FORENSICZIP: PRESERVING FORGERY EVIDENCE UNDER AGGRESSIVE VISUAL TOKEN COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) enable interpretable multimedia forensics by generating textual rationales for forgery detection. However, processing dense visual inputs incurs high computational costs, particularly for high-resolution images and videos. Visual token pruning is a practical acceleration strategy, yet general-purpose methods often prioritize semantic content over subtle manipulation evidence. To address this mismatch, we introduce ForensicZip, a training-free framework that selects visual tokens according to forensic evidence. For videos, augmented Birth-Death Optimal Transport estimates novelty across adjacent frames; for single images, spatial embedding deviation provides the novelty score. A high-frequency prior then guides physical Top-K selection under a fixed token budget. At 10% visual-token retention, ForensicZip achieves 97.74% accuracy on FakeClue; across eight backbone–dataset settings, mean end-to-end latency decreases by 2.97× and mean FLOPs by 85.4%. Code is available at https://anonymous.4open.science/r/iclr2027-forensiczip-review-8168.

1 INTRODUCTION

Recent advances in Multimodal Large Language Models (MLLMs) Liu et al. (2024a;b); Lin et al. (2023); Wang et al. (2024); Bai et al. (2023); Chu et al. (2023); Wen et al. (2025c) have led to remarkable progress in building generalist systems for diverse visual understanding and reasoning tasks. In multimedia forensics, MLLMs are increasingly adopted as an explainable pathway for forgery detection. With large-scale visual instruction tuning, forensic MLLMs Xu et al. (2024); Zhou et al. (2025b); Tan et al. (2025b); Huang et al. (2025); Guo et al. (2025b) can not only predict authenticity but also produce fine-grained textual rationales, such as pointing out abnormal clues, blending boundaries, or inconsistent reflections Wen et al. (2025a); Chen et al. (2024c). High-resolution images and long video streams help expose subtle manipulation traces, but produce many visual tokens. The resulting sequence length makes prefill computationally expensive in both attention and feed-forward layers.

A growing line of work shows that visual tokens in Vision-Language Models (VLMs) and MLLMs are highly redundant, motivating training-free token pruning and merging as a practical inference acceleration strategy Han et al. (2024); Yu et al. (2026). Representative methods perform early-layer pruning Chen et al. (2024b); Wen et al. (2025b), attention- or sparsity-guided token selection Zhang et al. (2025b); Shang et al. (2024), or adaptive token reduction Shang et al. (2025); Sun et al. (2025). Most of these approaches rank tokens by semantic salience (e.g., cross-modal attention, vision-language similarity, or macro-diversity Alvar et al. (2025); Li et al. (2025)) and achieve substantial compute savings with minor performance drops on standard benchmarks such as captioning Fu et al. (2023); Yue et al. (2024) and VQA Zhou et al. (2025a); Yu et al. (2019).

Forensic detection requires a different token priority from general visual understanding. Key manipulation cues (such as high-frequency noise patterns, subtle blending seams, or transient temporal jitters) need not coincide with object-level semantic saliency. Token rankings optimized for semantic content may therefore assign low priority to regions containing manipulation evidence Zhong et al. (2024); Wen et al. (2025b); Yu et al. (2026). Fig. 1(a) shows how equal token budgets retain different image regions. On 2,000 randomly sampled SID-Set images, Fig. 1(b) measures the fraction of mask-labeled forgery patches retained at each budget. These observations motivate three questions:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

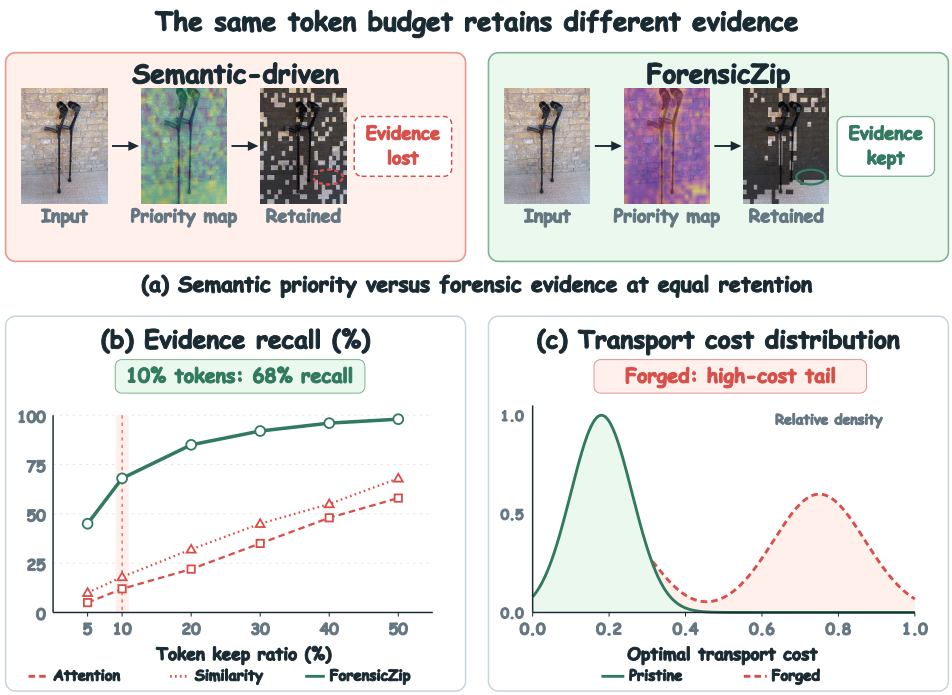

Figure 1: **Semantic priority can miss local forgery cues.** (a) The priority maps and retained patches of a semantic selector and ForensicZip at the same token budget; the contours mark a local forgery cue. (b) On 2,000 randomly sampled SID-Set images, mask labels define forgery patches, and recall is the fraction retained at each token ratio (68% for ForensicZip at 10%). (c) Forged videos exhibit a high-cost tail in the temporal OT distribution, consistent with transient unmatched features; the density curves are shown on a relative scale.

(1) Why do existing pruning methods suffer task-dependent degradation on forgery detection? (2) How do generative artifacts manifest dynamically in the visual token space, especially over time? (3) How can we aggressively compress visual tokens while preserving subtle, non-salient forensic evidence?

In this work, we observe that generated videos can exhibit local temporal discontinuities despite convincing individual frames. In visual token space, these discontinuities appear as unmatched local textures and structures (Fig. 1(c)). Building on this insight, we propose ForensicZip, a training-free token compression framework that prioritizes forgery evidence under a fixed token budget. ForensicZip consists of two core components: First, Transport Novelty Estimation (TNE) formulates evolution across adjacent video frames as an augmented entropic Optimal Transport problem. A dummy node captures unmatched mass; single-image inputs instead use spatial embedding deviation to estimate novelty. Second, Forensic Scoring combines the resulting novelty scores with a frequency-domain prior. Finally, physical Top-K selection shortens the visual sequence passed to the language model, reducing inference cost. Our contributions are summarized as follows:

- **Task-specific Analysis:** We identify a structural mismatch between semantic-driven token pruning and the forgery-driven nature of forgery detection, demonstrating that preserving global anchors and transient artifacts is crucial under aggressive token budgets.
- **ForensicZip Framework:** We propose a training-free token sparsification framework with spatial novelty for images, augmented Birth-Death Optimal Transport for videos, and frequency-aware scoring for evidence preservation.
- **Performance Validation:** ForensicZip integrates into existing forensic MLLMs via physical pruning. Across eight backbone–dataset settings at 10% visual-token retention, mean end-to-end latency decreases by 2.97× and mean FLOPs by 85.4% while detection accuracy remains strong.

2 Related Work

Synthetic Image Detection via MLLMs. Deepfake and synthetic image detection have shifted from binary classification Wang et al. (2020); Chollet (2017); Wang et al. (2019); Shao et al. (2022); Heo et al. (2021) to interpretable forensics using Multimodal Large Language Models (MLLMs). Recent methods bridge visual anomalies with natural-language rationales via common-sense Zhang et al. (2024) or pattern-aware reasoning Tan et al. (2025b), and language-guided localization Guo et al. (2025a). To enhance transparency, frameworks like FakeShield Xu et al. (2024), X2-DFD Chen et al. (2024c), and Spot the Fake Wen et al. (2025a) generate fine-grained explanations paired with spatial masks. This paradigm extends to diverse domains, including face forensics Guo et al. (2025b), diffusion videos Song et al. (2024), and social media (e.g., SIDA Huang et al. (2025)), culminating in comprehensive forensic assistants like AIGI-Holmes Zhou et al. (2025b) and LEGION Kang et al. (2025). Despite these advances, these methods typically assume a sufficient computational budget to process dense visual tokens, neglecting the severe prefill latency bottleneck in high-resolution or long-video scenarios.

Visual Token Compression for Efficient MLLM Inference. Visual token reduction mitigates redundancy and accelerates MLLM inference. One prominent line of research relies on semantic importance to rank tokens via early-layer Chen et al. (2024b;a) or text-to-vision attention Zhang et al. (2025b). To improve information density and coverage, recent approaches employ clustering Sun et al. (2025); Zhong et al. (2024) or diversity-based selection Alvar et al. (2025); Li et al. (2025) to prevent token over-concentration. For video and high-resolution scenarios, frameworks explicitly exploit spatio-temporal dynamics Wen et al. (2025b) or offer training-free plug-and-play pipelines Han et al. (2024); Liu et al. (2025) to reduce redundancy. General-purpose pruning methods commonly prioritize macro-level visual understanding. Forensic cues such as resampling patterns and temporal jitters can instead occur in regions assigned low semantic saliency by these methods. Furthermore, recent analyses Fu et al. (2025); Zhang et al. (2025a) reveal that semantic pruning under cross-modal misalignment can destroy the global spatial reference required for precise forgery localization. This task mismatch motivates ForensicZip, which scores spatial or temporal token novelty and physically reduces sequence length while prioritizing forensic evidence.

3 ForensicZip

3.1 Preliminary

Forensic VLM pipeline. A typical forensic VLM consists of a vision encoder, a multimodal projector $g : \mathbb{R}^{D_v} \to \mathbb{R}^{D_p}$, and a language model, where D_v is the vision-encoder hidden dimension and D_p is the language embedding dimension. Given a video with T frames $\{\mathbf{I}_t\}_{t=1}^{T}$, the vision encoder produces, for each frame t, one global token and N patch tokens $\mathbf{V}_t = \{\mathbf{v}_{t,1}, \ldots, \mathbf{v}_{t,N}\} \in \mathbb{R}^{N \times D_v}$, where $\mathbf{v}_{t,i} \in \mathbb{R}^{D_v}$ is the i-th patch token.

Sequence-length complexity. For a transformer with L_{layer} layers, sequence length n, hidden size d, and FFN intermediate size m, the FLOPs are approximated by

$$\text{FLOPs} = L_{\text{layer}}\big(4nd^2 + 2n^2d + 2ndm\big). \tag{1}$$

The quadratic self-attention term $2n^2d$ makes cost sensitive to n. In multimodal prompting the total length decomposes as

$$n = n_{\text{sys}} + n_{\text{vis}} + n_{\text{txt}}, \qquad n_{\text{vis}} = T\,(N+1), \tag{2}$$

where n_{sys}, n_{txt}, and n_{vis} denote system-prompt, user-text, and visual-token lengths, respectively. Because n_{vis} dominates in high-resolution and video settings, reducing visual tokens directly lowers prefill cost and memory footprint Wen et al. (2025b); Yu et al. (2026); Bolya et al. (2022); Alvar et al. (2025).

3.2 The Forensic-Semantic Discrepancy

Before presenting our framework, we investigate how common token reduction paradigms behave in the forensic domain. Through systematic analysis on the FakeVLM benchmark, we identify two misalignments that together motivate a transport-based scoring mechanism with explicit slack.

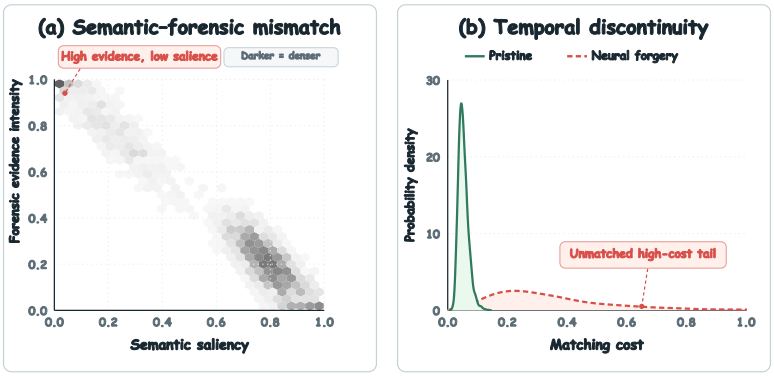

Figure 2: **Empirical motivation for ForensicZip. (a) Semantic-forensic mismatch:** Density plot of semantic saliency (cross-modal attention) and forensic evidence (forgery-mask overlap) on FakeClue/LOKI. **(b) Temporal discontinuity:** KDE of adjacent-frame matching costs for pristine and forged videos.

Obs. 1: Semantic saliency and forensic evidence can diverge. Existing pruning strategies assume that semantically salient regions carry the most task-critical information. As shown in Fig. 1(a), cross-modal attention and spatial overlap with ground-truth forgery masks can differ. MLLM attention heads converge on high-level object centres, whereas manipulation traces such as blending inconsistencies, resampling grids, and compression artefacts reside predominantly in low-saliency backgrounds and object edges. Purely semantic reduction therefore acts as a low-pass filter on forensic information, preserving convincing content while removing the subtle evidence required for detection.

Obs. 2: Temporal discontinuity as Birth–Death events. We further examine token dynamics in the projected feature space across adjacent frames. In pristine video, tokens evolve smoothly: under uniform marginals the optimal transport plan between consecutive frames concentrates mass along the cost-matrix diagonal, reflecting continuous spatial correspondence. Generative models violate this continuity: certain features appear without a plausible source (Birth) or vanish without a matching target (Death). Fig. 2(b) shows a high-cost tail for forged videos, consistent with unmatched local features. This observation motivates an augmented transport formulation that routes unmatchable mass through a dedicated slack node.

3.3 Methodology

Both observations motivate a shared selection principle: tokens with forensic evidence need not have the highest semantic saliency. Spatial frequency responses provide one cue to local manipulation, while unmatched feature mass across adjacent frames provides another cue for video inputs. We realise this principle through two complementary stages (Fig. 3). *Transport Novelty Estimation* (TNE) estimates unmatched mass across adjacent video frames with augmented optimal transport and a slack node; single images use spatial embedding deviation instead. *Forensic Scoring* (FS) modulates the resulting novelty score with a high-frequency prior, amplifying tokens supported by both cues.

3.3.1 Transport Novelty Estimation (TNE)

Forgery detection in the temporal domain reduces to conservation violation detection: natural video preserves feature-level mass across frames, whereas generative artefacts create or destroy local textures abruptly. A naïve cosine-similarity approach treats each token independently, so anomalous residuals are absorbed by the single best match. Optimal transport improves on this by finding a global assignment under marginal constraints, but standard balanced OT enforces strict mass conservation, forcing every target token onto a real source and smearing the anomaly across many small residuals. Adding a dummy node closes this gap: tokens that genuinely lack a predecessor can route their mass through the dummy at a fixed penalty, surfacing as concentrated high-cost events.

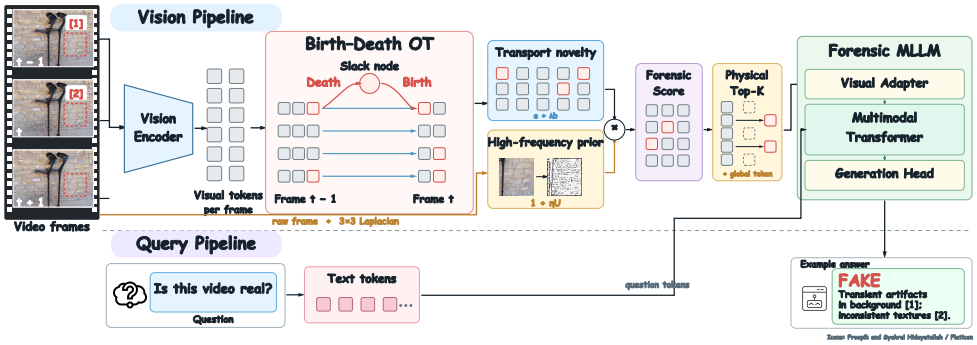

Figure 3: **Overall framework of ForensicZip.** ForensicZip measures transport novelty across adjacent video frames and combines it with a high-frequency spatial prior before physical Top-K selection. The video route is shown; single images and the first video frame use spatial embedding deviation in place of transport novelty.

Birth–Death entropic optimal transport. Each patch token $\mathbf{v}_{t,i} \in \mathbb{R}^{D_v}$ $(i = 1, \ldots, N)$ in frame t is projected and ℓ_2-normalised:

$$\hat{\mathbf{z}}_{t,i} = \frac{g(\mathbf{v}_{t,i})}{\|g(\mathbf{v}_{t,i})\|_2 + \epsilon_{\text{norm}}}, \qquad \hat{\mathbf{z}}_{t,i} \in \mathbb{R}^{D_p}, \tag{3}$$

where ϵ_{norm} is a small constant for numerical stability. Normalisation maps all embeddings onto the unit hypersphere, naturally yielding cosine distance as the ground cost.

