Synlab: A Spatio-Temporal Aggregation Framework for Collaborative Visual Misinformation Detection in Videos

Shuning Zhang[1] , **Linzhi Wang**[1] , **Ruokai Zhao**[2] , **Yiqun Xu**[1] , **Yuanyuan Wu**[3] , **Yabo Wang**[1] , **Simin Li**[4] , **Xin Yi**[1*] , **Hewu Li**[1]

[1]Tsinghua University
[2]Oxford University
[3]Shang Hai Jiao Tong University
[4]Beihang University

zhang.sn314@gmail.com, wang-lz22@mails.tsinghua.edu.cn, ruokai.zhao@st-hughs.ox.ac.uk, xuyiqun22@mails.tsinghua.edu.cn, buddy.yuan@sjtu.edu.cn, yb-wang22@mails.tsinghua.edu.cn, lisiminsimon@buaa.edu.cn, yixin@tsinghua.edu.cn, lihewu@cernet.edu.cn

Abstract

Identifying misinformation videos like deepfakes on social media is challenging due to their subtle manipulations across both time and space that are difficult for automated tools to reliably detect. These limitations require approaches that leverage collective user intelligence and detailed human feedback to pinpoint manipulations. Therefore, we introduce Synlab, an algorithmic framework for collaborative video misinformation identification. Synlab enables users to provide detailed spatio-temporal annotations, confidence scores, and textual rationales as structured multimodal input, crucial for complex video content. Our confidence-weighted spatio-temporal Intersection-over-Union (IoU) algorithm synthesizes diverse user annotations by iteratively merging them via spatio-temporal overlap and confidence, while also incorporating user reliability and summarizing rationales using language models to generate trustworthy consensus representations. A 7-day online study demonstrated Synlab achieved superior accuracy (97.20%) over alternative methods (No-Label: 80.80%, No-Agg: 95.80%). Further ablation studies confirmed benefits of confidence-weighting and user historical performance in the aggregation.

1 Introduction

The proliferation of online misinformation, particularly sophisticated video manipulations such as deepfakes, critically threatens informed public discourse [Vosoughi *et al.*, 2018; Wang *et al.*, 2019; Thomas, 2022]. Prevailing countermeasures, predominantly centralized platform governance [Hartwig *et al.*, 2024a], integrate automated detection, human moderation, and fact-checking initiatives [Alrashidi *et al.*, 2022; Arsht and Etcovitch, 2018; Bélair-Gagnon *et al.*, 2023] but face significant limitations.

Specifically, these platform-centric strategies face two fundamental challenges: (C1) limited detection accuracy, with automated tools struggling against novel manipulations and human moderation proving costly and prone to inconsistency; and (C2) suboptimal user experience, where operational opacity can diminish user agency and erode trust [Gorwa *et al.*, 2020]. These challenges highlight the need for complementary, user-empowering approaches for video verification.

Collective user intelligence offers a promising path to enhance both detection accuracy (C1) by aggregating diverse assessments, and user experience (C2) by fostering critical visual literacy and agency through direct involvement [Jahanbakhsh *et al.*, 2021; Pennycook *et al.*, 2021]. However, effectively translating varied spatio-temporal video annotations into reliable assessments is a key hurdle. Existing mechanisms often fail to consolidate conflicting video inputs for accurate collaborative verification [Hlaoua, 2024; Belgacem *et al.*, 2021; Suzuki, 2015], revealing a critical gap.

To address these video-specific challenges, this work introduces Synlab, an algorithmic framework for collaborative video misinformation identification through aggregating visual temporal annotations. Grounded in theories of Collective Intelligence (CI) [Wolpert and Tumer, 1999] and Social Influence (SI) [Cialdini and Goldstein, 2004], Synlab's design integrates interconnected annotation, aggregation and feedback processes. This structure fosters diverse, independent, and motivated user visual input while managing social dynamics.

The Synlab framework operationalizes these principles through a multi-stage algorithmic process (Figure 1). Initially, user provide rich, structured visual annotations. These primarily involve delineating specific visual spatio-temporal regions within the video footage suspected of manipulation and assigning semantic labels, confidence scores (0-100%), and textual rationales. The annotation is facilitated by an intuitive interface mirroring familiar paradigms such as on-screen bullet commenting. Subsequently, Synlab leverages a confidence-weighted 3D Intersection-over-Union (IOU)-based mechanism to address consensus conflict from diverse visual annotations. It iteratively merges annotations based on spatio-temporal overlap (with an IoU threshold empiri-

*Corresponding author.

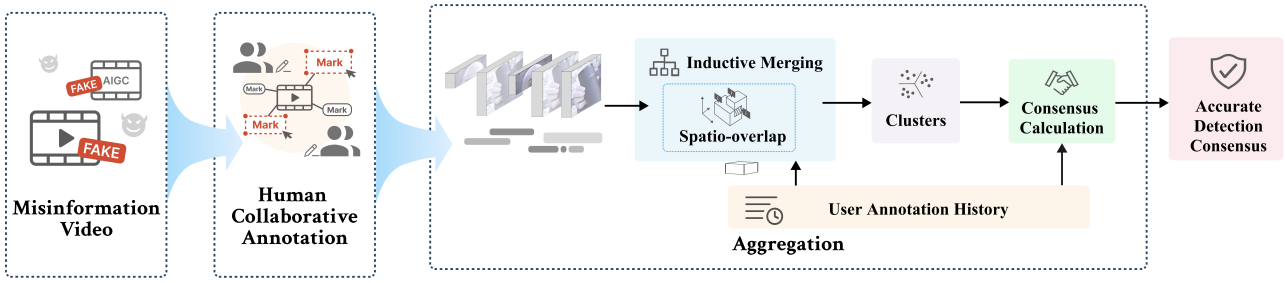

Figure 1: The algorithmic framework of this paper.

cally set to 40%) alongside user-assigned confidence scores. This confidence-weighting is theoretically grounded in the idea that higher confidence likely indicates greater accuracy. Synlab then determines predominant labels for consolidated visual regions by weighting both annotation confidence and user historical performance, thereby prioritizing insights from reliable annotators. It further synthesizes textual rationales using language model embeddings to generate coherent explanations for aggregated assessments. Finally, Synlab presents aggregated visual regions (computed as the weighted average of individual regions), consensus labels, aggregated confidence scores, and synthesized rationales hierarchically via a layered visualization strategy. This strategy uses color-coded overlays for immediate visual cues and enables interactive, hierarchical detail exploration.

