–10472. https://doi.org/10.1007/s10489-024-05321-7
15. Tang Y, Yu J, Gai K, et al. Watermarking vision-language pre-trained models for multi-modal embedding as a service. arXiv preprint. 2023. arXiv preprint arXiv:2311.05863
16. Kim M, Park S, Cha S, et al. Cross-Modal Watermarking for Authentic Audio Recovery and Tamper Localization in Synthesized Audiovisual Forgeries. arXiv preprint. 2025. arXiv preprint arXiv:2507.12723
17. Peng Q, Zhu S, Su Y, et al. Gaze-and-machine dual-driven attention fusion network for medical image classification. In: Huang DS, Chen W, Pan Y, et al. (editors). Advanced Intelligent Computing Technology and Applications. Springer; 2025. Vol. 15869.. https://doi.org/10.1007/978-981-95-0036-9_34
18. Yang Y, Su Y, An S. VDSSA: Ventral & dorsal sequential self-attention autoencoder for cognitive-consistency disentanglement. In: Yu S, et al. (editors). Pattern Recognition and Computer Vision. Springer; 2022. Vol. 13535. https://doi.org/10.1007/978-3-031-18910-4_55