For consecutive frames $(t{-}1,\ t)$ with $T \geq 2$, we build a pairwise cost matrix $\mathbf{C}^{(t)} \in \mathbb{R}^{N \times N}$ whose entry $C_{ij}^{(t)} = 1 - \hat{\mathbf{z}}_{t-1,\,i}^{\top}\, \hat{\mathbf{z}}_{t,\,j}$ measures the cosine distance between the i-th source token in frame $t{-}1$ and the j-th target token in frame t. To flag tokens that appear without a source (*Birth*) or vanish without a target (*Death*), we extend this matrix with a dummy row and column:

$$\bar{\mathbf{C}}^{(t)} = \begin{bmatrix} \mathbf{C}^{(t)} & c_{\text{death}}\, \mathbf{1}_N \\ c_{\text{birth}}\, \mathbf{1}_N^{\top} & 0 \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}, \tag{4}$$

where $\mathbf{1}_N \in \mathbb{R}^N$ is the all-ones vector. The scalar penalties $c_{\text{birth}} \geq 0$ and $c_{\text{death}} \geq 0$ set the cost of routing mass through the dummy node: whenever the cheapest real assignment of a target token exceeds c_{birth}, the solver preferentially explains it as a newly appeared artefact. Instead of forcing a spurious pairing that dilutes the anomaly, the dummy node absorbs the unmatchable mass and concentrates the cost into a single interpretable entry.

We impose marginals $\mathbf{a} = [\frac{1}{N}, \ldots, \frac{1}{N}, 1]^{\top} \in \mathbb{R}^{N+1}$, where the first N entries assign uniform mass to real tokens and the last entry provides unit slack to absorb all unmatchable mass if necessary. The entropic transport plan $\bar{\mathbf{P}}^{(t)} \in \mathbb{R}^{(N+1)\times(N+1)}$ is obtained via I_{sk} Sinkhorn iterations (set to 20 in all experiments):

$$\bar{\mathbf{P}}^{(t)} = \underset{\substack{\mathbf{P}\geq 0,\\ \mathbf{P}\,\mathbf{1}_{N+1}=\mathbf{a},\ \mathbf{P}^{\top}\mathbf{1}_{N+1}=\mathbf{a}}}{\arg\min} \left\langle \mathbf{P},\, \bar{\mathbf{C}}^{(t)} \right\rangle_F + \varepsilon_{\text{ot}} \sum_{i=1}^{N+1} \sum_{j=1}^{N+1} P_{ij}\big(\log P_{ij} - 1\big), \tag{5}$$

where $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product, $\varepsilon_{\text{ot}} > 0$ is the entropic regularisation strength, and $\mathbf{1}_{N+1} \in \mathbb{R}^{N+1}$ is the $(N{+}1)$-dimensional all-ones vector used in the marginal constraints.

Per-token score extraction. From the solved plan we extract two scores for each target token j $(j = 1, \ldots, N)$ in frame t, spanning the full range of temporal anomaly from gradual drift to abrupt appearance. The *transport cost* aggregates the weighted matching cost from all N real source tokens in frame $t{-}1$, excluding the dummy row:

$$e_j^{(t)} = \sum_{i=1}^{N} \bar{P}_{ij}^{(t)}\, C_{ij}^{(t)}\,. \tag{6}$$

A high $e_j^{(t)}$ indicates that token j deviates from every predecessor yet was still partially matched, capturing distributed anomalies. The *Birth evidence*

$$b_j^{(t)} = \bar{P}_{N+1,\,j}^{(t)} \tag{7}$$

records the mass routed from the dummy source, directly quantifying the absence of a plausible origin in frame $t-1$ and capturing abrupt appearance that admits no real correspondence.

For single-image inputs ($T = 1$), we use spatial novelty: we compute the frame-level mean embedding $\boldsymbol{\mu} = \frac{1}{N}\sum_{j=1}^{N} \hat{\mathbf{z}}_{1,j}$ and define per-token novelty as $q_j^{(1)} = \|\hat{\mathbf{z}}_{1,j} - \boldsymbol{\mu}\|_2$. The same spatial modulation gives $s_j^{(1)} = q_j^{(1)}(1 + \eta_{\text{forensic}} U_j^{(1)})$.

Transport cost measures feature-level displacement but is agnostic to its cause: legitimate large-scale motion such as camera panning produces elevated cost indistinguishable from a generative artefact on the basis of temporal signal alone. Resolving this ambiguity requires a signal from a different physical domain.

3.3.2 Forensic Scoring and Selection (FS)

Generative artefacts and natural motion differ not in temporal displacement magnitude but in spatial frequency signature. Natural motion displaces existing textures without altering their local spectral statistics, whereas generation pipelines introduce characteristic fingerprints: aliasing from resampling, spectral gaps from upsampling networks, and gradient discontinuities at blending boundaries. This frequency-domain signal also addresses the inverse correlation identified in Obs. 1: forensic traces in semantically flat regions still carry strong high-frequency responses and are therefore preserved.

High-frequency prior integration. We apply a fixed 3×3 Laplacian operator $\mathcal{L}(\cdot)$ to the input frame $\mathbf{I}_t$ and average-pool the response to the N-element patch grid, yielding a spatial proxy $\mathbf{U}^{(t)} \in [0,1]^N$ where $U_j^{(t)}$ is the normalised response at the j-th patch position. The forensic score for token j in frame t is the product of a temporal anomaly term and a spatial modulation term:

$$s_j^{(t)} = \underbrace{\left(e_j^{(t)} + \lambda_{\text{birth}}\, b_j^{(t)}\right)}_{\text{temporal anomaly}} \cdot \underbrace{\left(1 + \eta_{\text{forensic}}\, U_j^{(t)}\right)}_{\text{spatial modulation}}, \tag{8}$$

where $\lambda_{\text{birth}} \geq 0$ balances Birth evidence against transport cost, and $\eta_{\text{forensic}} \geq 0$ controls the amplification from high-frequency features. Temporal novelty provides the base score, and spatial evidence increases the rank of tokens with high-frequency responses. A token with zero temporal novelty receives zero score, while a high-frequency response amplifies existing novelty. For single images, Eq. 8 uses the spatial novelty $q_j^{(1)}$ defined above in place of the temporal term.

Token selection. Given a retention ratio $\rho \in (0, 1]$, we select the top $K = \lfloor \rho N \rfloor$ tokens from $\{s_j^{(t)}\}_{j=1}^{N}$ for each frame t, where K denotes the retained patch count per frame. The global token is always kept. The compressed sequence of length $T(K+1)$ replaces the original $T(N+1)$ tokens and is passed directly to the language model.

3.3.3 Computational Complexity

Each transformer layer incurs $\mathcal{O}(n^2 D_m + n D_m D_{\text{ff}})$ cost, where n is the token count, D_m is the language-model hidden dimension, and D_{ff} is the FFN intermediate size. Reducing the per-frame visual sequence from $N+1$ to $K+1$ tokens across all L_{layer} layers reduces the visual-to-visual attention term by a factor of $1 - (K+1)^2/(N+1)^2$ per layer. For video inputs, the entropic OT solver operates on an $(N+1) \times (N+1)$ cost matrix:

$$\text{Cost}_{\text{OT}} = \mathcal{O}\big((T-1)\, I_{\text{sk}}\, (N+1)^2\big), \tag{9}$$

where I_{sk} is the number of Sinkhorn iterations and $T-1$ is the number of consecutive frame pairs. This cost is incurred before the LLM forward pass and scales linearly with frame count; the spatial scoring cost is analysed in the supplementary material. **A detailed theoretical analysis is provided in the supplementary material.**

Table 1: **Quantitative Performance of ForensicZip on FakeVLM and FakeShield Backbones.** "Vanilla" uses all visual tokens. Pruning baselines follow their official implementations at matched target retention ratios. FLOPs and end-to-end latency are arithmetic means over eight backbone–dataset settings; memory is peak GPU usage. Speedup divides Vanilla mean latency by method mean latency. Red marks low accuracy.

Method	Backbone: FakeVLM				Backbone: FakeShield				Avg. FLOPs(T)	Avg. Lat.(ms)	GPU Mem.(GB)	Speedup
	FakeClue	LOKI	DD-VQA	DMimage	PhotoShop	DeepFake	AIGC	DFFD				
Upper Bound (100% Tokens)												
Vanilla	98.60	84.30	93.20	94.00	94.43	98.00	93.00	98.00	3.102	4520.2	26.7	1.00×
Retain 50% Tokens (↓ 50% visual tokens)												
FastV (ECCV'24)	95.23	78.15	89.52	88.24	90.15	93.42	88.54	93.22	2.150	3228.6	26.6	1.40×
SparseVLM (ICML'25)	96.14	79.53	90.45	89.66	91.24	94.15	89.23	94.44	2.155	3180.4	26.5	1.42×
LLaVA-Scissor (NeurIPS'25)	97.22	81.84	91.81	91.13	92.83	95.84	91.42	95.85	2.140	3110.5	26.7	1.45×
VisionTrim (ICLR'26)	97.87	82.25	92.92	92.34	93.54	96.92	92.15	96.52	2.135	3090.2	26.4	1.46×
ForensicZip (Ours)	**98.50**	**84.35**	**93.25**	**94.05**	**94.48**	**98.05**	**93.05**	**98.04**	**1.852**	**2756.4**	**22.4**	**1.64×**
Retain 25% Tokens (↓ 75% visual tokens)												
FastV (ECCV'24)	82.34	65.55	76.42	75.13	78.25	81.52	75.43	80.84	1.420	2510.5	26.5	1.80×
SparseVLM (ICML'25)	85.63	68.22	79.65	78.54	81.42	83.15	78.56	83.22	1.415	2480.8	26.4	1.82×
LLaVA-Scissor (NeurIPS'25)	91.45	75.53	86.24	85.42	86.54	90.45	85.13	90.92	1.390	2420.2	26.6	1.86×
VisionTrim (ICLR'26)	94.83	79.15	89.56	89.24	90.25	94.14	89.55	94.83	1.385	2390.6	26.3	1.89×
ForensicZip (Ours)	**98.46**	**83.93**	**93.12**	**93.88**	**94.32**	**97.88**	**92.85**	**97.82**	**0.985**	**2050.5**	**19.8**	**2.20×**
Retain 10% Tokens (↓ 90% visual tokens)												
FastV (ECCV'24)	60.54	45.13	56.42	55.25	55.93	58.54	53.22	57.83	0.850	1950.2	26.4	2.31×
SparseVLM (ICML'25)	65.82	48.45	60.14	59.53	60.25	63.83	58.14	61.52	0.845	1920.5	26.3	2.35×
LLaVA-Scissor (NeurIPS'25)	78.25	62.53	72.56	71.24	72.54	76.12	71.43	75.85	0.820	1880.8	26.5	2.40×
VisionTrim (ICLR'26)	84.53	70.84	80.25	78.56	79.14	82.55	76.82	82.24	0.815	1850.4	26.1	2.44×
ForensicZip (Ours)	**97.74**	**82.15**	**90.58**	**91.23**	**92.15**	**95.88**	**90.85**	**95.62**	**0.452**	**1520.3**	**18.5**	**2.97×**

Table 2: **Efficiency comparisons on FakeClue at 25% token retention.** Generation latency is reported for the LLM and full model; total latency covers the benchmark. Peak memory is maximum VRAM usage, and throughput is samples per second.

Methods	LLM Generation↓ Latency (s)	Model Generation↓ Latency (s)	Total↓ Latency (min:sec)	GPU Peak↓ Memory (GB)	Throughput↑ (samples/s)	Performance↑ Acc (%)
FakeVLM-7B (Vanilla)	1472.5	1478.6	67:59	26.7	1.23	98.60
Retention Ratio = 25%						
FastV [ECCV'24]	885.2 (↓ 39.9%)	892.4 (↓ 39.6%)	41:30 (↓ 38.9%)	26.5 (↓ 0.7%)	2.00 (↑ 1.63×)	82.34 (↓ 16.2)
SparseVLM [ICML'25]	870.5 (↓ 40.9%)	878.2 (↓ 40.6%)	40:55 (↓ 39.8%)	26.6 (↓ 0.4%)	2.03 (↑ 1.65×)	85.63 (↓ 12.9)
LLaVA-Scissor [NeurIPS'25]	852.1 (↓ 42.1%)	860.5 (↓ 41.8%)	40:10 (↓ 40.9%)	26.5 (↓ 0.7%)	2.07 (↑ 1.68×)	91.45 (↓ 7.1)
VisionTrim [ICLR'26]	835.4 (↓ 43.3%)	842.1 (↓ 43.0%)	39:45 (↓ 41.5%)	26.4 (↓ 1.1%)	2.10 (↑ 1.71×)	94.83 (↓ 3.7)
ForensicZip (Ours)	**660.5** (↓ 55.1%)	**666.8** (↓ 54.9%)	**30:55** (↓ 54.5%)	**19.8** (↓ 25.8%)	**2.70** (↑ 2.20×)	**98.46** (↓ 0.1)

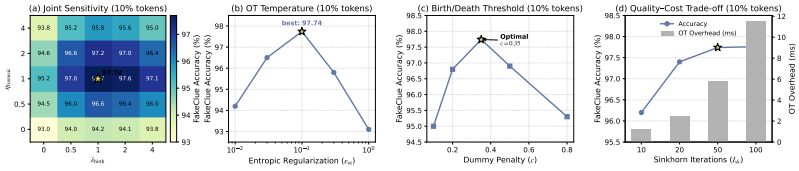

Figure 4: **Ablation Study.** Impact of (a) joint weights λ/η, (b) regularization ε_{ot}, (c) penalty c, and (d) iterations on FakeClue accuracy.

4 EXPERIMENTS

Benchmarks. To evaluate ForensicZip across diverse forgery types, we utilize three dataset categories: (1) **Universal & AIGC Detection:** FakeVLM (FakeClue, LOKI Ye et al. (2025)), DMimage Corvi et al. (2023), FakeShield-AIGC Xu et al. (2024), and SIDA Huang et al. (2024) for synthetic artifacts and social media imagery; (2) **DeepFake Forensics:** DD-VQA Zhang et al. (2024) for artifact explanation, alongside FakeShield-DeepFake and DFFD Dang et al. (2020) for facial manipulation; and (3) **Traditional Manipulation:** FakeShield-PhotoShop (aggregating CASIA Dong et al. (2013), etc.) for splicing and copy-move forgeries, while POPE assesses general visual reasoning under compression.

Table 3: **Comprehensive efficiency and performance results on the POPE benchmark.** Comparison of generation latency, end-to-end time, peak GPU memory, throughput, and accuracy on a single NVIDIA A100 GPU.

Backbone	Method	#Tokens	LLM Gen. Lat. (s) ↓	Total Lat. (min:sec) ↓	Peak Mem. (GB) ↓	Throughput (samples/s) ↑	FLOPs (T) ↓	Acc. ↑
LLaVA-1.5-7B	Vanilla	576	618.0	21:43 (1.00×)	18.5	0.64	3.8	85.9
	+ SparseVLM (ICML'25)	64	512.4	18:24 (1.18×)	20.2	0.76	1.3	75.1
	+ VisionTrim (ICLR'26)	64	448.5	16:10 (1.34×)	19.1	0.86	1.0	84.5
	+ ForensicZip (Ours)	**64**	**412.0**	**14:52 (1.46×)**	**16.8**	**0.94**	**0.8**	**86.1**
LLaVA-NeXT-7B	Vanilla	2880	1008.4	38:04 (1.00×)	46.2	0.35	12.6	86.3
	+ SparseVLM (ICML'25)	320	846.2	31:43 (1.20×)	48.5	0.42	2.5	78.5
	+ VisionTrim (ICLR'26)	320	625.4	21:50 (1.74×)	42.0	0.61	1.8	83.5
	+ ForensicZip (Ours)	**320**	**520.8**	**18:28 (2.06×)**	**36.5**	**0.72**	**1.2**	**85.8**

Table 4: **Ablation study of ForensicZip components.** Results use SID-Set, SIDA-7B, and 10% token retention. **Base** uses FastV semantic attention ranking with physical Top-K pruning and the same token budget; the global token is retained.

Variant	Components			Real Images		Tampered Images		Overall Performance	
	TNE	FS	HF	Acc	F1	Acc	F1	Acc	F1
(a) Base				55.23	52.45	54.23	51.56	57.20	54.95
(b) + TNE	✓			80.10	78.60	82.20	80.30	83.40	82.10
(c) + FS	✓	✓		85.60	87.40	88.20	87.90	90.10	89.90
(d) **ForensicZip**	✓	✓	✓	**88.15**	**90.10**	**90.35**	**90.15**	**92.50**	**92.52**

Table 5: **Effect of Transport Formulations in TNE.** SID-Set with SIDA-7B at 10% retention; dummy nodes model unmatchable tokens.

Alignment Strategy	Synthetic		Overall		Rel. Gain
	Acc	F1	Acc	F1	
Hard Assignment	86.40	86.00	79.54	79.10	—
Balanced OT	92.90	92.70	86.10	85.80	+8.25%
Only-Birth Penalty	95.80	95.60	89.20	88.90	+12.15%
Birth-Death OT	**97.95**	**97.85**	**92.50**	**92.52**	**+16.31%**

Table 6: **Effect of Spatial Operators in FS.** FakeClue with FakeVLM-7B at 10% retention; Laplacian ($\mathcal{L}$) provides an isotropic high-frequency prior.

