Appendix — From Binary Labels to Temporal Boundaries: EM and Constraint-Guided Weakly Supervised Forgery Localization

Xiaodong Zhu[1] Yuanming Zheng[1] Suting Wang[1] Junqi Yang[1]
Yuhong Yang[1,*] Weiping Tu[1] Zhongyuan Wang[1]
[1]NERCMS, School of Computer Science, Wuhan University
{xiaodongzhu, ameixa, wangsuting, yangjq, yangyuhong, tuweiping}@whu.edu.cn
wzy_hope@163.com

Contents

A. Detailed Description of Datasets

We conduct experiments on temporal forgery localizationusing two multimodal Deepfake datasets: LAV-DF [1] and AV-Deepfake1M [2].

LAV-DF [1] is a content-centric audio–visual Deepfake dataset specifically designed for temporal forgery localization. It consists of 136,304 videos (36,431 real and 99,873 fake) spanning 153 identities, with a total duration of approximately 325.5 hours. Unlike large-scale identity-swap datasets, LAV-DF focuses on fine-grained, segment-level manipulations: forged segments last only 0.650 seconds on average, while videos have an average duration of 8.579 seconds, making the localization task substantially more challenging than simple classification. The dataset uses only replacement-based manipulations. Audio forgeries are generated with SV2TTS [4], and visual forgeries with Wav2Lip [6]. Although these methods were once serviceable, their output quality is now well below modern synthesis standards, which inadvertently raises the difficulty for state-of-the-art detectors. Furthermore, LAV-DF relies on a rule-based antonym replacement scheme during transcript manipulation. This approach tends to introduce contextual inconsistencies and limited content diversity, often making the fake content feel artificially constrained.

AV-Deepfake1M [2] represents a more ambitious effort, providing a large-scale multimodal Deepfake dataset with 1,886 hours of video from 2,068 identities, captured across diverse real-world conditions. Beyond the replacement-style manipulations used in LAV-DF, AV-Deepfake1M incorporates insertion and deletion forgeries, enabling research on more complex and localized temporal manipulations. Forged segments are even shorter—only 0.326 seconds on average—while videos average 9.072 seconds, meaning forged content constitutes a much smaller proportion of the data. As a result, localization models cannot rely on extended manipulation intervals and must handle subtler, more realistic forgeries. The dataset leverages large language models (e.g., ChatGPT) to produce semantically coherent and diverse textual content. Visual forgeries are generated with TalkLip [7], and audio with VITS [5] and YourTTS [3], representing state-of-the-art open-source generation techniques. In contrast to LAV-DF's simplistic antonym-based manipulations, AV-Deepfake1M offers context-consistent, diverse, and higher-quality forgeries, forming a more realistic and demanding benchmark for modern multimodal Deepfake localization.

A detailed comparison of these datasets is provided in Table 1.

B. Solution of Temporal Consistency Refinement

We consider the linearly constrained Bregman optimization problem shown in Equation 1:

This objective enforces two consistency constraints: Each Q_t forms a valid categorical distribution, and the attention-weighted mean of the frame-level predictions aligns with the clip-level prior q.

*Corresponding author.

Dataset	Year	Manipulated Method	#Subjects	#Real	#Fake	#Total	Average Fake Length (sec)	Average Video Length (sec)
LAV-DF	2023	Content-driven RE/TTS	153	36,431	99,873	136,304	0.650	8.597
AV-Deepfake1M	2024	Content-driven RE/TTS	2,068	286,721	860,039	1,146,760	0.326	9.072

Table 1. Details for temporal forgery localization datasets. RE: Face reenactment, TTS: Text-to-speech

$$\min_{Q} \quad \sum_{t=1}^{T} \mathrm{KL}(Q_t||S_t) = \sum_{t=1}^{T}\sum_{c=0}^{m} Q_{t,c} \log \frac{Q_{t,c}}{S_{t,c}},$$
$$\text{s.t.} \quad Q_{t,c} \geq 0, \sum_{c=0}^{m} Q_{t,c} = 1, \forall t, \tag{1}$$
$$\frac{1}{T}\sum_{t=1}^{T} A_t Q_{t,c} = q_c, \forall c \in \mathcal{C} \setminus \{0\}.$$

We solve this problem using an Iterative Proportional Scaling (IPS) algorithm, which is training-free, has linear-time complexity, and converges to the unique minimizer of the KL divergence under these constraints. The detailed procedure is summarized in Algorithm 1, which alternates between row normalization (enforcing probability constraints) and column normalization (enforcing marginal alignment). The algorithm converges rapidly in practice, typically within $10 \sim 20$ iterations.