To evaluate Synlab, we collected visual spatio-temporal annotations from 48 participants over a 7-day period. This study utilized a custom experimental social platform mimicking Twitter, deploying various algorithmic configurations. Synlab demonstrated superior misinformation identification, achieving an overall accuracy of 97.20%. This significantly surpassed the configuration without peer label demonstration (denoted No-Label, 80.80% accuracy) and another showing individual visual annotations without Synlab's aggregation mechanism (denoted No-Agg, 95.80% accuracy). Furthermore, Synlab rapidly achieved high accuracy even with few visual annotations per item (93.44% with $N = 2$ per user, and 96.11% with $N = 3$ per user), outperforming non-aggregated approaches with similar limited input. These analyses also highlighted the efficacy of incorporating confidence-weighted user historical performance (R_{cw} setting) into the aggregation logic, which yielded the highest accuracy (97.20%) when considering at least three annotations ($N = 3$). **To sum up, this work contributes:**

• A comprehensive visual algorithmic framework for collaborative video misinformation identification, systematically integrating structured visual spatio-temporal annotation.

• A novel confidence-weighted 3D spatio-temporal aggregation algorithm to synthesize a reliable consensus from diverse visual assessments.

• Empirical validation through 7-day user visual annotations, demonstrating Synlab's superior accuracy (97.20%).

2 Background & Related Work

We review foundational literature on misinformation, its social influence and collaborative countermeasures. We then examine diverse identification and mitigation strategies, alongside user interface design principles for collaborative annotation, with a focus on video-specific challenges.

2.1 Identifying and Dealing with Misinformation

Researches have approached misinformation from technological, user-centered and cultural or community-based perspectives. Technological and socio-technical approaches investigate technical flaws and propose misinformation detection algorithms. Fernandez et al. [Fernandez and Alani, 2018] analyzed misinformation technologically and explored socio-technical advancements, while Sharma et al. [Sharma *et al.*, 2019] examined the technical challenges of fake news and dataset characteristics. AI has also been explored for mitigation. Jahanbakhsh et al. [Jahanbakhsh *et al.*, 2023] developed a personalized AI for content assessment prediction. However, technical solutions remain constrained by the sophistication of AI-generated videos, and users are also often reluctant to adopt or trust AI-driven detection systems [Hartwig *et al.*, 2024a]. User-centered interventions empower users in confronting misinformation. Kirchner et al. [Kirchner and Reuter, 2020] conducted a three-step study on user-centered approaches to counter misinformation on social media. Hartwig et al. [Hartwig *et al.*, 2024b] systematized these interventions and developed a taxonomy for intervention design. While AI assisted in processes like algorithmic labeling [Jahanbakhsh *et al.*, 2023; Lu *et al.*, 2022], current methods lack sufficient mechanisms to leverage collaborative intelligence for discerning subtle video manipulations. Cultural or community-based approaches examine cross-cultural susceptibility differences or focus on community-specific digital literacy initiatives. Heuer et al. [Heuer and Glassman, 2022] identified cross-cultural variations in misinformation detection, and Wilner et al. [Wilner *et al.*, 2023] engaged professionals to promote digital literacy in underserved communities. However, they lack specific design explorations for fine-grained spatio-temporal annotation, crucial for accurately identifying manipulations in video content.

2.2 Social Influence of Misinformation

Research focused on designing techniques and crowdsourcing platforms for public participation in content moderation.

Bozarth et al. [Bozarth *et al.*, 2023] studied content moderation workflows on social media platforms, noting that the moderation process focuses not only on content authenticity but also on user intent and potential harm. Jahanbakhsh et al. [Jahanbakhsh *et al.*, 2022] empower users through structured accuracy assessments and customizable content filters. Kaufman et al. [Kaufman *et al.*, 2022] demonstrated that crowdsourcing platforms can effectively assess content authenticity, especially for sensitive topics such as pandemics. However, these existing platforms, primarily designed for text content, lack interfaces and aggregation mechanisms for unique spatio-temporal complexity of video misinformation.

2.3 User Interface Design for Collaborative Annotating

Collaborative annotation techniques, often leveraged in crowdsourcing, have been examined across theoretical, application and technical dimensions. Theoretical-wise, Stureborg et al. [Stureborg *et al.*, 2023] studied whether and how concept hierarchies can inform the design of annotation interfaces to improve labeling quality and efficiency. Hartwig et al. [Hartwig *et al.*, 2024b] surveyed user-centered misinformation interventions and proposed taxonomies of collaborative annotations. Most studies examine individual components for single-user annotations rather than developing video-oriented theories. Application-wise, Bhuiyan et al. [Bhuiyan *et al.*, 2023] proposed comparative news annotation, while Wood et al. [Wood *et al.*, 2018] built a mobile application to support co-annotating online news articles. These implementations lack the specific focus on video annotation or governance. Technical-wise, Park et al. [Park *et al.*, 2024] discussed leveraging LLMs as interactive research tools to facilitate human-AI collaboration in annotating online risk data. Shabani et al. [Shabani *et al.*, 2021] leveraged humans' fact-checking skills by providing feedback on news stories about the source, etc. Jahanbakhsh et al. [Jahanbakhsh *et al.*, 2022; Jahanbakhsh and Karger, 2024] designed a prototype social media platform with in-browser signaling to support collaborative trusted annotations. However, all previous platforms did not consider the annotation on rich multimodal data such as videos. These modalities brought inherent challenges for annotation, which we aimed to cope with.