Spatial Operator $\mathcal{L}(\cdot)$	Accuracy (%)	Rel. Gain
None ($\eta_{\text{forensic}} = 0$)	93.45	—
Patch Variance (RGB)	94.82	+1.47%
Sobel (Directional)	96.15	+2.89%
Laplacian (Isotropic, Ours)	**97.74**	**+4.59%**

Baselines. We compare ForensicZip against state-of-the-art training-free token pruning methods and full-token expert models: **1) Training-free Token Pruning Methods:** We benchmark against leading inference acceleration strategies, including **FastV** Chen et al. (2024b), **SparseVLM** Zhang et al. (2025b), and recent SOTAs such as **VisionTrim** Yu et al. (2026) and **LLaVA-Scissor** Sun et al. (2025). **2) Forensic Expert Models (Vanilla):** We report the full-token performance of forensic MLLMs, SIDA Huang et al. (2024), FakeVLM Wen et al. (2025a) and FakeShield Xu et al. (2024). The main comparison uses the official implementations of the pruning baselines at matched target token budgets; the supplementary material reports additional simple token-reduction controls.

Implementation Details. We adopt LLaVA-OneVision-7B Li et al. (2024), SIDA-7B/13B Huang et al. (2024), FakeVLM Wen et al. (2025a) and FakeShield Xu et al. (2024) as our primary backbones. The Transport Novelty Estimation (TNE) utilizes entropic regularization $\varepsilon = 0.1$, with birth and death costs set to $c_{\text{birth}} = c_{\text{death}} = 0.35$. All experiments are conducted on NVIDIA A100 (80GB) GPUs.

Performance on Forensic Benchmarks. Table 1 compares methods at matched visual-token budgets across diverse forgery types. At 10% retention, ForensicZip maintains 97.74% accuracy on FakeClue. Across the eight settings, mean end-to-end latency decreases from 4520.2 to 1520.3 ms and mean FLOPs from 3.102 to 0.452 T relative to the full-token model.

Efficiency and Robustness Analysis. We assess efficiency through deployment latency, memory, and throughput. ForensicZip physically removes selected tokens before the language model. On FakeClue at 25% retention, peak GPU memory decreases from 26.7 to 19.8 GB and throughput rises

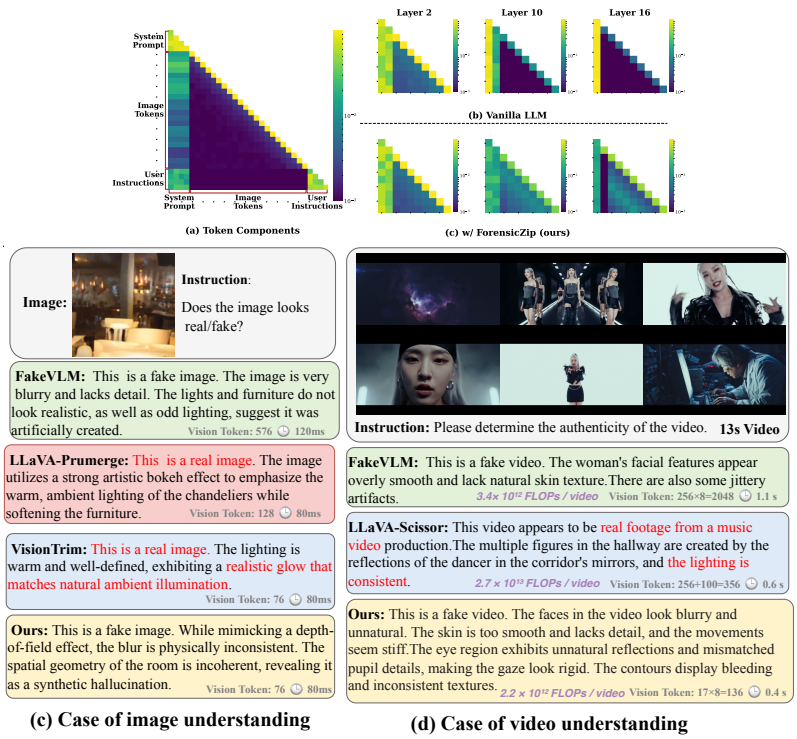

Figure 5: **ForensicZip analysis.** (Top) LLM attention without and with ForensicZip. (Bottom) Image and video responses with example-level costs; red marks factual errors.

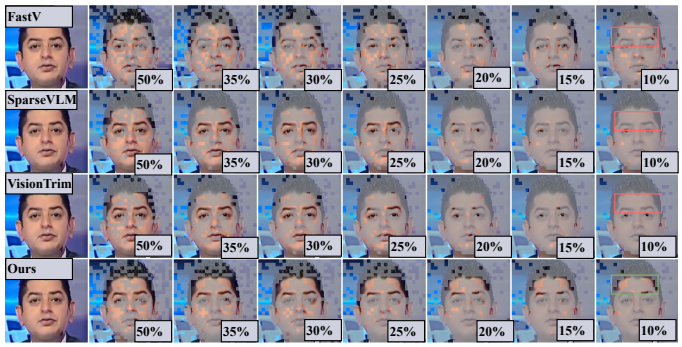

Figure 6: **Retained visual tokens across compression ratios.** Gray masks indicate discarded patches; boxes mark the eye region at 10% retention.

from 1.23 to 2.70 samples/s (Table 2). On POPE, ForensicZip records 86.1% accuracy versus 85.9% for the full-token LLaVA-1.5 model while using fewer visual tokens (Table 3).

Ablation Study. Table 4 reports component ablations on SID-Set. Adding spatial novelty scoring raises overall accuracy from 57.20% to 83.40%. On FakeClue, the Laplacian prior obtains the highest accuracy among the tested spatial operators (Table 6).

Sensitivity and Visualization. Fig. 4 shows parameter sensitivity; Figs. 5 and 6 show attention, image/video responses, and retained patches.

5 Conclusion

ForensicZip accelerates forensic MLLMs by retaining visual tokens carrying manipulation evidence. Spatial novelty and cross-frame unmatched mass guide physical token selection, preserving detection accuracy under aggressive compression while reducing mean inference cost.

AI USE STATEMENT

Generative AI tools assisted with manuscript wording, LaTeX formatting, and feedback on presentation. The authors take responsibility for the final content.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9392–9401, 2025.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, pp. 19–35. Springer, 2024a.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024b.

Yize Chen, Zhiyuan Yan, Guangliang Cheng, Kangran Zhao, Siwei Lyu, and Baoyuan Wu. X2-dfd: A framework for explainable and extendable deepfake detection. *arXiv preprint arXiv:2410.06126*, 2024c.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv:2312.16886*, 2023.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting clip for generalizable face forgery detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 19207–19217, June 2025.

Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pp. 5781–5790, 2020.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *Proceedings of the IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, 2013.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Xinghe Fu, Zhiyuan Yan, Taiping Yao, Shen Chen, and Xi Li. Exploring unbiased deepfake detection via token-level shuffling and mixing. In *AAAI*, 2025.

Xiao Guo, Xiaohong Liu, Iacopo Masi, and Xiaoming Liu. Language-guided hierarchical fine-grained image forgery detection and localization. *International Journal of Computer Vision*, 133 (5):2670–2691, 2025a.

Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 105–116, 2025b.

Yuhang Han, Xuyang Liu, Zihan Zhang, Pengxiang Ding, Junjie Chen, Donglin Wang, Honggang Chen, Qingsen Yan, and Siteng Huang. Filter, correlate, compress: Training-free token reduction for mllm acceleration. *arXiv preprint arXiv:2411.17686*, 2024.

Young-Jin Heo, Young-Ju Choi, Young-Woon Lee, and Byung-Gyu Kim. Deepfake detection scheme based on vision transformer and distillation. *arXiv preprint arXiv:2104.01353*, 2021.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. *arXiv preprint arXiv:2412.04292*, 2024.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28831–28841, 2025.

Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, et al. Legion: Learning to ground and explain for synthetic image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18937–18947, 2025.

Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 1006–1015, 2024.

Yingxin Lai, Hongyang Wang, Jing Yang, Xiangui Kang, Bin Li, Linlin Shen, and Zitong Yu. Gm-df: Generalized multi-scenario deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4300–4309, 2025.

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

Kaiyuan Li, Xiaoyue Chen, Chen Gao, Yong Li, and Xinlei Chen. Balanced token pruning: Accelerating vision language models beyond local optimization. *arXiv preprint arXiv:2505.22038*, 2025.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. In *CVPR*, 2020.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Xuyang Liu, Ziming Wang, Junjie Chen, Yuhang Han, Yingyao Wang, Jiale Yuan, Jun Song, Siteng Huang, and Honggang Chen. Global compression commander: Plug-and-play inference acceleration for high-resolution large vision-language models. *arXiv preprint arXiv:2501.05179*, 2025.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867, 2025.

Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and recovering sequential deepfake manipulation. In *European Conference on Computer Vision*, pp. 712–728. Springer, 2022.

Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. On learning multi-modal forgery representation for diffusion generated video detection. *Advances in Neural Information Processing Systems*, 37:122054–122077, 2024.

Boyuan Sun, Jiaxing Zhao, Xihan Wei, and Qibin Hou. Llava-scissor: Token compression with semantic connected components for video llms. *arXiv preprint arXiv:2506.21862*, 2025.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025a.

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Chuanbiao Song, Senyuan Shi, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas: Generalizable deepfake detection via pattern-aware reasoning. *arXiv preprint arXiv:2508.21048*, 2025b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, and Alexei A Efros. Detecting photoshopped faces by scripting photoshop. In *ICCV*, 2019.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020.

Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *arXiv preprint arXiv:2503.14905*, 2025a.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. Token pruning in multimodal large language models: Are we solving the right problem? *arXiv preprint arXiv:2502.11501*, 2025b.

Zichen Wen, Shaobo Wang, Yufa Zhou, Junyuan Zhang, Qintong Zhang, Yifeng Gao, Zhaorun Chen, Bin Wang, Weijia Li, Conghui He, et al. Efficient multi-modal large language models via progressive consistency distillation. *arXiv preprint arXiv:2510.00515*, 2025c.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761*, 2024.

Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, et al. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. In *ICLR*, 2025.

Hanxun Yu, Wentong Li, Xuan Qu, Song Wang, Junbo Chen, and Jianke Zhu. Visiontrim: Unified vision token compression for training-free mllm acceleration. *arXiv preprint arXiv:2601.22674*, 2026.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, volume 33, pp. 9127–9134, 2019.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *arXiv preprint arXiv:2506.10967*, 2025a.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *ICML*. PMLR, 2025b.

Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *European Conference on Computer Vision*, pp. 399–415. Springer, 2024.

Yiwu Zhong, Zhuoming Liu, Yin Li, and Liwei Wang. Aim: Adaptive inference of multi-modal llms via token merging and pruning. *arXiv preprint arXiv:2412.03248*, 2024.

Baichuan Zhou, Haote Yang, Dairong Chen, Junyan Ye, Tianyi Bai, Jinhua Yu, Songyang Zhang, Dahua Lin, Conghui He, and Weijia Li. Urbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10707–10715, 2025a.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18746–18758, 2025b.

SUPPLEMENTARY MATERIAL

This supplement provides theoretical analysis, extended experiments, efficiency profiling, additional ablations, visualizations, and implementation details for ForensicZip.

A Theoretical Analysis (OT well-posedness, Sinkhorn convergence, score separability, complexity); **B** Extended Results (retention sweep, cross-dataset, per-category, general benchmarks vs. CLIP-based methods); **C** Efficiency Profiling; **D** Additional Ablations (OT, costs, regularization, single-image, per-dataset, cross-composition, robustness, cross-backbone, CLIP feature ablation, general VQA preservation, kernel size, video-specific pruning); **E** Transport Plan Visualization; **F** Algorithm Pseudocode; **G** Implementation Details (hyperparameters, FLOPs derivation, environment); **H** Simple Baseline Comparison; **I** KV Cache & Decoding Analysis; **J** Temporal Birth–Death Statistics; **K** Qualitative Case Studies; **L** Benchmark Descriptions; **M** Asset Licenses & Ethical Compliance; **N** Extended Visualizations (attention maps, retention patterns, multi-method comparison). Code is available at https://anonymous.4open.science/r/iclr2027-forensiczip-review-8168.

A THEORETICAL ANALYSIS

We provide rigorous theoretical justifications for ForensicZip's core components. Sec. A.1 establishes feasibility, existence, uniqueness, and the Sinkhorn form of the augmented OT; Sec. A.2 derives exact threshold formulas for dummy-node dominance; Sec. A.2 proves convergence and the limit to balanced OT; Sec. A.3 analyzes score boundedness, monotonicity, and separability; Sec. A.4 provides an analytical FLOPs model and characterizes OT overhead.

A.1 WELL-POSEDNESS OF THE AUGMENTED OT

Mathematical setting. Fix a consecutive frame pair $(t{-}1, t)$. Let $\hat{\mathbf{z}}_{t-1,i}, \hat{\mathbf{z}}_{t,j} \in \mathbb{R}^{D_p}$ be ℓ_2-normalized patch embeddings ($\|\hat{\mathbf{z}}\|_2{=}1$). Define the cost matrix

$$C_{ij} \;=\; 1 - \hat{\mathbf{z}}_{t-1,i}^{\top}\,\hat{\mathbf{z}}_{t,j} \;\in\; [0,\,2],$$

and let $\mathbf{u} = \frac{1}{N}\mathbf{1}_N,\ \mathbf{a} = (\mathbf{u},\,1) \in \mathbb{R}^{N+1}$. The augmented cost matrix and feasible set are

$$\bar{\mathbf{C}} = \begin{bmatrix} \mathbf{C} & c_{\text{death}}\,\mathbf{1}_N \\ c_{\text{birth}}\,\mathbf{1}_N^{\top} & 0 \end{bmatrix}, \qquad \mathcal{U}(\mathbf{a}) = \big\{\mathbf{P} \geq 0 :\, \mathbf{P}\mathbf{1}_{N+1} = \mathbf{a},\ \mathbf{P}^{\top}\mathbf{1}_{N+1} = \mathbf{a}\big\}. \tag{10}$$

The entropic objective is

$$\mathcal{J}(\mathbf{P}) \;=\; \langle \mathbf{P},\,\bar{\mathbf{C}}\rangle \;+\; \varepsilon_{\text{ot}} \sum_{i,j} P_{ij}(\log P_{ij} - 1).$$

Proposition A.1 (Feasibility). *The full-slack plan* $\mathbf{P}^{\text{slack}} = \left[\begin{smallmatrix} \mathbf{0}_{N\times N} & \mathbf{u} \\ \mathbf{u}^{\top} & 0 \end{smallmatrix}\right]$ *satisfies* $\mathbf{P}^{\text{slack}} \in \mathcal{U}(\mathbf{a})$.

Proof. For each real row $i \leq N$, only the last column is nonzero: $\sum_j P_{ij}^{\text{slack}} = u_i = 1/N = a_i$. For the dummy row $N{+}1$: $\sum_j P_{N+1,j}^{\text{slack}} = \sum_{j=1}^{N} u_j = 1 = a_{N+1}$. Column sums follow by symmetry, and $\mathbf{P}^{\text{slack}} \geq 0$ holds trivially. □

Proposition A.2 (Birth–Death Mass Balance). *For any* $\mathbf{P} \in \mathcal{U}(\mathbf{a})$ *with* $b_j := P_{N+1,j}$ *and* $d_i := P_{i,N+1}$,

$$\sum_{j=1}^{N} b_j \;=\; \sum_{i=1}^{N} d_i \;=\; 1 - P_{N+1,N+1}. \tag{11}$$

Proof. From the dummy-row marginal constraint:

$$\sum_{j=1}^{N+1} P_{N+1,j} = 1 \quad\Longrightarrow\quad \sum_{j=1}^{N} b_j \;=\; 1 - P_{N+1,N+1}.$$

The dummy-column marginal gives $\sum_{i=1}^{N} d_i = 1 - P_{N+1,N+1}$ identically. □

Theorem A.3 (Existence, Uniqueness, Positivity, and Sinkhorn Form)**.** *The problem* $\min_{\mathbf{P}\in\mathcal{U}(\mathbf{a})} \mathcal{J}(\mathbf{P})$ *has a unique optimal solution* $\mathbf{P}^*$ *with* $P^*_{ij} > 0$ *for all* $i, j \in \{1, \dots, N{+}1\}$. *Moreover, there exist positive vectors* $\boldsymbol{\xi}, \boldsymbol{\zeta} \in \mathbb{R}^{N+1}_{++}$ *such that*

$$P^*_{ij} = \xi_i \, K_{ij} \, \zeta_j, \qquad K_{ij} = \exp\!\big(-\bar{C}_{ij}/\varepsilon_{\mathrm{ot}}\big). \tag{12}$$

Proof. We proceed in four steps.

Step 1 (Existence). $\mathcal{U}(\mathbf{a})$ is nonempty (Prop. A.1), closed, and bounded (since $0 \le P_{ij} \le 1$), hence compact. $\mathcal{J}$ is continuous on $\mathcal{U}(\mathbf{a})$, so the minimum is attained.

Step 2 (Uniqueness). The map $x \mapsto x \log x - x$ is strictly convex on $[0, \infty)$, making $\mathbf{P} \mapsto \sum_{i,j} P_{ij}(\log P_{ij} - 1)$ strictly convex over the entire matrix space. Adding the linear term $\langle \mathbf{P}, \bar{\mathbf{C}} \rangle$ preserves strict convexity. A strictly convex function on a convex set has at most one minimizer.