Algorithm 1: Iterative Proportional Scaling for Temporal Consistency Optimization

Input: $S \in \mathbb{R}^{T\times(m+1)}$, clip-level attribute $q \in \mathbb{R}^{m+1}$, maximum iterations K, latent attribute set $\mathcal{C}$, tolerance ϵ
Output: $Q^* \in \mathbb{R}^{T\times(m+1)}$

Initialize $Q \leftarrow S$, $Q_{\text{prev}} \leftarrow 0$;
for $k = 1$ **to** K **do**
  /* Row scaling: normalize each frame to sum to 1 */
  $Q_{t,c} \leftarrow \frac{Q_{t,c}}{\sum_{c'=0}^{m} Q_{t,c'}}, \quad \forall t;$
  /* Column scaling: match clip-level distribution q */
  $\hat{q}_c \leftarrow \frac{1}{T}\sum_{t=1}^{T} A_t Q_{t,c}, \quad \forall c \in \mathcal{C} \setminus \{0\};$
  $Q_{t,c} \leftarrow Q_{t,c} \cdot \frac{q_c}{\hat{q}_c}, \quad \forall t, \forall c \in \mathcal{C} \setminus \{0\};$
  /* Convergence check */
  if $\|Q - Q_{prev}\|_F < \epsilon$ **then**
    break
  $Q_{\text{prev}} \leftarrow Q;$
$Q^* \leftarrow Q;$

C. Closed-Form Solution of Graph Diffusion

We start from the recurrence relation that defines the confidence diffusion process:

$$\omega^{(t+1)} = \beta \mathcal{T}\, \omega^{(t)} + (1-\beta)\, \omega^{(0)}, \tag{2}$$

where $\mathcal{T}$ is the row-normalized transition matrix of the proposal graph and $\beta \in (0,1)$ controls the diffusion strength. This process aggregates supportive evidence from neighboring proposals while maintaining consistency with the initial confidence vector $\omega^{(0)}$.

Unfolding the recurrence iteratively yields:

$$\begin{aligned}
\omega^{(1)} &= \beta \mathcal{T}\, \omega^{(0)} + (1-\beta)\, \omega^{(0)}, \\
\omega^{(2)} &= \beta \mathcal{T}\, \omega^{(1)} + (1-\beta)\, \omega^{(0)}, \\
\omega^{(3)} &= (\beta\mathcal{T})^2 \omega^{(0)} + (1-\beta)\sum_{k=0}^{1} (\beta\mathcal{T})^k \omega^{(0)}, \\
&\vdots \\
\omega^{(t)} &= (\beta\mathcal{T})^t \omega^{(0)} + (1-\beta)\sum_{k=0}^{t-1} (\beta\mathcal{T})^k \omega^{(0)}.
\end{aligned} \tag{3}$$

When $t \to \infty$, the limit exists if the power series $\sum_{k=0}^{\infty}(\beta\mathcal{T})^k$ converges. The convergence condition is determined by the spectral radius $\rho(\beta\mathcal{T})$. Because $\mathcal{T}$ is row-stochastic, we have $\rho(\mathcal{T}) = 1$, and therefore $\rho(\beta\mathcal{T}) = \beta < 1$, ensuring that the series converges absolutely. Applying the Neumann series identity for a convergent matrix geometric series gives:

$$\sum_{k=0}^{\infty} (\beta\mathcal{T})^k = (I - \beta\mathcal{T})^{-1}, \tag{4}$$

Substituting this into the limit form of $\omega^{(t)}$ leads to the closed-form equilibrium solution:

$$\omega^* = (1-\beta)(I - \beta\mathcal{T})^{-1}\omega^{(0)}. \tag{5}$$

This closed-form solution provides an analytical interpretation of the diffusion process: the operator $(I - \beta\mathcal{T})^{-1}$ acts as a global propagation kernel that aggregates confidence scores across all connected proposals. A smaller β restricts the influence to local neighborhoods, while a larger β promotes long-range propagation, approaching a global equilibrium. Since $(I - \beta\mathcal{T})$ is non-singular for $\beta < 1$, the solution is unique and numerically stable.

