From Clips to Communities: Fusing Social Video into Knowledge Graphs for Localness-Aware LLMs

Zihan Gao
zihan.gao@wisc.edu
Information School, University of Wisconsin-Madison
Madison, Wisconsin, USA

Jiaying “Lizzy” Liu
jiayingliu@utexas.edu
School of Information, The University of Texas at Austin
Austin, Texas, USA

Yifei Xu
yxu@cs.ucla.edu
Computer Sceience, University of California, Los Angeles
Los Angeles, California, USA

Jacob Thebault-Spieker
jacob.thebaultspieker@wisc.edu
Information School, University of Wisconsin-Madison
Madison, Wisconsin, USA

Abstract

Understanding how people experience and assign meaning to places — what we term “localness” — is essential for designing systems that support orientation, belonging, and community integration. However, current computational approaches to modeling place often struggle to capture the multimodal, relational, and contextual richness of localness. In this paper, we introduce a novel workflow for extracting structured localness representations from geotagged short-form social videos using a Graph-Enhanced Retrieval-Augmented Generation (Graph-RAG) framework. Our method aligns and fuses multimodal video data from TikTok into a knowledge graph, retrieves relevant contextual evidence through a hybrid of graph-based and semantic retrieval, and then prompts large language models (LLMs) to infer fine-grained localness attributes across cognitive, relational, and physical domains. We evaluate our system using both human-annotated ground truth and qualitative analysis, demonstrating full coverage across all localness components, with especially strong performance in environmental and relational dimensions. Our findings reveal both the promises and the challenges of grounding AI systems in lived experiences of place, and we offer design implications for localness-aware technologies in digital placemaking, community informatics, and contextual computing.

CCS Concepts

• **Human-centered computing → Collaborative and social computing systems and tools**; • **Computing methodologies → Information extraction**.

Keywords

Localness; Knowledge graph; Large language models; Retrieval-augmented generation (RAG); Multimodal data; Place representation

ACM Reference Format:
Zihan Gao, Jiaying “Lizzy” Liu, Yifei Xu, and Jacob Thebault-Spieker. 2025. From Clips to Communities: Fusing Social Video into Knowledge Graphs for Localness-Aware LLMs. In *Companion of the Computer-Supported Cooperative Work and Social Computing (CSCW Companion ’25), October 18–22, 2025, Bergen, Norway*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3715070.3749277

1 Introduction

When people move or travel to a new city or neighborhood, they often struggle to understand the everyday texture of a place [10]: Where do locals actually eat? Is that park good for quiet reading or weekend festivals? What’s the vibe of the neighborhood — family-oriented, artsy, student-heavy? While search engines and location-based services offer facts like hours, ratings, or addresses, they rarely capture the emotional, cultural, or social layers that give a place its character [18, 29, 32]. Information is scattered across platforms — reviews on Google Maps, event listings on Facebook, posts on Reddit, or short videos on TikTok — making it difficult to form a cohesive understanding. Increasingly, people tend to turn to peer-generated content[29, 32], especially short videos, for richer insights into local environments [4, 10]. These videos offer immersive, narrative-rich portrayals of place: walking tours, daily routines, or event vlogs that convey how people see, hear, and interact with their surroundings. CSCW scholars have long emphasized the importance of such multimodal and experiential cues in understanding place meaning [5, 6, 10, 18, 33]. However, the lack of systematic methods to retrieve such scattered and unstructured information limits its broader utility for spatial orientation and community integration.

To address this gap, we center our work on the framework of *localness*: how people experience, relate to, and become situated in specific places through cultural practices, social ties, and embodied routines [11, 13, 15]. Localness captures more than physical location: it reflects feelings of belonging, knowledge of community norms, familiarity with language or landmarks, and emotional attachment, etc. [11]. We aim to build computational systems that can capture these multidimensional aspects of place in accessible ways. Our approach introduces a graph-enhanced, retrieval-augmented generation framework for encoding localness from social videos. Compared to traditional RAG approaches that focus primarily on semantic similarity for retrieval, our graph-enhanced approach

CSCW Companion ’25, Bergen, Norway

ACM ISBN 979-8-4007-1480-1/2025/10
https://doi.org/10.1145/3715070.3749277

explicitly models the relationships among place-related information, enabling retrieval that considers both semantic relevance and structural connectivity within the place narrative [16, 23, 39].