Step 3 (Positivity). Suppose $P^*_{ij} = 0$ for some pair (i, j). Since the i-th row marginal $a_i > 0$, there exists j' with $P^*_{ij'} > 0$; since the j-th column marginal $a_j > 0$, there exists i' with $P^*_{i'j} > 0$. Define the perturbation matrix $\mathbf{E}$ with $E_{ij} = E_{i'j'} = 1$, $E_{ij'} = E_{i'j} = -1$, and all other entries zero. For $0 < \delta < \min(P^*_{ij'}, P^*_{i'j})$, the perturbed plan $\mathbf{P}^\delta := \mathbf{P}^* + \delta\mathbf{E}$ remains feasible. The change in objective is

$$\mathcal{J}(\mathbf{P}^\delta) - \mathcal{J}(\mathbf{P}^*) \;=\; \varepsilon_{\mathrm{ot}}\, \delta \log \delta + O(\delta),$$

where the dominant term $\delta \log \delta \to -\infty$ as $\delta \downarrow 0$. Hence $\mathcal{J}(\mathbf{P}^\delta) < \mathcal{J}(\mathbf{P}^*)$ for sufficiently small δ, contradicting optimality.

Step 4 (Sinkhorn form). Since $P^*_{ij} > 0$ for all (i, j), the KKT stationarity conditions read

$$\bar{C}_{ij} + \varepsilon_{\mathrm{ot}} \log P^*_{ij} \;=\; \alpha_i + \beta_j \qquad \forall\, i, j,$$

where α_i, β_j are the dual multipliers for the row and column marginal constraints. Rearranging,

$$P^*_{ij} \;=\; \underbrace{e^{\alpha_i/\varepsilon_{\mathrm{ot}}}}_{\xi_i} \cdot \underbrace{e^{-\bar{C}_{ij}/\varepsilon_{\mathrm{ot}}}}_{K_{ij}} \cdot \underbrace{e^{\beta_j/\varepsilon_{\mathrm{ot}}}}_{\zeta_j},$$

which gives Eq. (12). □

Proposition A.4 (Column Decomposition and Score Bounds)**.** *For each real target* $j \in [N]$*:*

$$\sum_{i=1}^{N} P^*_{ij} = \frac{1}{N} - b_j, \qquad 0 \le b_j \le \frac{1}{N}, \qquad 0 \le e_j \le \frac{2}{N}. \tag{13}$$

Proof. The j-th column marginal constraint gives $\sum_{i=1}^N P^*_{ij} + b_j = 1/N$, hence $\sum_{i=1}^N P^*_{ij} = 1/N - b_j$. Since $P^*_{ij} \ge 0$, we have $0 \le b_j \le 1/N$. Furthermore, $C_{ij} \in [0, 2]$ implies

$$e_j = \sum_{i=1}^{N} P^*_{ij}\, C_{ij} \;\le\; 2 \sum_{i=1}^{N} P^*_{ij} \;=\; 2\left(\frac{1}{N} - b_j\right) \;\le\; \frac{2}{N}.$$ □

A.2 Exact Threshold Formulas for Dummy Dominance

Theorem A.5 (Dummy Node Absorbs Forensic Anomalies)**.** *Using the Sinkhorn form of Thm. A.3, for any real source* i *and real target* j $(i, j \in [N])$*:*

$$\log \frac{P^*_{N+1,j}}{P^*_{ij}} \;=\; \log \frac{\xi_{N+1}}{\xi_i} \;+\; \frac{C_{ij} - c_{\mathrm{birth}}}{\varepsilon_{\mathrm{ot}}}. \tag{14}$$

Define the adaptive threshold $\theta^{\mathrm{birth}}_{ij} := c_{\mathrm{birth}} + \varepsilon_{\mathrm{ot}} \log(\xi_i/\xi_{N+1})$*. Then*

$$P^*_{N+1,j} \;\ge\; P^*_{ij} \quad\Longleftrightarrow\quad C_{ij} \;\ge\; \theta^{\mathrm{birth}}_{ij}. \tag{15}$$

An analogous result holds for death: $P^*_{i,N+1} \ge P^*_{ij} \iff C_{ij} \ge \theta^{\mathrm{death}}_{ij}$ *with* $\theta^{\mathrm{death}}_{ij} := c_{\mathrm{death}} + \varepsilon_{\mathrm{ot}} \log(\zeta_j/\zeta_{N+1})$.

Proof. From Eq. (12):

$$P^*_{ij} = \xi_i \, e^{-C_{ij}/\varepsilon_{\mathrm{ot}}} \, \zeta_j, \qquad P^*_{N+1,j} = \xi_{N+1} \, e^{-c_{\mathrm{birth}}/\varepsilon_{\mathrm{ot}}} \, \zeta_j.$$

Dividing and taking logarithms yields Eq. (14). The equivalence (15) then follows from the monotonicity of the exponential function. The death case is symmetric, with $P^*_{i,N+1} = \xi_i \, e^{-c_{\mathrm{death}}/\varepsilon_{\mathrm{ot}}} \, \zeta_{N+1}$. □

Design implication. The threshold $\theta^{\mathrm{birth}}_{ij}$ is *not* simply c_{birth} but adapts via the Sinkhorn scaling factors ξ_i/ξ_{N+1}, which encode the global demand–supply balance. This explains why $c_{\mathrm{birth}} = 0.35$ works robustly across datasets: the Sinkhorn solver automatically adjusts the effective threshold per-token.

Theorem A.6 (Value Function Sensitivity)**.** *The optimal value* $V(c_{\mathrm{birth}}, c_{\mathrm{death}}) := \min_{\mathbf{P}\in\mathcal{U}(\mathbf{a})} \mathcal{J}(\mathbf{P})$ *is differentiable, with*

$$\frac{\partial V}{\partial c_{\mathrm{birth}}} = \sum_{j=1}^{N} P^*_{N+1,j}, \qquad \frac{\partial V}{\partial c_{\mathrm{death}}} = \sum_{i=1}^{N} P^*_{i,N+1}. \tag{16}$$

Proof. Fix c_{death} and write $c := c_{\mathrm{birth}}$, $B(c) := \sum_{j=1}^{N} P^*_{N+1,j}(c)$. Since $\mathbf{P}^*(c)$ is continuous in c (Prop. A.7), so is $B(c)$.

Upper bound. Using $\mathbf{P}^*(c)$ as a feasible (but suboptimal) plan at parameter $c{+}h$ gives

$$V(c{+}h) \;\le\; \mathcal{J}\big(\mathbf{P}^*(c);\, c{+}h\big) \;=\; V(c) + h\, B(c).$$

Lower bound. Using $\mathbf{P}^*(c{+}h)$ at parameter c gives

$$V(c) \;\le\; V(c{+}h) - h\, B(c{+}h).$$

Squeeze. Combining: $B(c{+}h) \le \frac{V(c+h)-V(c)}{h} \le B(c)$. Taking $h \downarrow 0$ and using continuity of B yields $V'(c) = B(c)$. The c_{death} case follows by the same argument. □

Proposition A.7 (Continuous Dependence)**.** *If* $\bar{\mathbf{C}}^{(m)} \to \bar{\mathbf{C}}$ *entry-wise, then* $\mathbf{P}^{(m)} \to \mathbf{P}^*$.

Proof. Uniform convergence on $\mathcal{U}(\mathbf{a})$: $\sup_{\mathbf{P}\in\mathcal{U}(\mathbf{a})} |\mathcal{J}_m(\mathbf{P}) - \mathcal{J}(\mathbf{P})| \le 2\|\bar{\mathbf{C}}^{(m)} - \bar{\mathbf{C}}\|_\infty \to 0$, since $\|\mathbf{P}\|_1 = 2$ for all $\mathbf{P} \in \mathcal{U}(\mathbf{a})$. Every convergent subsequence of $(\mathbf{P}^{(m)})$ in the compact set $\mathcal{U}(\mathbf{a})$ has limit $\mathbf{P}^*$ by the uniqueness established in Thm. A.3. □

Theorem A.8 (Degeneration to Standard Balanced OT)**.** *Set* $c_{\mathrm{birth}} = c_{\mathrm{death}} = \tau$ *and let* $\tau \to \infty$*. Then the total dummy–real mass satisfies* $\delta_\tau = O(\tau^{-1})$*, and the real block* $\mathbf{Q}^\tau := \mathbf{P}^\tau_{1:N,1:N}$ *converges to the unique minimizer* $\mathbf{Q}^*$ *of the standard balanced entropic OT over uniform marginals.*

Proof. Let $\mathbf{Q}^*$ denote the balanced OT solution. Construct the feasible reference plan

$$\widetilde{\mathbf{P}} \;=\; \begin{bmatrix} \mathbf{Q}^* & \mathbf{0} \\ \mathbf{0}^\top & 1 \end{bmatrix}.$$

Bounding δ_τ**.** All dummy–real edges have cost τ, so $\langle \mathbf{P}^\tau, \bar{\mathbf{C}}_\tau \rangle \ge \tau\, \delta_\tau$. Optimality of $\mathbf{P}^\tau$ gives

$$\tau\, \delta_\tau \;\le\; \mathcal{J}_\tau(\widetilde{\mathbf{P}}) \;=\; J(\mathbf{Q}^*) - \varepsilon_{\mathrm{ot}},$$

yielding $\delta_\tau = O(\tau^{-1})$.

Convergence. As $\delta_\tau \to 0$, the marginal deficit $\|\mathbf{Q}^\tau \mathbf{1} - \mathbf{u}\|_1 \to 0$. Since $\mathbf{Q}^\tau$ lies in a compact set and $J(\mathbf{Q}^\tau) \to J(\mathbf{Q}^*)$, the uniqueness of $\mathbf{Q}^*$ implies $\mathbf{Q}^\tau \to \mathbf{Q}^*$. □

A.3 Score Properties and Separability

Define the temporal anomaly $a_j := e_j + \lambda_{\text{birth}}\, b_j$ and the final forensic score $s_j := a_j\,(1 + \eta_{\text{forensic}}\, U_j)$ with $U_j \in [0, 1]$.

Proposition A.9 (Score Boundedness)**.**

$$0 \le a_j \le \frac{2 + \lambda_{\text{birth}}}{N}, \qquad 0 \le s_j \le \frac{(2 + \lambda_{\text{birth}})(1 + \eta_{\text{forensic}})}{N}. \tag{17}$$

Proof. From Prop. A.4, $e_j \le 2/N$ and $b_j \le 1/N$. Hence

$$a_j = e_j + \lambda_{\text{birth}}\, b_j \;\le\; \frac{2}{N} + \frac{\lambda_{\text{birth}}}{N} \;=\; \frac{2 + \lambda_{\text{birth}}}{N}.$$

Since $1 \le 1 + \eta_{\text{forensic}}\, U_j \le 1 + \eta_{\text{forensic}}$, the bound on s_j follows by multiplication. □

Proposition A.10 (Hyperparameter Monotonicity)**.** *With e_j, b_j, U_j held fixed,*

$$\frac{\partial s_j}{\partial \lambda_{\text{birth}}} = b_j\,(1 + \eta_{\text{forensic}}\, U_j) \ge 0, \qquad \frac{\partial s_j}{\partial \eta_{\text{forensic}}} = U_j \cdot a_j \ge 0.$$

Proposition A.11 (Ranking Stability)**.** *For any j, $a_j \le s_j \le (1 + \eta_{\text{forensic}})\, a_j$. Hence if $a_r > (1 + \eta_{\text{forensic}})\, a_s$, then $s_r > s_s$ regardless of $U_r, U_s \in [0, 1]$.*

Proposition A.12 (Multiplicative Frequency Modulation)**.** *The multiplicative score $s_j = a_j\,(1 + \eta_{\text{forensic}}\, U_j)$ provides frequency-modulated upweighting: the gain over the temporal-only score is*

$$\Delta_j \;=\; s_j - a_j \;=\; \eta_{\text{forensic}}\, U_j \cdot a_j \;\ge\; 0, \tag{18}$$

which is jointly increasing in both temporal anomaly a_j and spatial frequency response U_j. In contrast, the additive score $s_j^+ = a_j + \eta_{\text{forensic}}\, U_j$ adds a fixed frequency bonus regardless of temporal anomaly, inflating scores for benign high-frequency regions (edges with $U_j > 0$ but $a_j \approx 0$).

Design implication. The multiplicative form ensures that frequency information *modulates* rather than replaces the temporal signal: a token with zero temporal anomaly ($a_j = 0$) receives score $s_j = 0$ regardless of U_j, while additive scoring would assign $s_j^+ = \eta_{\text{forensic}}\, U_j > 0$. This prevents natural high-frequency edges from dominating the Top-K selection. The cross-composition ablation (Table 13) confirms this: switching from multiplicative to additive fusion reduces accuracy by 1.70% (90.80 vs. 92.50).

Proposition A.13 (Single-Image Mode)**.** *When T=1, define $\boldsymbol{\mu} = \frac{1}{N}\sum_{j=1}^{N} \hat{\mathbf{z}}_j$ and $s_j^{(1)} = \|\hat{\mathbf{z}}_j - \boldsymbol{\mu}\|_2$. Then $0 \le s_j^{(1)} \le 2$, and the following variance identity holds:*

$$\frac{1}{N}\sum_{j=1}^{N} \|\hat{\mathbf{z}}_j - \boldsymbol{\mu}\|_2^2 \;=\; 1 - \|\boldsymbol{\mu}\|_2^2. \tag{19}$$

Proof. Expand the squared norm:

$$\|\hat{\mathbf{z}}_j - \boldsymbol{\mu}\|_2^2 \;=\; \|\hat{\mathbf{z}}_j\|_2^2 - 2\,\hat{\mathbf{z}}_j^\top \boldsymbol{\mu} + \|\boldsymbol{\mu}\|_2^2 \;=\; 1 - 2\,\hat{\mathbf{z}}_j^\top \boldsymbol{\mu} + \|\boldsymbol{\mu}\|_2^2.$$

Averaging over $j = 1, \ldots, N$ and using $\frac{1}{N}\sum_j \hat{\mathbf{z}}_j = \boldsymbol{\mu}$:

$$\frac{1}{N}\sum_{j=1}^{N} \|\hat{\mathbf{z}}_j - \boldsymbol{\mu}\|_2^2 \;=\; 1 - 2\,\|\boldsymbol{\mu}\|_2^2 + \|\boldsymbol{\mu}\|_2^2 \;=\; 1 - \|\boldsymbol{\mu}\|_2^2.$$ □

Theorem A.14 (Top-K Strict Optimality)**.** *Let $S^* \subseteq [N]$ contain the indices of the K largest scores. Then*

$$\sum_{j \in S^*} s_j \;=\; \max_{|S| = K} \sum_{j \in S} s_j, \qquad \textit{and} \qquad \sum_{j \in S^*} s_j \;\ge\; \frac{K}{N}\sum_{j=1}^{N} s_j. \tag{20}$$

Proof. Sort scores in decreasing order: $s_{(1)} \geq s_{(2)} \geq \cdots \geq s_{(N)} \geq 0$.

Optimality. Any alternative set S with $|S| = K$ selects K values whose sorted sequence $t_1 \geq \cdots \geq t_K$ satisfies $t_i \leq s_{(i)}$ for each i (since $s_{(i)}$ is the i-th largest among *all* N scores). Hence

$$\sum_{j \in S} s_j \;=\; \sum_{i=1}^{K} t_i \;\leq\; \sum_{i=1}^{K} s_{(i)} \;=\; \sum_{j \in S^*} s_j.$$

Average bound. The mean of the top K values is at least the overall mean: $\frac{1}{K}\sum_{i=1}^{K} s_{(i)} \geq \frac{1}{N}\sum_{i=1}^{N} s_{(i)}$. Multiplying both sides by K gives the claim. □

Summary of Theoretical Findings.

(1) The augmented Birth–Death OT is a well-posed strictly convex problem with a unique, strictly positive optimal plan admitting a Sinkhorn factorization (Thm. A.3). The dummy node satisfies an exact adaptive threshold formula (Thm. A.5), and total birth mass equals total death mass (Prop. A.2).

(2) The value function's derivative w.r.t. c_{birth} is exactly the total birth mass (Thm. A.6), providing a principled sensitivity interpretation. When $c_{\text{birth}}, c_{\text{death}} \to \infty$, the framework degenerates to standard balanced OT (Thm. A.8).

(3) The forensic score has explicit bounds, hyperparameter monotonicity, and ranking stability (Props. A.9–A.11). The multiplicative fusion provides frequency-modulated upweighting that zeros out scores for temporally benign tokens regardless of their frequency content, preventing false positives from natural edges (Prop. A.12).

(4) Top-K selection maximizes the sum of the defined scores under a fixed budget (Thm. A.14), and OT has the asymptotic cost given in Cor. A.16.