3 Synlab: Design and Implementation

3.1 Algorithmic Design Guidelines

Synlab aims to enhance the identification of video misinformation by effectively processing and synthesizing human-annotated visual data. Rooted in the theoretical foundations of Collective Intelligence (CI) and Social Influence (SI), Synlab addresses the inherent complexity of dynamic video content via a three-stage algorithmic pipeline: (i) the structuring of visual annotations, (ii) the spatio-temporal aggregation of observations, and (iii) the generation of synthesized visual feedback. This framework converts diverse, fine-grained spatio-temporal annotations into reliable consensus. The following design principles guide the development of our algorithms:

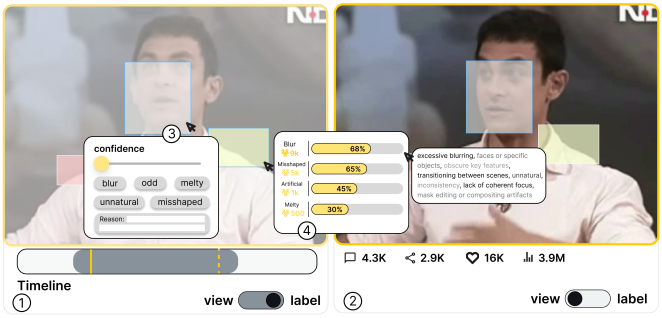

Figure 2: The interface for Synlab, where ① users toggled labeling to draw regions on the interface, with its time range being the temporal regions. They input labels and confidence. The default viewing state is shown in ②.

(G1) Encode Dynamic Temporal Visual Information: The framework must effectively encode and process the dynamic, sequential visual information in video, ensuring that the user-identified spatio-temporal manipulations are accurately captured and represented from visual annotations.

(G2) Mitigate Bias and Assess Reliability in Visual Aggregation: The aggregation algorithm should integrate principled mechanisms to attenuate biases arising from heterogeneous visual inputs, while dynamically evaluating the reliability of individual annotators. By modulating the influence of spatio-temporal annotations based on assessed trustworthiness, the framework enhances the overall robustness of the aggregated representation.

(G3) Efficiently Process Fine-grained Visual Data: The algorithms should be optimized to efficiently process and integrate highly granular spatio-temporal visual annotations, thereby preserving the richness of human perceptual input necessary for the analysis of complex and nuanced visual manipulations.

(G4) Provide Algorithmic Support for Visual Evidence Presentation: Feedback generation algorithms should ensure the transparent demonstration of aggregated visual findings and their underlying spatio-temporal evidence.

3.2 Synlab: System Architecture and Algorithms

Synlab, implemented as a web-based plugin for collaborative video annotation within social media contexts, is structured around three core algorithmic components: (1) an intuitive spatio-temporal annotation interface, (2) a novel aggregation algorithm that synthesizes collective intelligence, and (3) a principled cold-start management strategy. Users identify misinformation by marking rectangular spatio-temporal regions on the video player (Figure 2①). Each region is supplemented with a semantic label (either predefined or user-defined), a confidence score (0–100%), and optional textual rationales. Synlab visualizes the submitted annotations on a timeline, where hovering over each region reveals the corresponding aggregated consensus labels.

Annotation and Interaction System

The Synlab frontend, implemented in JavaScript, utilizes the HTML Canvas API to render a transparent overlay on

top of the video player. This overlay enables users to create and visualize spatio-temporal annotations. User interactions—such as toggling annotation mode, specifying time segments, and submitting annotation data—are handled by client-side JavaScript. The plugin architecture is operating system–independent, requiring only a modern web browser with support for JavaScript and the HTML Canvas API (e.g., current versions of Chrome, Firefox, Edge, or Safari). Compatibility has been verified across several major video-centric social media platforms, including YouTube, Twitter, and Bilibili.

Annotation data is transmitted asynchronously to the backend upon submission. The backend is developed in Python, utilizing the Flask web framework to handle API requests and execute the aggregation logic. Numerical computations integral to the aggregation process, such as the confidence-weighted Intersection-over-Union (IoU) calculations, are performed using NumPy.

Confidence-Weighted Annotation Aggregation Algorithm

To effectively leverage collective user intelligence, Synlab introduces a novel confidence-weighted annotation aggregation algorithm (detailed in Algorithm 1). This algorithm significantly advances beyond traditional methods by adopting a confidence-weighted inductive strategy that incorporates user-assigned confidence and historical reliability, crucial for discerning subtle manipulations in complex spatio-temporal data. The aggregation process has two phases:

Phase 1: Merging Annotations.

1. Sorting: Annotations (A) are initially sorted in descending order based on their confidence scores.

2. Iterative Merging: The algorithm iterates through the sorted annotations. For each annotation (a), it attempts to merge it with existing aggregated regions (R).

• A merge occurs if the 3D Intersection-over-Union (IoU) between a's region and an existing region r's region meets or exceeds a predefined threshold (IoU_{thresh}), balancing label granularity with semantic coherence. The threshold is empirically set to 40%.

• The **AppendAnnotationByLabel** function manages the addition of a's information to r. If a's label matches an existing label stream in r, the function appends a's confidence score and rationale to that stream. Otherwise it contributes a distinct label stream within r.