We focus on TikTok as our primary data source due to its short-form, narrative-rich structure, its affordances for multimodal storytelling, and its widespread global adoption — particularly among youth and diverse communities [3, 8, 27, 41]— which enables access to vernacular place experiences often underrepresented in traditional data sources [4, 36, 37]. Specifically, starting with TikTok videos, we systematically construct a knowledge graph that captures the narrative and relational context of how places are depicted, and use this graph to guide LLMs in generating structured evidence-grounded localness attributes. This approach enables us to preserve both the sensory richness and the social fabric essential to place-based reasoning [1, 19, 21]. This study is guided by two research questions:

- **RQ1: How can multimodal social media content be systematically analyzed to construct place knowledge graphs?**
- **RQ2: Can graph-enhanced retrieval and prompt-guided LLMs reliably extract grounded, fine-grained localness attributes?**

Our contributions are threefold:

- This study introduces a methodological framework for constructing place knowledge graphs from multimodal user-generated content;
- We develop a graph-enhanced retrieval-augmented generation approach for encoding localness as place representation;
- This study provides empirical insights of the strengths and limitations of the decoding social media content and encoding localness workflow.

Ultimately, we aim to support the development of location-based services that help people understand, belong, and thrive in the places they call *home*.

2 Workflow of Capturing Localness from Multimodal Social Media Content

Our goal is to extract interpretable, context-rich representations of localness from TikTok videos. To do so, we propose a Graph-RAG framework that integrates multimodal feature extraction, structured knowledge representation, and LLM reasoning (Figure 1).

2.1 Multimodal Posts to Knowledge Graph

2.1.1 Modality Alignment. The first step is to convert multimodal TikTok posts, including video, audio, and text, into a unified text format to support downstream knowledge graph construction and retrieval. Inspired by prior studies that apply MLLMs for video analysis [24, 25], we first segment videos into keyframes, and use GPT-4o to generate a scene description for each keyframe. For fine-grained feature extraction, we detect objects, landmarks, and on-screen text in each keyframe and represent all outputs in text form. We use Whisper to transcribe audios and segment them according to the corresponding time windows of each keyframe. Full modality alignment details are provided in Appendix A. Eventually, all multimodal TikTok content is extracted and converted into text, including post text, frame transcripts, detected landmarks, on-screen text, and scene descriptions.

2.1.2 Geographic Entity Recognition and Verification. To associate the extracted elements with locations, we use spaCy’s NER to extract named locations and cultural entities, and GPT-4o to filter ambiguous mentions like “the Square” or “the Heights” with contextual reasoning. We verify the extracted locations to ensure they are clearly mapped to real locations. We use the Google Places API to map them to official names, formatted addresses, coordinates, and related metadata. We then query Wikidata for detailed descriptions and compute their semantic similarity with the corresponding frame transcripts and scene descriptions for cross-validation. If their cosine similarity exceeds a certain threshold, we consider the location verified. We reuse the vague mentions identified by GPT-4o by linking them to the verified place as local or contextual aliases. This preserves local language while grounding it in real-world geography.

2.1.3 Fusing Entities and Relationships into Knowledge Graph. The modality alignment and geographic entity recognition process naturally populates a rich set of entities and their relationships. For example, a geographic entity extracted from a scene description is linked to its corresponding keyframe. A complete list of entities and relationships is provided in Appendix B. We construct the knowledge graph using `NetworkX MultiDiGraph` in PostgreSQL, where entities and relationships serve as nodes and edges. To enable semantic retrieval in addition to graph-based queries (Section 2.2.2), we also generate embeddings for all entities and relationships in this step.