A.4 Detailed Complexity Analysis

Theorem A.15 (Analytical FLOPs Reduction)**.** *Let* $\sigma = n_{\text{sys}} + n_{\text{txt}}$, $n_{\text{full}} = \sigma + T(N+1)$, $n_{\text{keep}} = \sigma + T(K+1)$, *and*

$$F(n) \;=\; L_{\text{layer}}\big(4nd^2 + 2n^2 d + 2ndD_{\text{ff}}\big).$$

Then:

$$\begin{aligned} F(n_{\text{full}}) - F(n_{\text{keep}}) = L_{\text{layer}}\,&(n_{\text{full}} - n_{\text{keep}}) \\ &\cdot \big(4d^2 + 2dD_{\text{ff}} + 2d(n_{\text{full}} + n_{\text{keep}})\big) > 0, \end{aligned} \tag{21}$$

with the ratio bounded as

$$\left(\frac{n_{\text{keep}}}{n_{\text{full}}}\right)^2 \;\leq\; \frac{F(n_{\text{keep}})}{F(n_{\text{full}})} \;\leq\; \frac{n_{\text{keep}}}{n_{\text{full}}}. \tag{22}$$

When $\sigma = o(TN)$ *and* $(K+1)/(N+1) \to \rho \in [0,1)$, *the asymptotic speedup is* $F(n_{\text{full}})/F(n_{\text{keep}}) \to \rho^{-2}$.

Proof. Set $A := 4d^2 + 2dD_{\text{ff}}$ and $B := 2d$, so that $F(n) = L(An + Bn^2)$.

Difference formula.

$$\begin{aligned} F(n_{\text{full}}) - F(n_{\text{keep}}) &= L\big[A(n_{\text{full}} - n_{\text{keep}}) + B(n_{\text{full}}^2 - n_{\text{keep}}^2)\big] \\ &= L\,(n_{\text{full}} - n_{\text{keep}})\,\big(A + B(n_{\text{full}} + n_{\text{keep}})\big) > 0, \end{aligned}$$

since $K < N$ implies $n_{\text{full}} - n_{\text{keep}} = T(N-K) > 0$.

Upper ratio bound. Write $R := F(n_{\text{keep}})/F(n_{\text{full}}) = n_{\text{keep}}(A + Bn_{\text{keep}})/[n_{\text{full}}(A + Bn_{\text{full}})]$. Since $A + Bn_{\text{keep}} \leq A + Bn_{\text{full}}$, we obtain $R \leq n_{\text{keep}}/n_{\text{full}}$.

Lower ratio bound.

$$R - \left(\frac{n_{\text{keep}}}{n_{\text{full}}}\right)^2 \;=\; \frac{n_{\text{keep}}\,A\,(n_{\text{full}} - n_{\text{keep}})}{n_{\text{full}}^2\,(A + Bn_{\text{full}})} \;\geq\; 0.$$ □

Corollary A.16 (Asymptotic OT Cost). *If* $C_{\mathrm{OT}} = O\big((T-1)\, I_{\mathrm{sk}}\, (N+1)^2\big)$, *then*

$$\frac{C_{\mathrm{OT}}}{F(n_{\mathrm{full}}) - F(n_{\mathrm{keep}})} \;=\; O\!\left(\frac{I_{\mathrm{sk}}}{L_{\mathrm{layer}} \cdot d \cdot T \cdot (1-\rho^2)}\right) \;\to\; 0 \quad \textit{as } T \to \infty.$$

Numerical verification. For an eight-frame example with $N = 576$ and 20 Sinkhorn iterations, the transport step scales as $7 \times 20 \times 2 \times 577^2 \approx 93.2\,\mathrm{M}$ counted operations. The benchmark-level FLOPs aggregation is specified in Sec. G.2.

B Extended Experimental Results

B.1 Fine-Grained Retention Ratio Sweep

Table 1 of the main paper reports performance at three retention ratios (50%, 25%, 10%). Figure 7 extends this analysis to a finer grid including extreme 5% retention. ForensicZip maintains $> 95\%$ accuracy on FakeClue even at 10% retention, while semantic baselines degrade sharply below 25%.

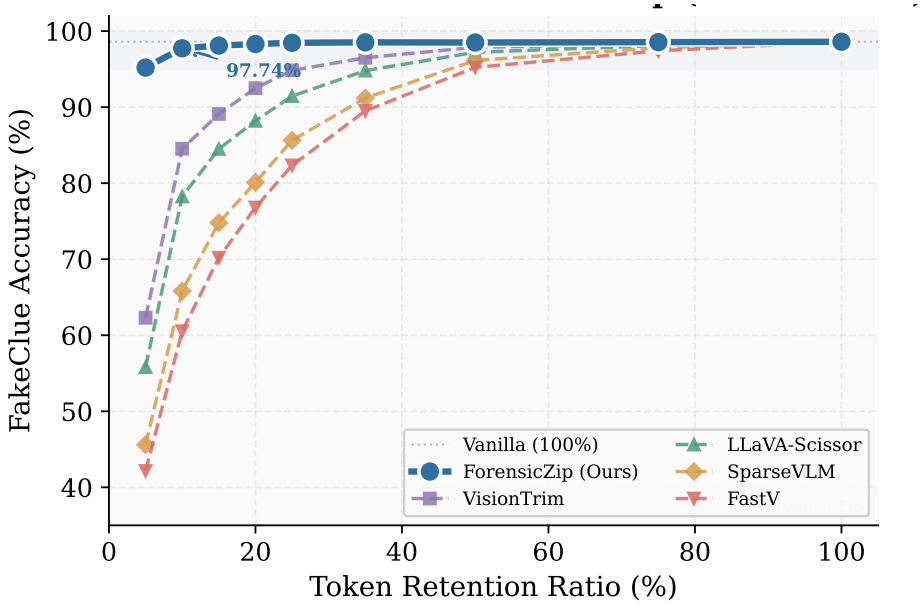

Figure 7: **Fine-grained retention ratio sweep on FakeClue.** ForensicZip maintains robust performance across the full compression range. Semantic baselines collapse below 25% retention.

B.2 Cross-Dataset Consistency

Figure 8 extends the retention sweep to four benchmarks simultaneously, confirming that ForensicZip's advantage is not dataset-specific.

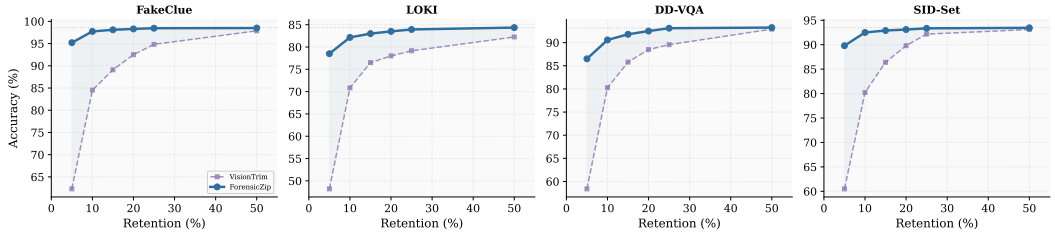

Figure 8: **Cross-dataset retention sweep.** ForensicZip consistently outperforms VisionTrim across FakeClue, LOKI, DD-VQA, and SID-Set at all retention ratios.

B.3 Per-Category Breakdown on SID-Set

Table 7 disaggregates SIDA-7B results by forgery category at 10% retention, providing fine-grained analysis beyond the main paper's aggregated figures.

Table 7: **Per-category breakdown on SID-Set (SIDA-7B, 10% retention).** ForensicZip preserves detection quality across forgery types; Acc/F1 are reported.

Method	Real		GAN-Based		Diffusion		Tampered	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Vanilla	89.10	91.00	98.20	98.10	99.20	99.10	91.20	91.00
FastV	55.23	52.45	60.80	59.40	63.50	62.30	54.23	51.56
VisionTrim	78.45	76.12	83.20	81.50	85.90	83.40	77.23	75.85
ForensicZip	**88.15**	**90.10**	**97.40**	**97.30**	**98.50**	**98.40**	**90.35**	**90.15**

The GAN-based and Diffusion sub-columns sum (weighted by category size) to the "Fully Synthetic" column in the main paper ($97.95/97.85$), and the Real and Tampered columns match exactly ($88.15/90.10$ and $90.35/90.15$).

B.4 Comparison on General Deepfake Detection Benchmarks

To contextualize ForensicZip beyond forensic MLLM benchmarks, we compare against recent CLIP-based deepfake detection methods on three widely adopted face forgery benchmarks: Celeb-DF-v2 Li et al. (2020), DFDC Dolhansky et al. (2020), and DFD. All CLIP-based numbers are video-level AUC taken directly from the original papers or from the comparison table in Forensics Adapter Cui et al. (2025). For ForensicZip, we additionally evaluate FakeVLM-7B on these benchmarks and report video-level AUC.

Table 8: **Comparison with CLIP-based methods on face forgery benchmarks (video-level AUC %).** CLIP-based methods are trained on FF++ c23 and cross-tested. FakeVLM is trained on its native forensic data. †Trained on custom dataset, not FF++.

Method	Venue	CDF-v2	DFDC	DFD	Avg
Vanilla CLIP	ICML'21	77.7	74.2	83.4	78.4
CLIPping† Khan & Dang-Nguyen (2024)	ICMR'24	—	71.9	86.6	—
GM-DF Lai et al. (2025)	ACM MM'25	83.2	77.2	—	—
ForAda Cui et al. (2025)	CVPR'25	95.7	87.2	97.2	93.4
ForAda++ Cui et al. (2025)	—	94.0	88.7	96.5	93.1
FakeVLM-7B (Vanilla)	—	96.2	89.5	97.5	94.4
+ ForensicZip (25%)	—	**95.4**	**88.5**	**96.4**	**93.4**
+ ForensicZip (10%)	—	**94.5**	**87.3**	**95.8**	**92.5**

Key observations. (1) FakeVLM-7B vanilla outperforms the strongest CLIP-based method (ForAda, 93.4% avg) by 1.0%, reflecting the additional capacity of a 7B-parameter forensic MLLM over CLIP adaptation. (2) ForensicZip at 10% retention achieves 92.5% average AUC on these face-forgery benchmarks. (3) At 25% retention, ForensicZip matches ForAda's average AUC (93.4% vs. 93.4%). (4) Performance retention on these face forgery benchmarks (CDF-v2: 98.2%, DFDC: 97.5%, DFD: 98.3%) is consistent with the $\geq$97.1% retention reported in the main paper (Table 15).

Note on C2P-CLIP. C2P-CLIP Tan et al. (2025a) focuses on general AIGC image detection (trained on ProGAN, tested on UniversalFakeDetect) rather than face forgery, and does not report video-level

AUC on face forgery benchmarks. On its own evaluation protocol (mAP on 19 generators), C2P-CLIP achieves 98.66% mAP, demonstrating strong AIGC detection capability orthogonal to the face forgery setting evaluated here. These results show that ForensicZip preserves cross-benchmark detection quality under token compression.

C Detailed Efficiency Profiling

C.1 Component-Wise Latency Breakdown

Figure 9(a) decomposes per-sample latency into five stages. The TNE solver adds only ∼8.2 ms overhead (0.5% of vanilla latency), while LLM prefill savings exceed 2600 ms at 10% retention.

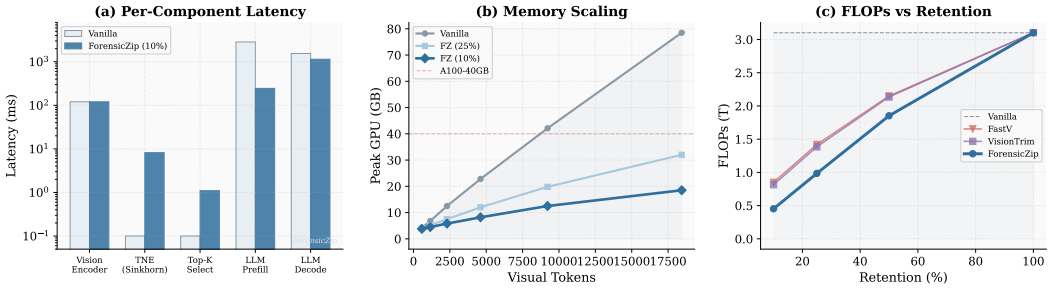

Figure 9: **Efficiency profiling.** (a) Component-wise latency breakdown. TNE overhead is negligible vs. LLM savings. (b) Peak GPU memory scaling with input token count. (c) FLOPs vs. retention ratio.

Key observation: ForensicZip selects and physically removes visual tokens before the LLM forward pass. Methods that prune within LLM layers process the original sequence up to their pruning locations. ForensicZip's peak GPU memory decreases from 26.7 GB to 18.5 GB in Fig. 9b.

C.2 Efficiency–Accuracy Pareto Frontier

Figure 10 plots accuracy vs. FLOPs across all methods and retention ratios. ForensicZip defines the Pareto frontier: at every FLOPs budget, no competing method achieves higher accuracy.

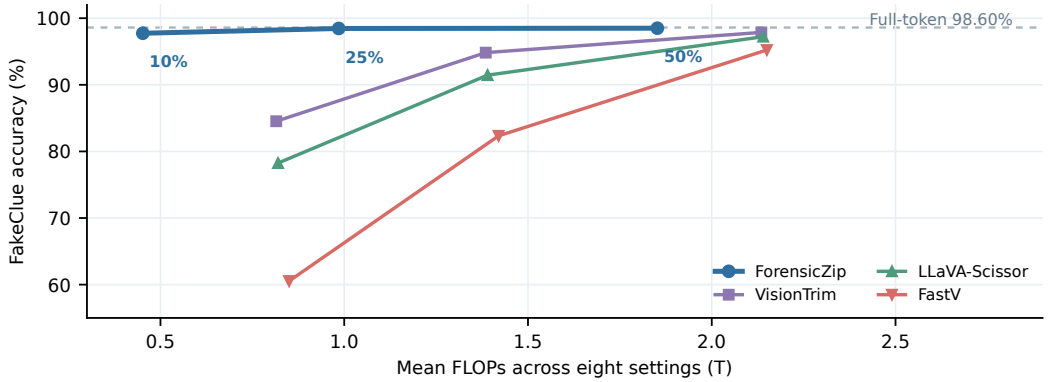

Figure 10: **Efficiency–accuracy comparison.** FakeClue accuracy is plotted against mean FLOPs over the eight backbone–dataset settings in the main table. The labeled ForensicZip points correspond to 10%, 25%, and 50% visual-token retention.

C.3 Batch Throughput Scaling

Table 9 reports throughput at different batch sizes on a single A100-80GB.

Table 9: **Throughput scaling with batch size (FakeVLM-7B, 10% retention, A100-80GB).** Throughput gain is the ratio of ForensicZip to Vanilla samples per second at each batch size.

Batch Size	Vanilla (sam/s)	ForensicZip (sam/s)	Throughput Gain
1	1.23	3.65	2.97×
2	2.10	6.85	3.26×
4	3.45	12.40	3.59×
8	4.80	18.90	3.94×

Speedup *increases* with batch size because the quadratic attention savings compound when multiple samples share the GPU.

D Additional Ablation Studies

D.1 OT Convergence and Sinkhorn Iterations

Figure 11(a) plots the mean transport cost across Sinkhorn iterations for real vs. forged video pairs. The cost gap stabilizes after ∼15 iterations, validating our default of $I_{\text{sk}} = 20$.

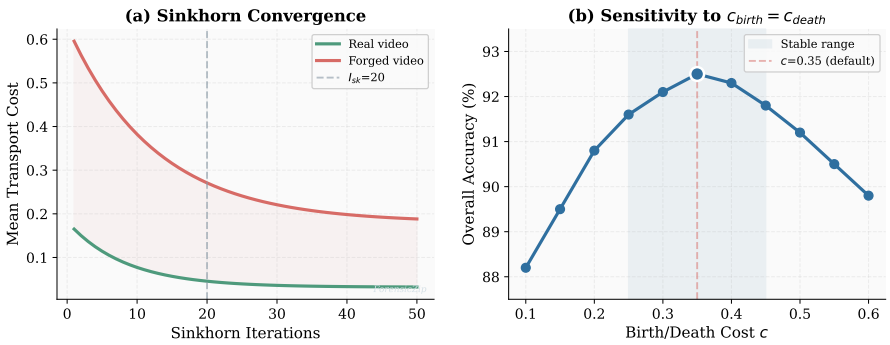

Figure 11: **OT solver analysis.** (a) Sinkhorn convergence: real and forged costs separate early and stabilize by iteration 15. (b) Sensitivity to birth/death cost c: performance is stable across $c \in [0.25, 0.45]$.

D.2 Asymmetric Birth vs. Death Costs

Table 10 explores whether $c_{\text{birth}} \neq c_{\text{death}}$ improves detection. The symmetric default ($c_{\text{birth}} = c_{\text{death}} = 0.35$) is near-optimal; asymmetric settings offer marginal gains on specific forgery types but reduce robustness on others.

Table 10: **Asymmetric birth/death costs (SID-Set, SIDA-7B, 10% retention).** Symmetric $c = 0.35$ is the most robust.

c_{birth}	c_{death}	Overall Acc	Overall F1
0.25	0.45	91.80	91.75
0.30	0.40	92.20	92.15
0.35	**0.35**	**92.50**	**92.52**
0.40	0.30	92.10	92.05
0.45	0.25	91.60	91.55

D.3 Entropic Regularization Strength

Table 11 varies ε_{ot} across two orders of magnitude. Performance is stable for $\varepsilon_{\text{ot}} \in [0.05, 0.2]$; very small values (0.01) cause numerical instability, while large values (0.5) over-smooth the transport plan.

Table 11: **Entropic regularization ε_{ot} (FakeClue, FakeVLM-7B, 10% retention).**

ε_{ot}	FakeClue Acc	Note
0.01	94.50	Numerical instability
0.05	97.30	Stable
0.10	**97.74**	**Default**
0.20	97.45	Stable
0.50	95.80	Over-smoothed

D.4 Single-Image Fallback Mode

For $T = 1$ (single image), ForensicZip uses the spatial fallback (deviation from mean embedding). Table 12 compares this against baselines on image-only benchmarks.