• The **ConfWeightedAvg** function updates the spatial extent of r, calculating the merged region's coordinates as a weighted average of all annotations contributing to r. This method, unlike Non-Maximum Suppression (NMS), retains broad collective intelligence by weighting contributions rather than solely selecting the highest confidence annotation, thereby mitigating risks from potential errors in isolated high-confidence inputs.

• If an annotation a does not achieve the IoU_{thresh} with any existing region in R, it initiates a new aggregated region cluster via the **CreateNewAggregatedRegion** function.

Phase 2: Calculating Aggregated Attributes.

Once all annotations are processed, the algorithm calculates the final attributes for each aggregated region r in R.

1. For each label identified within an aggregated region r:

• Weighted Score Calculation: For each individual annotation contributing to this label, our method computes a weighted score by multiplying the annotation's confidence by the user's average historical annotation confidence (retrieved via **GetUserHist**).

Algorithm 1 Confidence-weighted Visual Annotation Aggregation

Require: Annotations $A = \{a_i\}$, IoU_{thresh}, T
$A_{sorted} \leftarrow$ SortDesc(A, by confidence)
$R \leftarrow []$
▷ Phase 1: Merge annotations
for all $a \in A_{sorted}$ **do**
  $merged \leftarrow$ False
  for all $r \in R$ **do**
    if IoU(a.region, r.region) $\geq IoU_{thresh}$ **then**
      AppendAnnotationByLabel(r, a)
      r.annotations.append(a)
      r.region $\leftarrow$ ConfWeightedAvg(r.annotations)
      $merged \leftarrow$ True; **break**
    end if
  end for
  if not $merged$ **then**
    $newAggR \leftarrow$ CreateNewAggregatedRegion(a)
    R.append($newAggR$)
  end if
end for
▷ Phase 2: Calculate aggregated attributes
for all $r \in R$ **do**
  for all label, data in r.labelData.items() **do** ▷ Iterate through each label in the region
    $confs \leftarrow$ [pair.confidence for pair in data]
    $reasons \leftarrow$ [pair.reason for pair in data]
    $wScores \leftarrow []$
    for all pair in data **do**
      $hist \leftarrow$ GetUserHist(pair.user)
      $wScores$.append(pair.confidence $\times$ Avg($hist$))
    end for
    r.aggInfo[label].score $\leftarrow$ Avg($wScores$)
    r.aggInfo[label].conf $\leftarrow$ Avg(Top($T, confs$))
    r.aggInfo[label].reason $\leftarrow$ LM_Agg($reasons$)
  end for
end for
return R

• Final Label Score: The aggregated score for this label (r.aggInfo[label].score) is the average of these weighted scores. The predominant label for the aggregation region r is determined as the label achieving the highest final aggregated score. This approach considers both the self-assessed confidence for the current annotation and the user's demonstrated capability derived from historical performance.

• Aggregated Confidence: The aggregated confidence for the label (r.aggInfo[label].conf) is calculated as the average of the top-T (empirically set to $T = 5$) highest individual confidence scores among annotations contributing to this label.

This ensures that the aggregated confidence is supported by multiple high-confidence users.

• Aggregated Rationale: Textual rationales associated with the label are aggregated using language model-based embeddings (LM_Agg), similar to prior work, to synthesize a representative reason while mitigating the influence of idiosyncratic or erroneous individual textual inputs.

The algorithm returns the set of aggregated regions R, each populated with its determined predominant label, aggregated confidence and synthesized rationale.

Cold-start Annotation Generation Strategy

To address the cold-start challenges in video annotation (i.e., no prior user data), Synlab initiates the process with a constrained random generation of seed annotations. This strategy aims to mitigate anchoring bias and promote users' critical evaluation from inception, presenting these seeds as reference points or provocations rather than definitive guides, thereby facilitating efficient task management.

The generation is governed by empirically-derived constraints on crucial parameters: first, bounding boxes are strategically positioned away from frame boundaries and constrained to a mid-to-small size range. Second, annotations are temporally dispersed across distinct video segments, maintaining a controlled density (e.g., typically one to three labels per approximately five segments). Third, attributes such as initial confidence scores, label types and accompanying rationales are diversified. Fourth, a deliberate mix of both plausible and ostensibly incorrect examples are included to stimulate rigorous user assessment. Collectively, these constraints guide parameter sampling, ensuring the initial seeded environment actively supports the objectives of the subsequent user labeling phase.

Demonstration

For visualizing aggregated results, the backend calculates aggregated confidence (C_{agg}) and inter-annotator agreement (A_{agg}) metrics for each consolidated region. Synlab then maps these metrics to specific colors based on predefined thresholds: green denotes high consensus ($C_{agg} \geq 75$ and $A_{agg} \geq 80$), red indicates low confidence or agreement ($C_{agg} \leq 40$ or $A_{agg} \leq 50$), and orange represents intermediate values. The frontend displays these colors as semi-transparent overlays (40% opacity) on the video, providing immediate visual feedback on the collective evaluation of video segments. Users can hover over aggregated annotation regions to see labels and details, with further hovering on individual labels revealing their rationales.

4 Evaluating Synlab Within Social Media Environment

To assess the effectiveness of Synlab in identifying misinformation videos within social media environments, we conducted an evaluation study comparing Synlab and alternative techniques. Guided by previous literature [Jahanbakhsh *et al.*, 2022; Jahanbakhsh and Karger, 2024], we collected user data over a seven-day period.