D. Additional Experiments

D.1. Impact of Attribute Prior Initialization

#	Initialization methods	mAP (0.1:0.7)	mAR (0.1:0.7)
1	random distribution	42.4	59.6
2	gaussian distribution	42.5	59.6
3	uniform distribution	42.7	59.8

Table 2. The impact of different initialization methods for the attribute prior π_c on TFL performance. The best mAP/mAR scores are highlighted in red, and the second-best scores in blue.

In this section, we investigate the impact of different initialization methods for the attribute prior on the localization results. As shown in Table 2, we initialize the attribute prior π_c with a random distribution (Line 1), a Gaussian distribution (Line 2), and a uniform distribution (Line 3). The results in the table indicate that the uniform distribution achieves the best mAP and mAR scores. However, the differences between the uniform distribution and the other two initialization methods are minimal. This suggests that regardless of the prior initialization, the EM optimization process during LAD enables the model to learn the best way to differentiate between various fake attributes.

D.2. Impact of EMA Update Weights

Our method GEM-TFL involves two Exponential Moving Average (EMA) update coefficients during the classification phase: 1) the coefficient δ used in the LAD module to update the attribute prior π_c, and 2) the coefficient β used in the GPR module to update the proposal weights. In this section, we conduct ablation studies on both EMA coefficients. The experimental results are shown in Figure 2(a) and Figure 2(b), respectively.

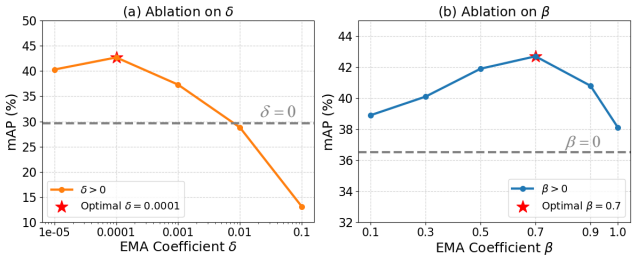

Figure 1. Ablation study of the EMA update coefficient δ in the LAD module and the EMA update coefficient β in the GPR module.

As shown in Figure 2(a), the optimal value of the coefficient δ is 0.0001. When δ is increased to 0.01 or 0.1, the performance drops significantly and even falls below the $\delta = 0$ case, where the attribute prior is not updated. This behavior can be attributed to the fact that a large δ leads to unstable attribute assignment in the early training stage, which may cause dimensional collapse and subsequently distort the attribute prior, forming a detrimental feedback loop. In contrast, smaller values of δ result in a more stable optimization process and effectively avoid such failure modes.

As shown in Figure 2(b), the optimal value of the coefficient β is 0.7. When β is small, the proposal weights are strongly influenced by their initial values (i.e., the OIC scores) during the iterative refinement process. As β increases, this dependency becomes weaker. Notably, any $\beta > 0$ yields approximately a 2% improvement in mAP over the $\beta = 0$ setting. This confirms that leveraging inter-proposal relationships effectively refines the proposal weights, mitigates proposal fragmentation, and consequently improves the quality of the pseudo labels.

D.3. Qualitative Analysis

Figure 2. Visualization of ground truth, UMMAFormer (fully supervised), WMMT and GEM-TFL.

We visualize predicted forgery segments from an AV-Deepfake1M video in Figure 2, comparing the ground truth, fully supervised UMMAFormer, weakly supervised WMMT, and our GEM-TFL (with and without LP). The fully supervised model aligns most closely with the ground truth, benefiting from frame-level labels. WMMT produces scattered and inaccurate proposals due to weak supervision. In contrast, GEM-TFL generates fewer fragmented segments and achieves more accurate localization, validating the effectiveness of our two-stage design and proposed modules.

References

[1] Zhixi Cai, Shreya Ghosh, Abhinav Dhall, Tom Gedeon, Kalin Stefanov, and Munawar Hayat. Glitch in the matrix: A large scale benchmark for content driven audio–visual forgery detection and localization. *Computer Vision and Image Understanding*, 236:103818, 2023. 1

[2] Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, Tom Gedeon, and Kalin Stefanov. Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset. In *ACM MM*, pages 7414–7423, 2024. 1

[3] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot

voice conversion for everyone. In *ICML*, pages 2709–2720. PMLR, 2022. 1

[4] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *NeurIPS*, 31, 2018. 1

[5] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *ICML*, pages 5530–5540. PMLR, 2021. 1

[6] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM MM*, pages 484–492, 2020. 1

[7] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *CVPR*, pages 14653–14662, 2023. 1