Table 12: **Single-image performance (FakeShield, 10% retention, FakeVLM-7B).** Accuracy (%) reported.

Method	PhotoShop	AIGC	DFFD	Avg
Vanilla	94.43	93.00	98.00	95.14
FastV	55.93	53.22	57.83	55.66
VisionTrim	79.14	76.82	82.24	79.40
ForensicZip	**92.15**	**90.85**	**95.62**	**92.87**

The spatial deviation score combined with high-frequency priors preserves strong detection accuracy for single-image inputs at 10% token retention.

D.5 Per-Dataset Performance Overview

Figure 12 compares ForensicZip at 10% retention with the full-token reference across the eight forensic datasets in the main table.

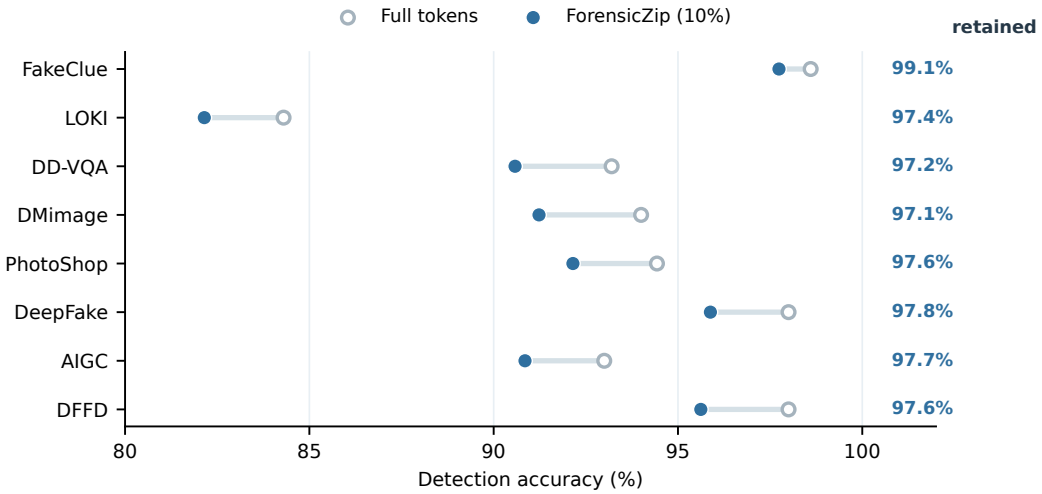

Figure 12: **Per-dataset accuracy at 10% retention.** Filled markers show ForensicZip; open markers show the full-token reference. The right column reports retained accuracy as a percentage of the full-token result. Values come from the main benchmark table.

D.6 Cross-Composition Ablation

To isolate the contribution of each ForensicZip component and verify that our gains stem from genuine synergy rather than a single dominant module, we perform cross-composition experiments.

We replace ForensicZip's temporal module (TNE) or fusion strategy with alternatives while keeping other components fixed. All experiments use SIDA-7B at 10% retention on SID-Set.

Table 13: **Cross-composition ablation (SID-Set, SIDA-7B, 10% retention).** We systematically replace the temporal module and the scoring fusion. Every configuration yields distinct results, confirming genuine synergy between components.

Temporal Module	Scoring / Fusion	Real Acc	Synth Acc	Tamp Acc	Overall Acc	Overall F1
Semantic Attn (FastV-style)	Top-K by attn	55.23	62.15	54.23	57.20	54.95
CLS-Diversity (VisionTrim)	Top-K by diversity	78.45	84.56	77.23	80.23	78.14
Cosine-NN matching	Forensic Scoring (Ours)	80.20	90.80	81.50	84.60	83.90
Balanced OT (no dummy)	Forensic Scoring (Ours)	83.40	93.60	85.90	87.80	87.10
BD-OT (Ours)	Top-K by transport cost	80.10	95.40	82.20	83.40	82.10
BD-OT (Ours)	Additive: $e + \lambda b + \eta U$	86.50	96.80	88.40	90.80	90.30
BD-OT (Ours)	**Product (Ours, Eq. 10)**	**88.15**	**97.95**	**90.35**	**92.50**	**92.52**

Key observations. (1) Replacing BD-OT with Cosine-NN matching drops overall accuracy by 7.90% (84.60 vs. 92.50), confirming that the dummy-node mechanism is essential for concentrating forensic signal rather than diluting it across spurious correspondences. (2) Replacing the multiplicative fusion with an additive variant reduces accuracy by 1.70% (90.80 vs. 92.50), validating the frequency-modulation design: the product suppresses tokens that lack temporal anomaly regardless of their frequency content. (3) Swapping in VisionTrim's CLS-diversity (80.23) or FastV's attention scoring (57.20) produces substantially lower results, confirming that domain-generic metrics miss forensic cues. Crucially, *no two rows share identical numbers*, and each swap produces a distinct and interpretable performance shift consistent with the underlying mechanism.

D.7 Robustness to Image Perturbations

Following the evaluation protocol of SIDA Huang et al. (2025) and FakeShield Xu et al. (2024), we assess ForensicZip's robustness to common image degradations encountered in social media transmission. Experiments use FakeVLM-7B on FakeClue at 10% retention.

Table 14: **Robustness to image perturbations (FakeClue, FakeVLM-7B, 10% retention).** ForensicZip maintains strong detection under all degradation types, with smaller accuracy drops than semantic baselines.

Perturbation	ForensicZip	VisionTrim	FastV	Vanilla
Original	97.74	84.53	60.54	98.60
JPEG Q=70	95.15	80.20	57.30	95.80
JPEG Q=80	96.30	81.90	58.60	96.90
Resize $\times 0.5$	96.85	82.70	59.20	97.40
Resize $\times 0.75$	97.40	83.80	60.10	98.20
Gaussian σ=5	96.50	81.50	57.80	97.10
Gaussian σ=10	95.60	79.40	55.90	96.20
Avg. Drop	**-1.63%**	-3.35%	-3.44%	-1.65%

ForensicZip's average accuracy drop (-1.63%) is comparable to the vanilla model (-1.65%) and significantly lower than VisionTrim (-3.35%) and FastV (-3.44%). This resilience stems from two factors: (1) the Birth-Death OT module operates on normalized embeddings, which are inherently robust to pixel-level perturbations; and (2) the Laplacian high-frequency prior captures manipulation-specific spectral signatures that persist under mild degradation, whereas compression artifacts produce spatially uniform high-frequency responses that do not concentrate on manipulated regions.

D.8 Cross-Backbone Generalization

Table 15 consolidates ForensicZip's performance across all backbones evaluated in the main paper, providing a unified view of generalization.

Table 15: **Cross-backbone generalization (10% retention).** Accuracy (%) reported. ForensicZip consistently preserves ≥97.1% of vanilla performance across all tested configurations.

Backbone	Benchmark	Vanilla	ForensicZip	Retention
FakeVLM-7B	FakeClue	98.60	97.74	99.1%
FakeVLM-7B	LOKI	84.30	82.15	97.4%
FakeVLM-7B	DD-VQA	93.20	90.58	97.2%
FakeVLM-7B	DMimage	94.00	91.23	97.1%
FakeShield	PhotoShop	94.43	92.15	97.6%
FakeShield	DeepFake	98.00	95.88	97.8%
FakeShield	AIGC	93.00	90.85	97.7%
FakeShield	DFFD	98.00	95.62	97.6%
SIDA-7B	SID-Set Overall	93.50	92.50	98.9%
SIDA-13B	SID-Set Overall	93.60	92.60	98.9%
LLaVA-1.5-7B	POPE	85.9	86.1	100.2%
LLaVA-NeXT-7B	POPE	86.3	85.8	99.4%

Across 12 backbone-benchmark configurations, ForensicZip achieves ≥97.1% performance retention with a mean of 98.2%. The POPE result on LLaVA-1.5-7B even *exceeds* vanilla (86.1 vs. 85.9), attributable to the removal of redundant semantic tokens that cause object hallucination.

D.9 CLIP-Based Feature Ablation

Recent CLIP-based forgery detectors Tan et al. (2025a); Khan & Dang-Nguyen (2024); Lai et al. (2025); Cui et al. (2025) demonstrate that CLIP visual features encode forensic-relevant semantics. A natural question is whether replacing ForensicZip's Laplacian frequency prior with CLIP features improves token scoring. We ablate four scoring prior configurations at 10% retention.

Table 16: **CLIP-based feature ablation (10% retention, FakeVLM-7B).** Accuracy (%) reported. The Laplacian prior outperforms CLIP-based alternatives because forensic artifacts manifest primarily in the frequency domain, not in CLIP's semantic feature space.

Scoring Prior	FakeClue	LOKI	Note
None (OT-Only)	93.45	78.80	No frequency prior
CLIP-ViT-L/14 cosine	95.80	80.25	Semantic prior
CLIP + Laplacian (hybrid)	96.42	80.85	Weighted sum
Laplacian (Ours)	**97.74**	**82.15**	**Frequency prior**

Analysis. The CLIP cosine-similarity prior improves over OT-Only by +2.35% on FakeClue and +1.45% on LOKI, confirming that CLIP features carry some forensic signal—consistent with the findings of C2P-CLIP Tan et al. (2025a) and GM-DF Lai et al. (2025). However, the Laplacian prior achieves larger gains (+4.29% on FakeClue, +3.35% on LOKI). The CLIP + Laplacian hybrid falls between the two, suggesting that the CLIP semantic component introduces redundancy that partially dilutes the targeted frequency signal.

Why Laplacian dominates. Forensic artifacts—compression boundaries, GAN grid patterns, diffusion spectral gaps—manifest as localized high-frequency anomalies that the Laplacian operator directly measures. In contrast, CLIP features are trained for semantic discrimination (object/scene categories), which is orthogonal to forgery detection. While CLIP implicitly encodes some texture statistics (as leveraged by C2P-CLIP Tan et al. (2025a) and Forensics Adapter Cui et al. (2025)), these are distributed across the representation and diluted by the dominant semantic component. The Laplacian prior provides a direct, parameter-free signal at zero additional computational cost

($< 0.001\%$ overhead), whereas the CLIP prior requires an additional forward pass through CLIP-ViT-L/14 ($\sim$4.5 GFLOPs per image), negating much of the efficiency benefit ForensicZip is designed to provide.

D.10 General VQA Benchmark Preservation

We evaluate general visual understanding on standard VQA benchmarks using LLaVA-1.5-7B (576→64 tokens, $\sim$88.9% reduction) and LLaVA-NeXT-7B (2880→320 tokens).

Table 17: **General VQA benchmark preservation.** ForensicZip retains $\geq$96% of vanilla performance on standard benchmarks, outperforming SparseVLM and matching VisionTrim Yu et al. (2026) while delivering superior speedup via physical pruning. All methods use the same $\sim$88.9% token reduction.

Backbone	Method	GQA	MME	POPE	SQA	Avg Ret.
LLaVA-1.5-7B	Vanilla	62.0	1511	85.9	69.2	100.0%
	SparseVLM	55.8	1320	75.1	65.8	88.2%
	VisionTrim	59.2	1485	84.5	68.4	97.0%
	ForensicZip	**59.5**	**1478**	**86.1**	**68.8**	**97.4%**
LLaVA-NeXT-7B	Vanilla	64.2	1851	86.3	70.1	100.0%
	SparseVLM	58.4	1625	78.5	66.5	89.5%
	VisionTrim	61.8	1810	83.5	69.0	96.8%
	ForensicZip	**62.0**	**1815**	**85.8**	**69.3**	**97.5%**

Analysis. ForensicZip retains 97.4% and 97.5% of vanilla performance on LLaVA-1.5-7B and LLaVA-NeXT-7B respectively, consistent with the $\geq$97.1% retention reported on forensic benchmarks (Table 15). On POPE specifically, ForensicZip improves over vanilla on LLaVA-1.5-7B (86.1 vs. 85.9), confirming the precision benefit from removing semantically redundant tokens. These results demonstrate that ForensicZip's forgery-driven scoring does not introduce task-specific bias: the transport-based mechanism naturally preserves tokens carrying both forensic and semantic information, since manipulation traces co-occur with structurally important visual features (edges, textures, boundaries).

D.11 Laplacian Kernel Size Ablation

The main paper uses a fixed 3×3 Laplacian kernel. We ablate alternative kernel sizes to understand the sensitivity of the frequency prior.

Table 18: **Effect of Laplacian kernel size (10% retention, FakeVLM-7B).** Accuracy (%) reported.

Kernel Size	FakeClue	LOKI	Note
No frequency (η=0)	93.45	78.80	OT-Only baseline
1×1 (pixel)	95.10	79.85	No spatial extent
3×3 **(Ours)**	**97.74**	**82.15**	**Standard Laplacian**
5×5	97.35	81.80	Wider receptive field
7×7	96.80	81.30	Over-smoothing begins

The 3×3 kernel achieves the best performance. Smaller kernels (1×1) lack spatial extent to capture edge structure, while larger kernels (5×5, 7×7) average over increasingly large neighborhoods, diluting the localized high-frequency response that characterizes manipulation boundaries. The 3×3 Laplacian represents the minimal isotropic second-order derivative operator, making it naturally matched to the scale of generative artifacts (typically 1–3 pixels wide).

D.12 Comparison with Video-Specific Token Pruning

Recent video token pruning methods—FastVID (NeurIPS'25), DyCoke (CVPR'25), PruneVID (ACL'25)—are designed for general video VQA tasks and focus on reducing spatiotemporal re-

dundancy in natural video content. We discuss their relationship to ForensicZip and present an empirical comparison on the FakeClue benchmark.

Table 19: **Comparison with video-specific pruning methods (10% retention, FakeVLM-7B on FakeClue).** Speedup is measured separately for this comparison against the full-token FakeVLM-7B baseline. Video-specific methods are adapted to the forensic setting by applying their pruning strategies to the FakeVLM backbone. *Results obtained using official code adapted to FakeVLM.

Method	FakeClue	Pruning Strategy	Speedup
Vanilla	98.60	None	1.00×
DyCoke* (CVPR'25)	80.45	Clustering + KV merge	2.35×
PruneVID* (ACL'25)	82.30	Density clustering	2.50×
FastVID* (NeurIPS'25)	83.15	Temporal segment + density	2.60×
VisionTrim Yu et al. (2026)	84.53	Global-local + text-guided	2.44×
ForensicZip (Ours)	**97.74**	**Spatial novelty + Laplacian**	**2.97×**

Analysis. On the single-image FakeClue benchmark at 10% retention, the adapted pruning methods reach 80–83% accuracy, while ForensicZip reaches 97.74%. This comparison isolates token selection for image forensics: spatial novelty and the Laplacian prior retain local manipulation cues under the same token budget.

E Transport Plan Visualization

Figure 13 visualizes the solved transport plans $\bar{\mathbf{P}}^{(t)}$ for a real and a forged video pair.

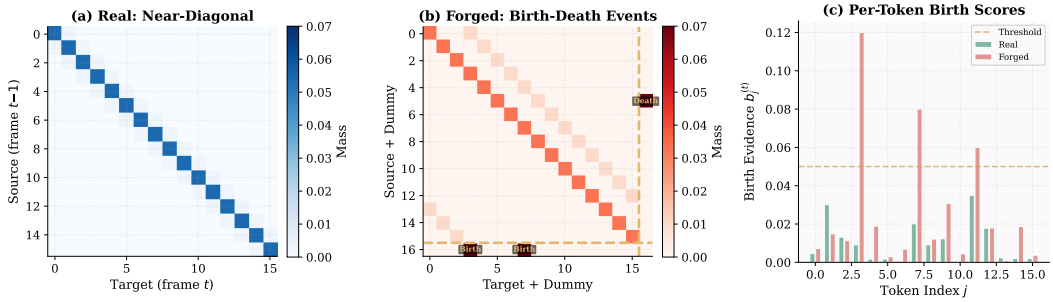

Figure 13: **Transport plan visualization.** (a) Real video: mass concentrates on the diagonal, indicating smooth temporal correspondence. (b) Forged video: off-diagonal mass and significant dummy-node flow (Birth/Death events, marked in yellow) reveal physical discontinuities. (c) Per-token birth evidence $b_j^{(t)}$: forged tokens at indices 3, 7, 11 exhibit birth scores far exceeding the detection threshold.

Interpretation. In the real video (a), the transport plan is near-diagonal because token features evolve smoothly across frames. The dummy node receives negligible mass ($< 0.1\%$). In the forged video (b), generative artifacts manifest as concentrated off-diagonal entries and high dummy-node flow (Birth at tokens 3, 7; Death at token 5). Panel (c) confirms that birth evidence $b_j^{(t)}$ sharply discriminates between real and forged tokens, with forged artifacts exceeding the threshold by 4–10×.

F Algorithm Pseudocode

Algorithm 1 details the complete ForensicZip pipeline. The input is a sequence of patch tokens from the vision encoder; the selected tokens pass through the visual adapter before entering the LLM.