4.1 Experiment Material and Platform

The experiment material contained three prominent misinformation video datasets: FakeSV [Qi *et al.*, 2023], FakeTT [Papadopoulou *et al.*, 2019] and FVC [Bu *et al.*, 2024] dataset. The selection was guided by three key criteria: (1) enabling a rigorous comparison of AI-driven versus human-perceived ground truth accuracy, where these datasets were commonly benchmarked with AI algorithms [Qi *et al.*, 2023; Bu *et al.*, 2024; Zeng *et al.*, 2024], (2) ensuring thematic diversity to broaden the scope of misinformation analysis, and (3) prioritizing datasets with familiar, real-world content to enhance user engagement and relevance. To balance experimental scope with participant load considerations given our recruitment constraints, we randomly sampled 50 videos from each dataset, totaling 150 videos, whose number is similar to prior practices [Jahanbakhsh *et al.*, 2022; Jahanbakhsh and Karger, 2024].

We developed an online social media platform as the experiment platform due to ethical concerns, mimicking the layout of Twitter. We advocated to participants the platforms and let them become familiar with the platform in advance.

4.2 Data Collection Process

The study has *technique* as the only one between-subjects factor. We compared Synlab with two alternative techniques, inspired by previous work [Jahanbakhsh *et al.*, 2022; Jahanbakhsh and Karger, 2024] and specifically tailored to video misinformation annotation context.

• Synlab: it is implemented as in Section 3.

• No-Label: This variant follows Synlab's annotation and aggregation process but does not display other users' annotations on the videos.

• No-Agg: The annotation process is similar to Synlab, but users' annotations were not aggregated and thus the demonstration showed each users' aggregation separately, similar to the previous systems [Jahanbakhsh and Karger, 2024]. However, the demonstration form is adapted to video-based forms as Synlab, because the original form is targeted at textual annotations and not suitable for video misinformation.

Participants needed to view several videos a day on the platform and annotate suspicious ones. They needed to annotate at least five videos a day, determined according to previous literature [Jahanbakhsh *et al.*, 2021; Jahanbakhsh and Karger, 2024] to balance users' fatigue and align with their daily viewing practices.

We recruited 48 participants (17 males, 29 females, with a mean age of 22.9, SD=2.3), with a mix of expertise levels ranging from novice users to experts in media literacy, information science, and AI. The diversity of participants ensured that the study reflected a broad spectrum of social media users' perspectives. Participants were each compensates 210RMB for their time, and informed consent was obatined before participation. The study was approved by our university's Institutional Review Board (IRB).

4.3 Results

Our evaluation showed Synlab's effectiveness in enhancing video misinformation identification accuracy and fostering

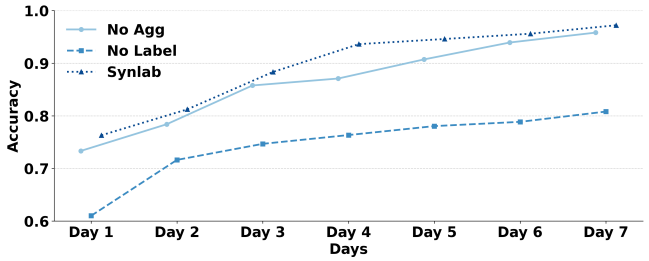

Figure 3: Average accuracy of different techniques across seven days.

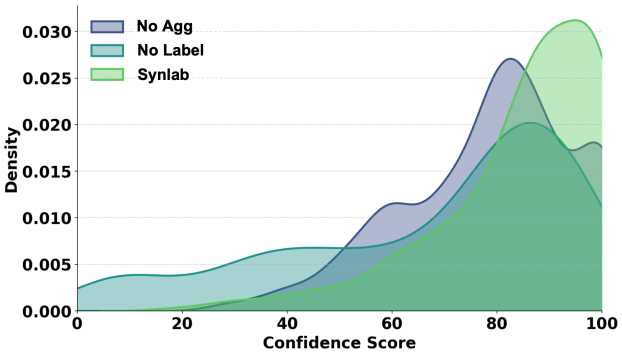

Figure 4: A Kernel Density Estimation (KDE) plot illustrating the confidence distribution of the annotations for different techniques.

user engagement, aligning with our core design goals. Statistical analyses employed One-way Analysis of Variance (one-way ANOVA) for accuracy and confidence metrics, and Wilcoxon tests for subjective ratings, with appropriate post-hoc comparisons (Tukey HSD and Nemenyi, respectively).

Annotation Accuracy

Annotation accuracy, defined as the correct identification of misinformation in videos, was a key metric in evaluating Synlab. Synlab achieved a mean accuracy of 97.20% across datasets (Table 1), significantly surpassing the *No Label* condition ($\Delta = +16.40\%$) and modestly improving upon the *No Agg* condition ($\Delta = +1.40\%$). A one-way ANOVA confirmed significant accuracy differences across techniques ($F_{2,45} = 4.031$, $p < .05$). Post-hoc analyses revealed that Synlab's accuracy was significantly higher than that of the *No Label* condition ($p < .05$), highlight the efficacy of Synlab's integrated aggregation and demonstration in producing reliable collective judgments.

Figure 3 illustrates the 7-day improvement in annotation accuracy across all techniques. Synlab demonstrated rapid initial accuracy gains, nearing saturation by Day 5. In contrast, the *No Agg* technique reached 95.80% accuracy by Day 7 at a slower pace, while the *No Label* technique achieved a final accuracy of 80.80% on Day 7. Time significantly affected accuracy for Synlab ($F_{2,45} = 3.904$, $p < .05$), *No Agg* ($F_{2,45} = 4.153$, $p < .05$), and *No Label* ($F_{2,45} = 5.304$, $p < .01$). However, we found no significant inter-day differences in post-hoc tests, which suggests a general improvement trend rather than distinct daily accuracy leaps.