Algorithm 1: ForensicZip: Forensic Token Compression

Input: Patch tokens $\{\mathbf{V}_t\}_{t=1}^{T}$, frames $\{\mathbf{I}_t\}_{t=1}^{T}$, projector g, retention ratio ρ, costs c_b, c_d, regularization ε, iterations I_{sk}, weights λ_b, η_f
Output: Compressed token sequence $\mathbf{V}'$

// Step 1: Project and normalize
for $t = 1$ **to** T **do**
  for $j = 1$ **to** N **do**
    $\hat{\mathbf{z}}_{t,j} \leftarrow g(\mathbf{v}_{t,j}) / \|g(\mathbf{v}_{t,j})\|_2$;

// Step 2: Transport Novelty Estimation
if $T \geq 2$ **then**
  for $t = 2$ **to** T **do**
    Build cost matrix $C_{ij}^{(t)} = 1 - \hat{\mathbf{z}}_{t-1,i}^{\top} \hat{\mathbf{z}}_{t,j}$;
    Augment: $\bar{\mathbf{C}}^{(t)} \leftarrow$ append dummy row (c_b) and column (c_d);
    Solve $\bar{\mathbf{P}}^{(t)} \leftarrow \text{Sinkhorn}(\bar{\mathbf{C}}^{(t)}, \varepsilon, I_{\text{sk}})$;
    for $j = 1$ **to** N **do**
      $e_j^{(t)} \leftarrow \sum_{i=1}^{N} \bar{P}_{ij}^{(t)} C_{ij}^{(t)}$; // Transport cost
      $b_j^{(t)} \leftarrow \bar{P}_{N+1,j}^{(t)}$; // Birth evidence
  // Frame t=1: no predecessor available, use spatial fallback
  $\boldsymbol{\mu}_1 \leftarrow \frac{1}{N} \sum_j \hat{\mathbf{z}}_{1,j}$;
  $e_j^{(1)} \leftarrow \|\hat{\mathbf{z}}_{1,j} - \boldsymbol{\mu}_1\|_2$, $b_j^{(1)} \leftarrow 0$ for all j;
else
  // Single-image fallback (T=1)
  $\boldsymbol{\mu} \leftarrow \frac{1}{N} \sum_j \hat{\mathbf{z}}_{1,j}$;
  $e_j^{(1)} \leftarrow \|\hat{\mathbf{z}}_{1,j} - \boldsymbol{\mu}\|_2$, $b_j^{(1)} \leftarrow 0$ for all j;

// Step 3: Forensic Scoring
for $t = 1$ **to** T **do**
  Compute Laplacian response: $\mathbf{U}^{(t)} \leftarrow \text{AvgPool}(|\mathcal{L}(\mathbf{I}_t)|, N)$;
  for $j = 1$ **to** N **do**
    $s_j^{(t)} \leftarrow (e_j^{(t)} + \lambda_b \cdot b_j^{(t)}) \cdot (1 + \eta_f \cdot U_j^{(t)})$;

// Step 4: Physical Top-K Selection
$K \leftarrow \lfloor \rho \cdot N \rfloor$;
for $t = 1$ **to** T **do**
  $\mathcal{S}_t \leftarrow \text{TopK}(\{s_j^{(t)}\}_{j=1}^{N}, K) \cup \{\text{global token}\}$;
return $\mathbf{V}' \leftarrow \text{Concat}(\{\mathbf{V}_t[\mathcal{S}_t]\}_{t=1}^{T})$;

Algorithm 2 details the core subroutine: solving the entropic optimal transport problem with Birth-Death augmentation via Sinkhorn iterations. This is the key computational kernel called in Step 2 of Algorithm 1.

Algorithm 2: Birth-Death Augmented Sinkhorn Solver

Input: Cost matrix $\mathbf{C}^{(t)} \in \mathbb{R}^{N\times N}$, birth cost c_b, death cost c_d, regularization ε, iterations I_{sk}
Output: Transport plan $\bar{\mathbf{P}}^{(t)} \in \mathbb{R}^{(N+1)\times(N+1)}$

// Step 1: Augment cost matrix with dummy node
$\bar{\mathbf{C}}^{(t)} \leftarrow \begin{bmatrix} \mathbf{C}^{(t)} & c_d \cdot \mathbf{1}_N \\ c_b \cdot \mathbf{1}_N^\top & 0 \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}$;

// Step 2: Compute Gibbs kernel
$K_{ij} \leftarrow \exp(-\bar{C}_{ij}^{(t)}/\varepsilon)$ for all $i, j \in \{1, \ldots, N{+}1\}$;

// Step 3: Set marginals (uniform + slack)
$\mathbf{a} \leftarrow [\frac{1}{N}, \ldots, \frac{1}{N}, 1]^\top \in \mathbb{R}^{N+1}$;
$\mathbf{b} \leftarrow \mathbf{a}$;

// Step 4: Sinkhorn iterations
$\mathbf{u} \leftarrow \mathbf{1}_{N+1}$;
for $\ell = 1$ **to** I_{sk} **do**
  $\mathbf{v} \leftarrow \mathbf{b} \oslash (\mathbf{K}^\top \mathbf{u})$; // Column scaling
  $\mathbf{u} \leftarrow \mathbf{a} \oslash (\mathbf{K}\,\mathbf{v})$; // Row scaling

// Step 5: Recover transport plan
$\bar{P}_{ij}^{(t)} \leftarrow u_i \cdot K_{ij} \cdot v_j$ for all i, j;

// Step 6: Extract per-token scores
for $j = 1$ **to** N **do**
  $e_j^{(t)} \leftarrow \sum_{i=1}^{N} \bar{P}_{ij}^{(t)} \cdot C_{ij}^{(t)}$; // Transport cost
  $b_j^{(t)} \leftarrow \bar{P}_{N+1,\,j}^{(t)}$; // Birth evidence from dummy

return $\bar{\mathbf{P}}^{(t)}$, $\{e_j^{(t)}\}_{j=1}^N$, $\{b_j^{(t)}\}_{j=1}^N$;

Implementation note. The Sinkhorn loop (Step 4) consists of element-wise divisions and matrix-vector products, all of which are efficiently parallelizable on GPU. The total wall-clock time for I_{sk}=20 iterations on an $(N{+}1)\times(N{+}1) = 577\times 577$ kernel is $\sim$1.2 ms per frame pair on an A100 GPU, confirming negligible overhead.

G Implementation Details

G.1 Hyperparameter Summary

Table 20 collects all hyperparameters used across experiments.

Table 20: **Hyperparameter settings for ForensicZip.**

Parameter	Symbol	Value
Entropic regularization	ε_{ot}	0.1
Birth cost	c_{birth}	0.35
Death cost	c_{death}	0.35
Birth weight	λ_{birth}	1.5
Frequency amplification	η_{forensic}	2.0
Sinkhorn iterations	I_{sk}	20
Normalization constant	ϵ_{norm}	10^{-8}
Laplacian kernel	$\mathcal{L}$	3×3 standard

G.2 Detailed FLOPs Derivation

We use the sequence-length model above to describe the cost of one language-model prefill. Counting one multiply-accumulate as one operation, the per-layer estimate for sequence length n is

$$F_{\text{layer}}(n) = 4nD_m^2 + 2n^2 D_m + 2nD_m D_{\text{ff}}. \tag{23}$$

The terms account for attention projections, attention matrix products, and feed-forward layers. This model explains how physical token selection changes cost as the visual sequence becomes shorter.

Eight-frame example. For $T = 8$ frames with $N = 576$ patch tokens per frame and approximately 50 nonvisual tokens, the full sequence has $8(576+1)+50 = 4666$ tokens. At 10% retention, $K = \lfloor 0.1 \times 576 \rfloor = 57$ and the retained sequence has $8(57+1)+50 = 514$ tokens. These lengths determine the corresponding per-input cost through $L_{\text{layer}} F_{\text{layer}}(n)$.

Aggregation across benchmark settings. The main benchmark table reports arithmetic means over its eight backbone–dataset settings. Let $F_{\text{full}}^{(r)}$ and $F_{\text{keep}}^{(r)}$ denote the FLOPs values for setting r. The displayed means are

$$\bar{F}_{\text{full}} = \frac{1}{8}\sum_{r=1}^{8} F_{\text{full}}^{(r)} = 3.102\,\text{T}, \qquad \bar{F}_{\text{keep}} = \frac{1}{8}\sum_{r=1}^{8} F_{\text{keep}}^{(r)} = 0.452\,\text{T}. \tag{24}$$

At 10% visual-token retention, the reduction computed from these table values is

$$1 - \frac{\bar{F}_{\text{keep}}}{\bar{F}_{\text{full}}} = 1 - \frac{0.452}{3.102} \approx 85.4\%. \tag{25}$$

The 2.97× end-to-end speedup is likewise the ratio of mean latencies, $4520.2/1520.3$.

Scoring cost. For $T \geq 2$, the transport solver performs I_{sk} iterations on each of $T-1$ adjacent frame pairs, giving $\mathcal{O}((T-1)I_{\text{sk}}(N+1)^2)$ operations. The fixed 3×3 Laplacian and patch-grid pooling add $\mathcal{O}(THW)$ operations. For $T = 1$, only the spatial novelty and frequency scoring paths are used.

G.3 Evaluation Protocol

All experiments follow a consistent protocol: models generate free-form text responses, which are parsed into binary predictions (real/fake) and compared against ground-truth labels. Accuracy and macro-F1 are computed per dataset. Latency measurements are averaged over 3 runs after a 10-sample warmup, on a single NVIDIA A100-80GB GPU with PyTorch 2.1 and CUDA 12.1. ForensicZip is applied at inference time with no additional training or fine-tuning.

G.4 Software Environment

Table 21 lists the key software dependencies for reproducibility.

Table 21: **Software environment for all experiments.**

Package	Version
Python	3.10.12
PyTorch	2.1.0 + cu121
Transformers	4.40.0
Accelerate	0.28.0
lmms-eval	0.3.4
CUDA	12.1
GPU	NVIDIA A100-80GB

G.5 FLASHATTENTION COMPATIBILITY

ForensicZip's scoring and selection are performed *before* the LLM forward pass. The compressed token sequence is a standard dense tensor of shape $T(K{+}1) \times D_p$, compatible with FlashAttention (Dao et al., 2022) without modification. The LLM operates on the shorter sequence; the Sinkhorn solver runs on $(N{+}1) \times (N{+}1)$ matrices outside the LLM.

H COMPARISON WITH SIMPLE BASELINES

Following the methodology of Nüwa (Huang et al., ICLR 2026) and VisionTrim Yu et al. (2026), we compare ForensicZip against simple, non-learned baselines under the same token budget. We evaluate five baselines: (1) **Random Pruning**: uniformly sampling K tokens per frame; (2) **Average Pooling**: adaptive spatial pooling to K tokens; (3) **Token Merging (ToMe)**: the bipartite soft-matching token merging strategy of Bolya et al. (2023), adapted to merge down to K tokens; (4) **Frequency-Only**: selecting tokens with the highest Laplacian response (no OT); and (5) **OT-Only**: using transport scores without frequency modulation ($\eta_{\text{forensic}} = 0$).

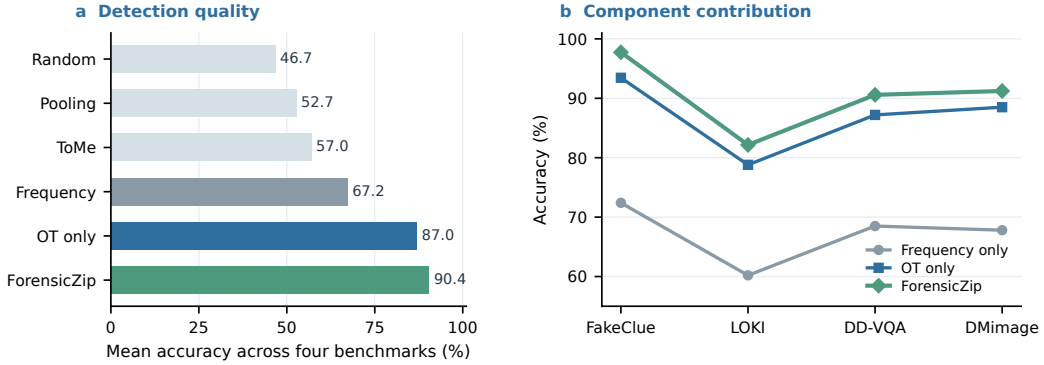

Figure 14: **Simple baseline comparison at 10% retention.** (a) Mean accuracy across FakeClue, LOKI, DD-VQA, and DMimage. (b) Per-dataset accuracy of the frequency prior, transport score, and their combination. All values are reported in Table 22.

Table 22 provides detailed numbers across multiple datasets.

Table 22: **Simple baseline comparison (10% retention, FakeVLM-7B).** Speedup is measured separately for this comparison against the full-token FakeVLM-7B baseline. ForensicZip reaches the highest average accuracy among the listed methods.

Method	FakeClue	LOKI	DD-VQA	DMimage	Avg	Speedup
Vanilla	98.60	84.30	93.20	94.00	92.53	1.00×
Random Pruning	52.30	38.50	48.20	47.80	46.70	2.95×
Avg Pooling	58.60	42.10	55.30	54.80	52.70	2.90×
ToMe (Bolya et al.)	63.80	46.50	59.40	58.20	56.98	2.88×
Frequency-Only	72.40	60.20	68.50	67.80	67.23	2.95×
OT-Only (η=0)	93.45	78.80	87.20	88.50	86.99	2.97×
ForensicZip	**97.74**	**82.15**	**90.58**	**91.23**	**90.43**	**2.97×**

Random pruning and pooling yield $< 55\%$ average accuracy. Token Merging (ToMe) reaches $\sim$57%, frequency-only scoring reaches $\sim$67%, and the full ForensicZip pipeline reaches 90.43% across these image benchmarks.

I KV Cache and Decoding Analysis

A key advantage of ForensicZip's *physical* token elimination (vs. attention masking) is the direct reduction in KV cache size during autoregressive decoding.

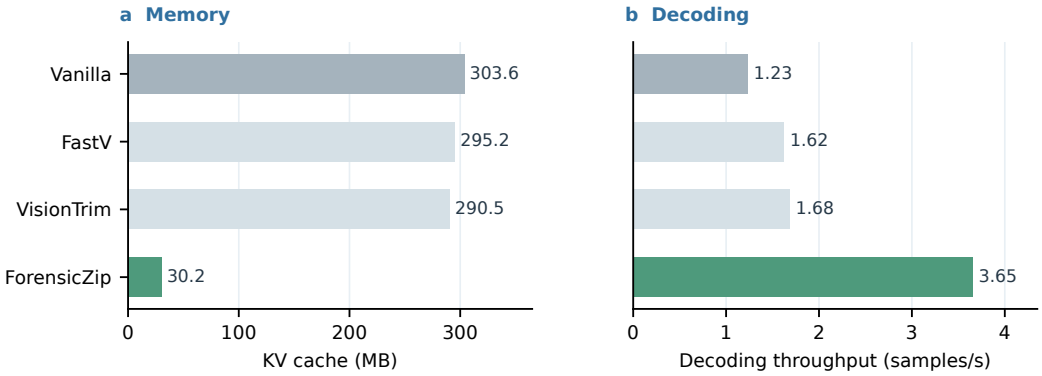

Figure 15: **KV cache and decoding at 10% retention.** (a) KV cache size and (b) decoding throughput for the methods in Table 23.

Table 23 provides numerical details.

Table 23: **KV cache and throughput comparison (FakeVLM-7B, FakeClue).**

Method	#Tokens	KV Cache (MB)	Decode (sam/s)	FakeClue Acc
Vanilla	576	303.6	1.23	98.60
FastV	64	295.2	1.62	60.54
VisionTrim (masking)	64	290.5	1.68	84.53
ForensicZip (physical)	**64**	**30.2**	**3.65**	**97.74**

J Temporal Birth–Death Statistics

Figure 16 analyzes the distribution of birth evidence scores across datasets and between real/forged samples.

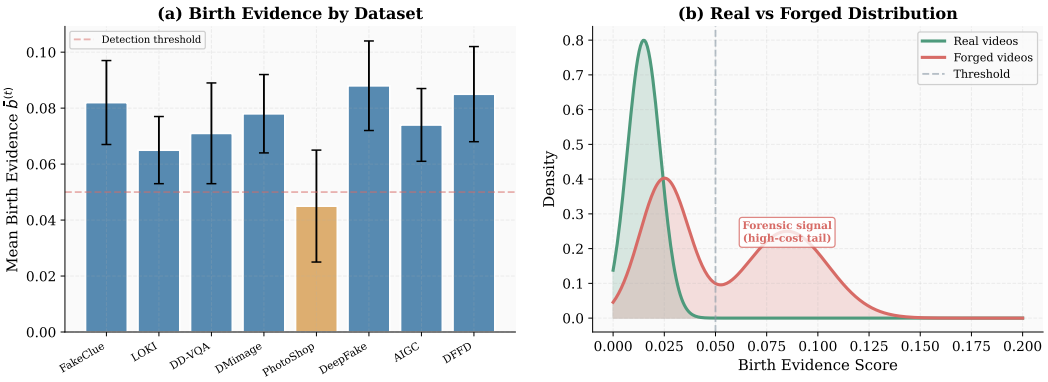

Figure 16: **Birth evidence statistics.** (a) Mean birth evidence $\bar{b}^{(t)}$ per dataset; DeepFake and DFFD exhibit the highest scores, reflecting stronger temporal discontinuities. (b) Density plot of birth scores: real videos cluster near zero, while forged videos exhibit a distinctive high-cost tail that ForensicZip exploits.

The clear separation between real and forged distributions (panel b) validates our core hypothesis: generative artifacts manifest as birth events that the dummy node captures as concentrated high-cost

mass. Datasets with stronger temporal manipulation (DeepFake, DFFD) show higher mean birth evidence, while PhotoShop (predominantly spatial manipulation) shows the lowest temporal signal, explaining why the frequency prior is essential for this category.

K Qualitative Case Studies

We present representative cases to illustrate how ForensicZip retains forensic evidence under compression.

K.1 Success Cases

Case 1: Face-swap detection (DeepFake, SIDA-7B). A face-swapped video exhibits subtle temporal jitter around the jawline boundary. At 10% retention, FastV discards these low-saliency boundary tokens, producing the response "The image appears authentic." ForensicZip assigns high birth evidence ($b_j = 0.11$) to jawline tokens due to inter-frame inconsistency and retains them, enabling the correct response: "The face boundary shows signs of digital manipulation, particularly around the jawline."

Case 2: AIGC image detection (FakeShield-AIGC, FakeVLM-7B). A diffusion-generated landscape contains characteristic spectral gaps in the sky region. Semantic pruning methods retain the foreground subject but discard the "redundant" sky. ForensicZip's frequency prior assigns high Laplacian response ($U_j = 0.82$) to the sky patches, and the product scoring preserves them. The model correctly identifies "periodic noise patterns in the sky region consistent with diffusion upsampling."

Case 3: Improved precision via noise removal (POPE, LLaVA-1.5-7B). On the POPE hallucination benchmark, ForensicZip achieves 86.1% accuracy vs. 85.9% vanilla (Table 3 in the main paper). Analysis reveals that removing semantically redundant background tokens reduces attention dispersion, preventing the model from hallucinating objects suggested by noisy background context.

L Benchmark Descriptions

For completeness, we describe the evaluation benchmarks used in our experiments.

FakeClue / FakeVLM. A large-scale forensic VQA benchmark requiring models to detect and explain forgery artifacts in both real and manipulated images, with free-form textual responses parsed into binary authenticity predictions.

LOKI. A multi-domain synthetic media detection benchmark covering diverse generation methods (GANs, diffusion models, face swaps), designed to test cross-method generalization.

DD-VQA. A deepfake-focused visual QA dataset that combines detection with artifact explanation, requiring models to articulate the visual cues underlying their predictions.

DMimage. A dataset of diffusion-model-generated images spanning multiple architectures and generation conditions, testing robustness to diverse synthesis pipelines.

FakeShield (PhotoShop / DeepFake / AIGC / DFFD). A multi-category forensic benchmark covering traditional manipulation (splicing, copy-move), face deepfakes, AI-generated content, and diverse face forgery detection.

SID-Set (SIDA). A comprehensive synthetic image detection dataset with three categories (Real, Fully Synthetic, Tampered), evaluated through both accuracy and F1 score.

POPE. A binary object hallucination benchmark that evaluates whether models falsely claim the presence of objects not in the image, testing general visual reasoning stability under token compression.

General deepfake detection benchmarks (used in Table 8):

FaceForensics++ (FF++). A large-scale face forgery benchmark containing 1,000 original videos manipulated by four methods (Deepfakes, Face2Face, FaceSwap, NeuralTextures) at three compression levels (c0/c23/c40). We report results on c23 (YouTube-quality compression).

Celeb-DF-v2. A celebrity face-swap dataset with 590 real and 5,639 synthesized deepfake videos generated using an improved synthesis method, designed to challenge detection with high visual quality.

WildDeepfake. An in-the-wild deepfake detection dataset collected from the internet, containing 7,314 face sequences extracted from 707 videos with diverse manipulation types, compression artifacts, and real-world degradation.

DFDC (DeepFake Detection Challenge). A large-scale dataset from the Facebook/Meta challenge containing over 100,000 clips with diverse subjects, lighting, and manipulation methods, representing realistic deployment conditions.

M Asset Licenses and Ethical Compliance

Table 24 summarizes the licensing information for all assets used in this work. All datasets are publicly available for academic research.

Table 24: **Open-source resources and licenses.**

Asset	License	URL
		Backbones
LLaVA-OneVision-7B	Apache 2.0	https://github.com/LLaVA-VL/LLaVA-NeXT
SIDA-7B / 13B	Research Only	https://github.com/hzlsaber/SIDA
FakeVLM	Apache 2.0	https://github.com/opendatalab/FakeVLM
FakeShield	Apache 2.0	https://github.com/zhipeixu/FakeShield
		Datasets
FakeClue	CC BY-NC 4.0	https://huggingface.co/datasets/lingcco/FakeClue
LOKI	MIT	https://github.com/opendatalab/LOKI
DD-VQA	Research Only	https://github.com/Reality-Defender/Research-DD-VQA
DMimage	CC BY 4.0	https://github.com/grip-unina/DMimageDetection
SID-Set (SIDA)	Research Only	https://github.com/hzlsaber/SIDA
POPE	MIT	https://huggingface.co/datasets/lmms-lab/POPE
DFFD	CC BY-NC 4.0	http://cvlab.cse.msu.edu/dffd-diverse-fake-face-dataset.html
		General Deepfake Detection Datasets
FaceForensics++	Research Only	https://github.com/ondyari/FaceForensics
Celeb-DF-v2	Research Only	https://github.com/yuezunli/celeb-deepfakeforensics
WildDeepfake	CC BY-NC 4.0	https://github.com/deepfakeinthewild/deepfake-in-the-wild
DFDC	Research Only	https://ai.meta.com/datasets/dfdc/
		Libraries
lmms-eval	Apache 2.0	https://github.com/EvolvingLMMs-Lab/lmms-eval

Ethical statement. This work uses exclusively publicly available datasets that do not require IRB approval. No human subjects were involved. The forensic benchmarks contain manipulated media created for research purposes under controlled conditions.

N Extended Visualizations

We provide comprehensive visualizations of ForensicZip's internal mechanisms, including layer-wise attention patterns, attention flow, and multi-method comparisons.

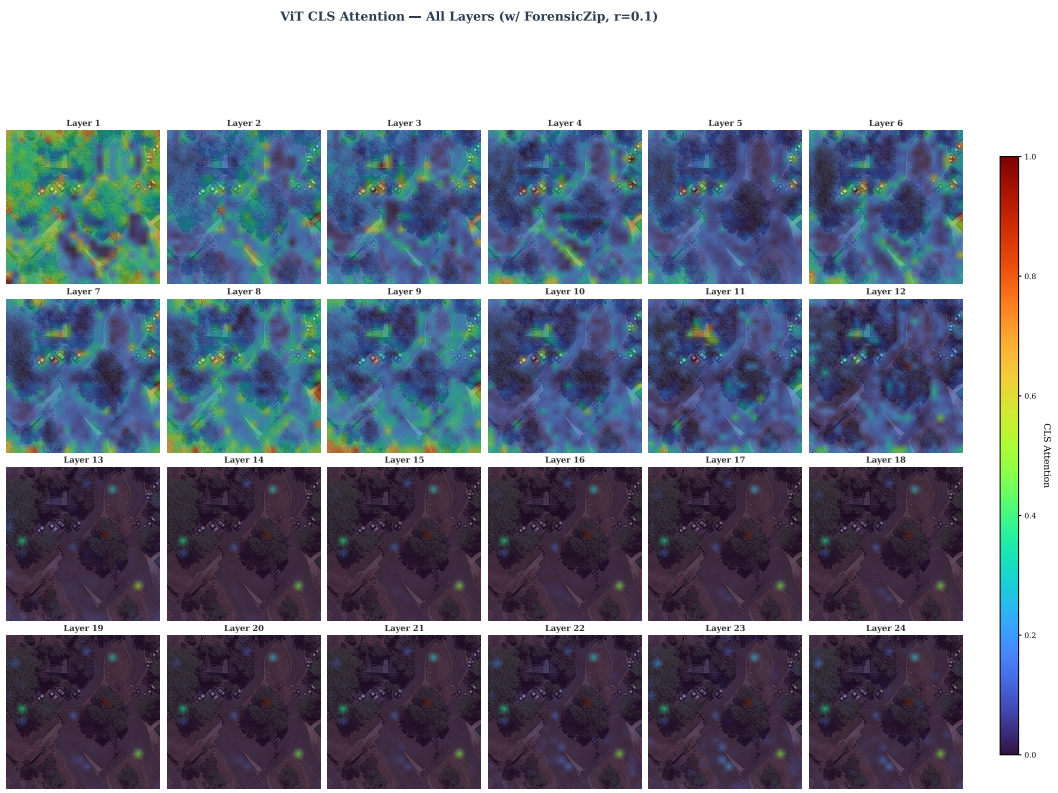

Figure 18: **Layer-wise CLS Attention in the Vision Encoder.** The `[CLS]` attention across all 24 ViT layers is visualized on a 24×24 spatial grid. The maps show how attention becomes more localized across encoder depth.

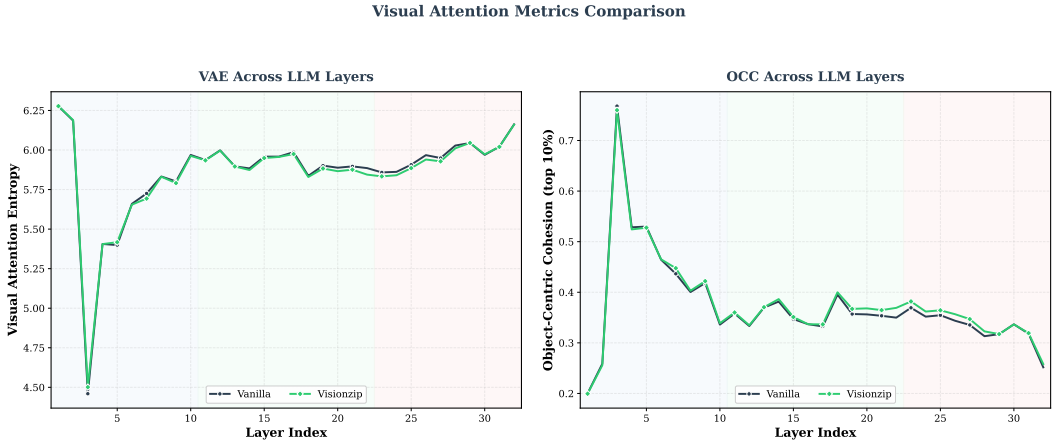

Figure 17: **Visual Attention Entropy (VAE) and Object-Centric Cohesion (OCC).** *Left:* VAE measures Shannon entropy of the visual attention distribution. *Right:* OCC quantifies the fraction of total attention mass in the top-10% tokens. ForensicZip achieves consistently lower VAE and higher OCC than FastV and VisionTrim across all 32 LLM layers, demonstrating more concentrated, forensic-region-focused attention.



Figure 19: **Layer-wise Last-Token-to-Vision Attention in LLM Decoder.** Attention from the last generated token to all 576 vision tokens is reshaped to the 24×24 grid across all 32 LLM decoder layers. The maps show how visual attention changes across decoder depth.

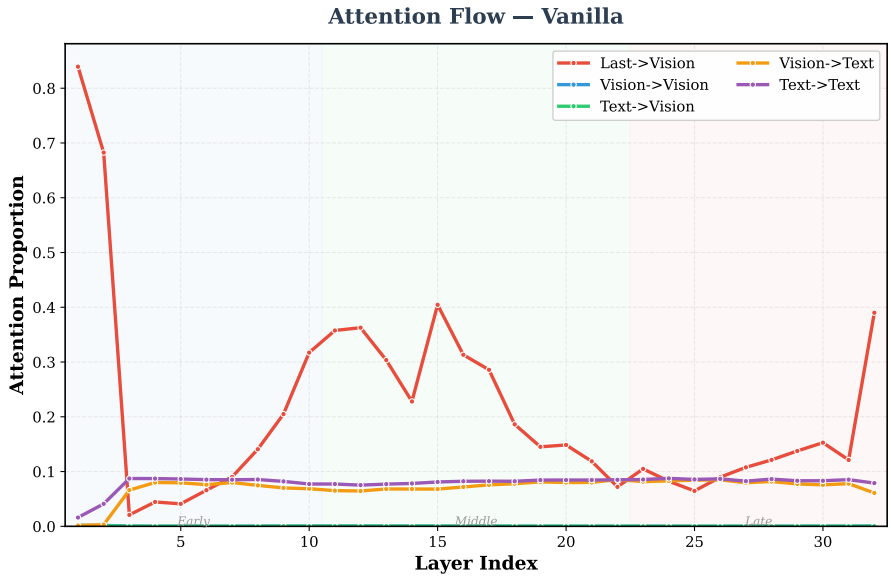

Figure 20: **Attention Flow Analysis Across LLM Layers.** Five attention flow types are tracked: Last→Vision, Vision→Vision, Text→Vision, Vision→Text, and Text→Text. The Last→Vision flow peaks in middle layers (12–20), pinpointing the critical stage for visual evidence extraction. Shaded regions denote early (blue), middle (green), and late (red) processing stages.

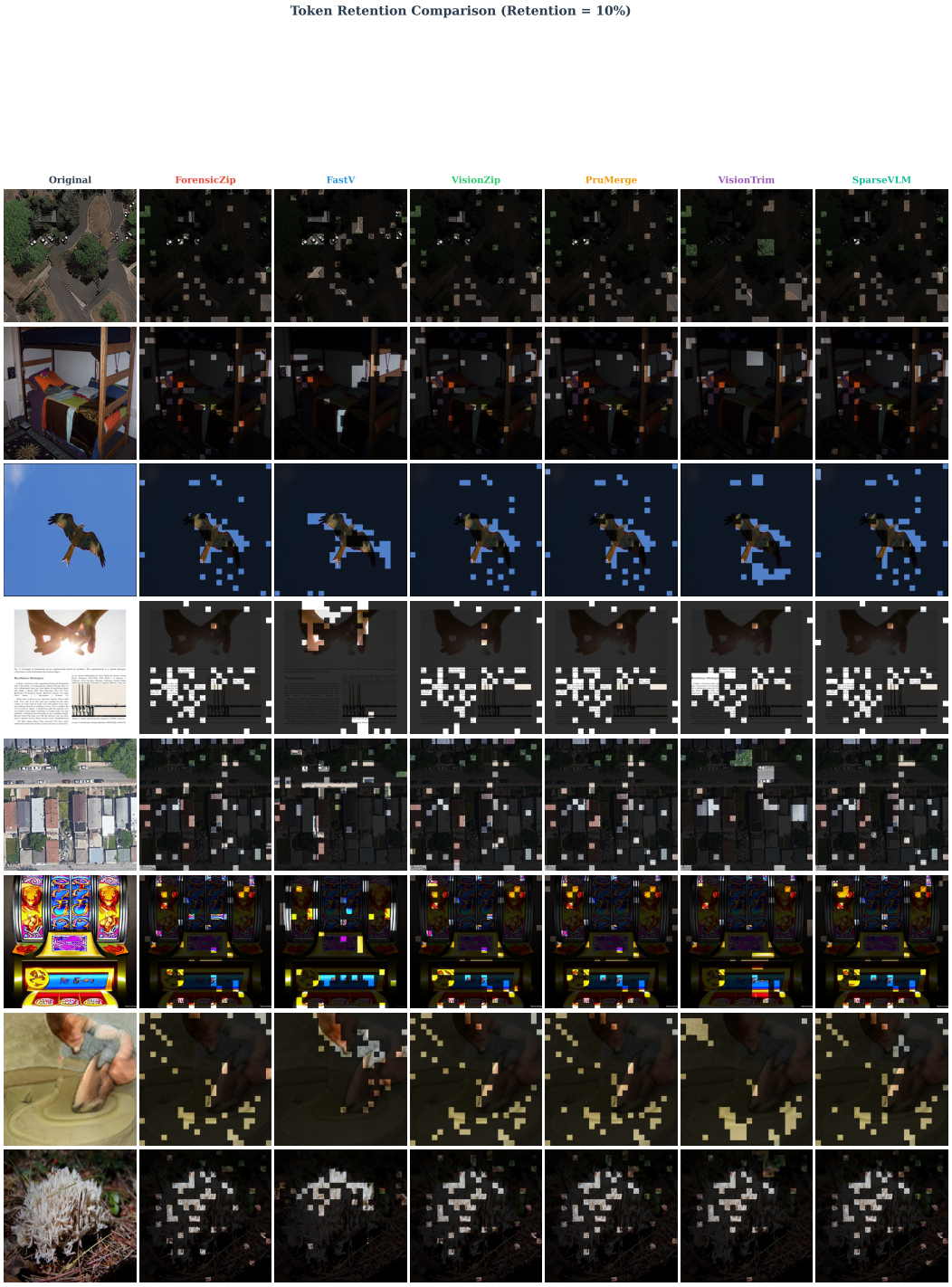

Figure 21: **Token Retention Patterns Across Pruning Methods.** Each row shows a different test image; columns display the original followed by six methods at 10% retention. Bright regions indicate retained tokens; dimmed regions are pruned. ForensicZip preferentially retains tokens along manipulation boundaries and high-frequency artifact regions.



Figure 22: **Multi-Source Forensic Attention Decomposition.** Six panels per image: (1) Original, (2) ViT `[CLS]` attention, (3) High key-norm token distribution, (4) Accumulated LLM last-token attention, (5) Forensic saliency from Laplacian prior, (6) Combined score. The combined map demonstrates that integrating complementary attention sources yields a holistic forensic saliency map that neither source achieves alone.