Synlab yielded more convergent labeling results, likely enhancing its higher accuracy. *Artificial* labels predominated for Synlab, accounting for 33.97% of its total labels. In contrast, the *No Agg* condition most frequently showed *Mismatch* (18.88%) and *Artificial* (18.07%) labels, while *Mismatch* (22.8%) and *Artificial* (15.6%) were the leading labels for the *No Label* condition.

Synlab's high accuracy through collaborative user annotation proves compelling against contemporary algorithmic methods for detecting video misinformation on identical datasets. For instance, SV-FEND achieved an accuracy of 79.31% on FakeSV [Qi *et al.*, 2023]. FakingRecipe reported accuracies of 85.35% on FakeSV and 79.15% on FakeTT [Bu *et al.*, 2024], and MMVD attained 82.64% on FakeSV and 90.36% on FVC [Zeng *et al.*, 2024]. These results underscore that Synlab effectively facilitates user-initiated collaborative annotation for identifying video misinformation, suggesting its viability for prominent social media platforms like TikTok [Hartwig *et al.*, 2024a], Twitter, and Bilibili.

User confidence ratings provided insight into perceived certainty and the system's capacity to support accurate judgment (Figure 4). Participants using Synlab reported the highest mean confidence at 84.89% (SD=15.89%), significantly higher compared to the *No Agg* condition (mean=78.97%, SD=16.36%) and markedly greater than the *No Label* condition (mean=67.48%, SD=27.33%), which exhibited considerable uncertainty. Statistical tests confirmed a significant effect of the technique on confidence ratings ($F_{2,45} = 4.05$, $p < .05$). Further examination of Synlab's confidence distribution reveals a tendency towards higher certainty: the lowest 30% of ratings averaged 70.00%, the middle 40% averaged 90.00%, and the highest 30% achieved 100.00% confidence. In contrast, *No Agg* showed ratings of 60.00%, 82.00% and 100.00% for the lowest, middle and highest percentages, respectively, while *No label* reflected greater user uncertainty with corresponding averages of only 30.00%, 80.00% and 90.00%. These results suggest that *Synlab*'s design effectively supports users in making decisive assessments.

Core Aggregation Algorithm Evaluation

To scrutinize the efficacy of our aggregation mechanism, we evaluated the accuracy of classifying a video as "Fake" contingent upon receiving a minimum number of annotations (N). As delineated in Table 2, Synlab's aggregation strategy shows a marked superiority as annotation density increases. While the *No-Agg* condition shows a baseline accuracy, Synlab's accuracy escalates rapidly, achieving 0.9344 with just $N = 2$ annotations and nearing perfection at 0.9611 with $N = 3$ annotations. Both techniques reach perfect accuracy when $N \geq 4$ annotations are present. This trajectory compellingly illustrates that Synlab's confidence-weighted spatio-temporal IoU algorithm effectively synthesizes diverse user inputs into reliable consensus, particularly outperforming non-aggregated approaches when a critical mass of few annotations becomes available. This rapid convergence to high accuracy with a small number of annotations underscores the algorithm's efficiency and its potential for timely misinformation identification.

Table 1: The accuracy of Synlab and other alternative techniques, where ± denoted one standard deviation across videos. Overall denoted the averaged accuracy on three datasets.

	FakeSV	FakeTT	FVC	Overall
No Agg	$94.50\%_{\pm 6.00\%}$	$80.60\%_{\pm 7.90\%}$	$100.00\%_{\pm 0.00\%}$	$95.80\%_{\pm 4.20\%}$
No Label	$80.50\%_{\pm 11.60\%}$	$80.80\%_{\pm 3.10\%}$	$80.00\%_{\pm 7.40\%}$	$80.80\%_{\pm 8.70\%}$
Synlab	$96.70\%_{\pm 4.3\%}$	$94.00\%_{\pm 11.30\%}$	$100.00\%_{\pm 0.00\%}$	$97.20\%_{\pm 4.80\%}$

Table 2: The aggregated accuracy of different techniques across different minimum number of annotations (N). We increased N until the results converged.

Conditions	**No-Agg**	**Synlab**
N = 1	0.8272	0.8111
N = 2	0.8769	0.9344
N = 3	0.9434	0.9611
N = 4	0.9480	0.9720
N = 5	0.9580	0.9720

Effect of User Historical Performance

Synlab leverage user historical performance to refine its aggregation process by assessing and considering annotator reliability. We assess annotator reliability based on True Positives (TP) for correctly identified "Fake" videos and False Positives (FP) for "Real" videos misclassified as "Fake", comparing three distinct settings:

Simple Precision (R_{sp}): Calculated as $R_{sp} = \frac{TP}{TP+FP}$, this metric equally weights all judgments, disregarding user confidence or prior beliefs.

Confidence-Weighted Precision (R_{cw}): This refines R_{sp} by incorporating user-expressed certainty, using $R_{cw} = \frac{\sum_{i \in \text{TP annotations}} \text{confidence}_i}{\sum_{i \in \text{TP annotations}} \text{confidence}_i + \sum_{j \in \text{FP annotations}} \text{confidence}_j}$. It prioritizes high-confidence correct annotations and heavily penalizes high-confidence errors.

Bayesian Reliability (R_{bb}): A Beta(1,1) distribution forms the basis for this reliability estimate $R_{bb} = \frac{\alpha_{prior}+TP}{\alpha_{prior}+\beta_{prior}+TP+FP} = \frac{1+TP}{2+TP+FP}$. The use of this Beta(1,1) prior signifies maximal initial uncertainty (a uniform prior) regarding user reliability and provides stable estimates, especially for sparse data, as its influence diminishes with accumulating annotations.

Table 3: Aggregated accuracy of Synlab under different historical performance calculation settings (R_{sp}: Simple Precision, R_{cw}: Confidence-Weighted Precision, R_{bb}: Bayesian Reliability) with varying minimum annotations (N). We increased N until the results converged.

Conditions	R_{sp}	R_{cw}	R_{bb}
N = 1	0.8272	0.8250	0.8220
N = 2	0.8750	0.8814	0.8650
N = 3	0.9362	0.9720	0.9005
N = 4	0.9550	0.9720	0.9328

Table 3 presented the empirical impact of these reliability settings on Synlab's aggregated accuracy, contingent on the minimum number of reliable annotations. The results underscore that the methodology chosen for assessing user reliability substantially influences the system's performance. While R_{sp} offers a marginal advantage when only a single annotation is required ($N = 1$, accuracy 0.8272), the R_{cw} setting consistently yields superior accuracy for $N \geq 2$. This advantage is particularly salient at $N = 3$, where R_{cw} achieves an accuracy of 0.9720, significantly outperforming R_{sp} (0.9362) and R_{bb} (0.9005). R_{cw} maintains its high performance at N=4, achieving 0.9720 accuracy (versus 0.9550 for R_{sp} and 0.9328 for R_{bb}). Our findings show that incorporating user-expressed confidence via R_{cw} setting effectively discerns reliable annotators from historical data, underscoring the importance of nuanced feedback.

5 Discussions and Future Work

Synlab offers a robust framework for effectively leveraging collective user input, a critical objective in collaborative systems [Jahanbakhsh and Karger, 2024]. Its core capability of algorithmically synthesizing diverse, fine-grained user inputs into a reliable consensus distinguishes it from platforms that lack specialized algorithmic tools for granular content auditing [Hartwig *et al.*, 2024a]. This capability is highly extensible, with potential applications including enhancing crowdsourced data labeling for machine learning [Chang *et al.*, 2017], developing nuanced peer-review systems [Pareek and Goncalves, 2024], and implementing advanced community-based content moderation platforms [Jahanbakhsh *et al.*, 2022].

While developed for video, Synlab's algorithms are adaptable to static visual media [Castano *et al.*, 2019] or text-based content [Suzuki, 2015]. Such adaptations would primarily necessitate modifications to the annotation interface and, where applicable, adjustments for handling temporal components. Given the prevalence of multimodal content [Griffith and Papacharissi, 2010; Abas, 2011], Synlab's methodology can analyze each modality independently or integrate analyses through defined linking mechanisms.

Acknowledgements

This work was supported by the Natural Science Foundation of China under Grant No. 62472243, 62132010 and 2022YFB3105201. This work was also supported by Deng Feng Fund.

Ethical Statement

We acknowledged that our paper has ethical concerns and tried our best to address these concerns. We followed Menlo

report [Bailey *et al.*, 2012] and Belmont report [Beauchamp and others, 2008] in organizing, designing and carrying out studies. Participants in our study has freedom in quitting the study without any reason, and was properly compensated according to the local wage standard. At the designing stage, our study encouraged participants to discern misinformation videos and critically think about the misinformation. Participants also agreed unanimously they reflected more critically after using the systems. After the study, we debriefed the study's content, the fake videos to the participants and encouraged them to critically examine the misinformation videos.

References

[Abas, 2011] Suriati Abas. Blogging: A multimodal perspective. *Changing demands, changing directions. Proceedings ascilite Hobart*, pages 13–20, 2011.

[Alrashidi *et al.*, 2022] Bedour Alrashidi, Amani Jamal, Imtiaz Khan, and Ali Alkhathlan. A review on abusive content automatic detection: approaches, challenges and opportunities. *PeerJ Computer Science*, 8:e1142, 2022.

[Arsht and Etcovitch, 2018] Andrew Arsht and Daniel Etcovitch. The human cost of online content moderation. *Harvard Journal of Law and Technology*, 2, 2018.

[Bailey *et al.*, 2012] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. The menlo report. *IEEE Security & Privacy*, 10(2):71–75, 2012.

[Beauchamp and others, 2008] Tom L Beauchamp et al. The belmont report. *The Oxford textbook of clinical research ethics*, pages 149–155, 2008.

[Bélair-Gagnon *et al.*, 2023] Valérie Bélair-Gagnon, Rebekah Larsen, Lucas Graves, and Oscar Westlund. Knowledge work in platform fact-checking partnerships. 2023.

[Belgacem *et al.*, 2021] Khadidja Belgacem, Mouna Kenoui, Feriel Bouguerra, Mohammed Laidi, Amine Semrani, and Celia Sellah. Collaborative visualization and annotations of dicom images for real-time web-based telemedicine system. In *2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI)*, pages 1–6. IEEE, 2021.

[Bhuiyan *et al.*, 2023] Md Momen Bhuiyan, Sang Won Lee, Nitesh Goyal, and Tanushree Mitra. Newscomp: Facilitating diverse news reading through comparative annotation. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.

[Bozarth *et al.*, 2023] Lia Bozarth, Jane Im, Christopher Quarles, and Ceren Budak. Wisdom of two crowds: Misinformation moderation on reddit and how to improve this process—a case study of covid-19. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–33, 2023.

[Bu *et al.*, 2024] Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. Fakingrecipe: Detecting fake news on short video platforms from the perspective of creative process. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1351–1360, 2024.

[Castano *et al.*, 2019] Silvana Castano, Alfio Ferrara, and Stefano Montanelli. Leveraging crowd skills and consensus for collaborative web-resource labeling. *Future Generation Computer Systems*, 95:790–801, 2019.

[Chang *et al.*, 2017] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 2334–2346, 2017.

[Cialdini and Goldstein, 2004] Robert B Cialdini and Noah J Goldstein. Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55(1):591–621, 2004.

[Fernandez and Alani, 2018] Miriam Fernandez and Harith Alani. Online misinformation: Challenges and future directions. In *Companion proceedings of the the web conference 2018*, pages 595–602, 2018.

[Gorwa *et al.*, 2020] Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945, 2020.

[Griffith and Papacharissi, 2010] Maggie Griffith and Zizi Papacharissi. Looking for you: An analysis of video blogs. *First Monday*, 2010.

[Hartwig *et al.*, 2024a] Katrin Hartwig, Tom Biselli, Franziska Schneider, and Christian Reuter. From adolescents' eyes: Assessing an indicator-based intervention to combat misinformation on tiktok. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2024.

[Hartwig *et al.*, 2024b] Katrin Hartwig, Frederic Doell, and Christian Reuter. The landscape of user-centered misinformation interventions-a systematic literature review. *ACM Computing Surveys*, 56(11):1–36, 2024.

[Heuer and Glassman, 2022] Hendrik Heuer and Elena Leah Glassman. A comparative evaluation of interventions against misinformation: Augmenting the who checklist. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2022.

[Hlaoua, 2024] Lobna Hlaoua. An overview of aggregation methods for social networks analysis. *Knowledge and Information Systems*, pages 1–28, 2024.

[Jahanbakhsh and Karger, 2024] Farnaz Jahanbakhsh and David R Karger. A browser extension for in-place signaling and assessment of misinformation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024.

[Jahanbakhsh *et al.*, 2021] Farnaz Jahanbakhsh, Amy X Zhang, Adam J Berinsky, Gordon Pennycook, David G Rand, and David R Karger. Exploring lightweight interventions at posting time to reduce the sharing of misin-

formation on social media. *Proceedings of the ACM on human-computer interaction*, 5(CSCW1):1–42, 2021.

[Jahanbakhsh *et al.*, 2022] Farnaz Jahanbakhsh, Amy X Zhang, and David R Karger. Leveraging structured trusted-peer assessments to combat misinformation. *Proceedings of the ACM on Human-computer Interaction*, 6(CSCW2):1–40, 2022.

[Jahanbakhsh *et al.*, 2023] Farnaz Jahanbakhsh, Yannis Katsis, Dakuo Wang, Lucian Popa, and Michael Muller. Exploring the use of personalized ai for identifying misinformation on social media. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–27, 2023.

[Kaufman *et al.*, 2022] Robert A Kaufman, Michael Robert Haupt, and Steven P Dow. Who's in the crowd matters: Cognitive factors and beliefs predict misinformation assessment accuracy. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–18, 2022.

[Kirchner and Reuter, 2020] Jan Kirchner and Christian Reuter. Countering fake news: A comparison of possible solutions regarding user acceptance and effectiveness. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW2):1–27, 2020.

[Lu *et al.*, 2022] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. The effects of ai-based credibility indicators on the detection and spread of misinformation under social influence. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–27, 2022.

[Papadopoulou *et al.*, 2019] Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Ioannis Kompatsiaris. A corpus of debunked and verified user-generated videos. *Online information review*, 43(1):72–88, 2019.

[Pareek and Goncalves, 2024] Saumya Pareek and Jorge Goncalves. Peer-supplied credibility labels as an online misinformation intervention. *International Journal of Human-Computer Studies*, 188:103276, 2024.

[Park *et al.*, 2024] Jinkyung Park, Pamela Wisniewski, and Vivek Singh. Leveraging large language models (llms) to support collaborative human-ai online risk data annotation. *arXiv preprint arXiv:2404.07926*, 2024.

[Pennycook *et al.*, 2021] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595, 2021.

[Qi *et al.*, 2023] Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 2023.

[Shabani *et al.*, 2021] Shaban Shabani, Zarina Charlesworth, Maria Sokhn, and Heiko Schuldt. Sams: human-in-the-loop approach to combat the sharing of digital misinformation. In *Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering*. CEUR workshop proceedings, 2021.

[Sharma *et al.*, 2019] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42, 2019.

[Stureborg *et al.*, 2023] Rickard Stureborg, Bhuwan Dhingra, and Jun Yang. Interface design for crowdsourcing hierarchical multi-label text annotations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.

[Suzuki, 2015] Ryo Suzuki. Poster: Interactive and collaborative source code annotation. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, volume 2, pages 799–800. IEEE, 2015.

[Thomas, 2022] Naomi Thomas. Doctors worry that online misinformation will push abortion-seekers toward ineffective, dangerous methods. *CNN*, 2022. Available at: https://www.cnn.com.

[Vosoughi *et al.*, 2018] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[Wang *et al.*, 2019] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine*, 240:112552, 2019.

[Wilner *et al.*, 2023] Tamar Wilner, Kayo Mimizuka, Ayesha Bhimdiwala, Jason C Young, and Ahmer Arif. It's about time: Attending to temporality in misinformation interventions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

[Wolpert and Tumer, 1999] David H Wolpert and Kagan Tumer. An introduction to collective intelligence. *arXiv preprint cs/9908014*, 1999.

[Wood *et al.*, 2018] Gavin Wood, Kiel Long, Tom Feltwell, Scarlett Rowland, Phillip Brooker, Jamie Mahoney, John Vines, Julie Barnett, and Shaun Lawson. Rethinking engagement with online news through social and visual coannotation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[Zeng *et al.*, 2024] Zhi Zeng, Minnan Luo, Xiangzheng Kong, Huan Liu, Hao Guo, Hao Yang, Zihan Ma, and Xiang Zhao. Mitigating world biases: A multimodal multiview debiasing framework for fake news video detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6492–6500, 2024.