2.2 Encoding Localness Attributes

2.2.1 Localness Framework. We introduce the graph-enhanced RAG framework for inferring *localness attributes* — fine-grained characteristics associated with specific places. Our methodology leverages the structured Localness Conceptual Framework from Gao et al. [11], which organizes localness into three interconnected domains: *Physical*, *Cognitive*, and *Relational*. Each domain consists of clearly defined dimensions and components. The *Relational* domain captures social connections and emotional bonds (e.g., community engagement, friendships). The *Physical* domain involves direct interaction with a place (e.g., long-term residence, community gardening). The *Cognitive* domain focuses on local knowledge and cultural understanding (e.g., knowing local history, navigating without assistance). This structured, hierarchical framework enables precise and intuitive annotation by illustrating how various aspects of localness manifest in different place.

2.2.2 Hybrid Retrieval and Prompt Engineering. We combine two key components: hybrid retrieval and framework-guided prompt engineering, to generate grounded and interpretable localness attributes. **Hybrid retrieval** integrates two complementary strategies to gather relevant evidence for a given location: (1) graph retrieval collects topologically adjacent nodes from the knowledge graph in up to 2 hops, (2) semantic retrieval identifies semantically similar nodes (e.g., other videos describing similar experiences) with text embeddings. This ensures both relational and conceptual

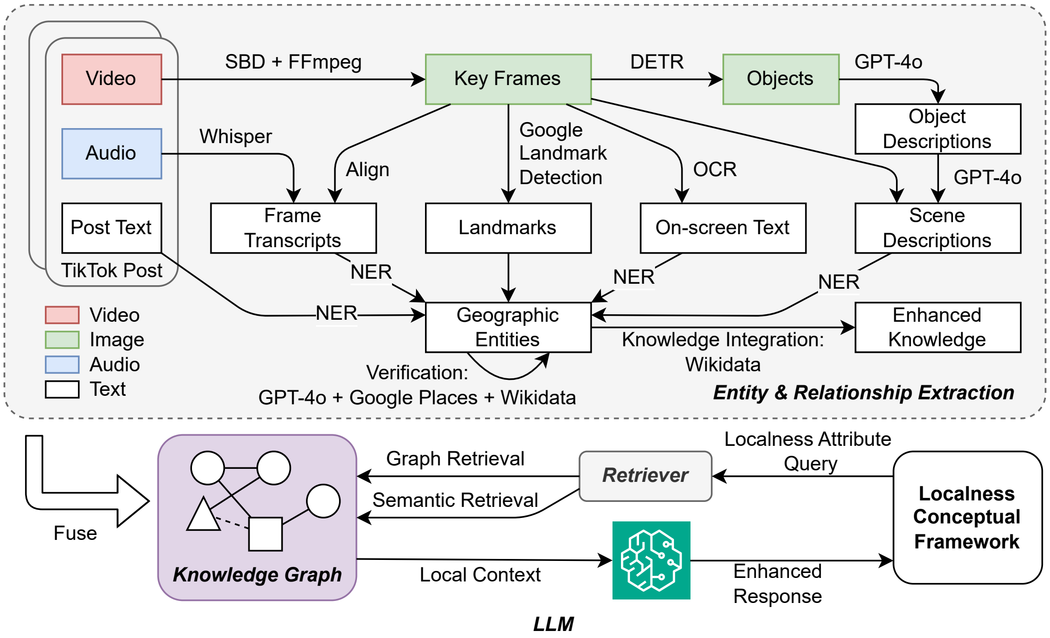

Figure 1: Workflow Visualization

relevance in the retrieved evidence. For example, for a small neighborhood café, the retrieved context might include: transcripts where users describe the café as a "quiet local spot to study"; scene descriptions showing people working on laptops; spatial relationships linking the café to a nearby bookstore and park frequently mentioned in the same video; and audio cues such as regional speech patterns or ambient jazz music. **Prompt engineering** is structured around the Localness Conceptual Framework. This framework governs both the format of the prompt and the expected output structure. At inference time, we embed the retrieved evidence into a prompt template that instructs the LLM to extract localness attributes that are specific to the target location solely based on the provided context. The prompt includes detailed instructions to ensure quality control: the LLM must distinguish between direct observation and inference, cite concrete supporting evidence (e.g., quotes, timestamps, visual descriptions), and assign confidence levels. This hybrid strategy ensures that LLM outputs are not only semantically rich and relationally aware, but also evidentially grounded

2.3 Human Evaluation

To assess the effectiveness of our framework in capturing localness, we conducted human evaluation using the evidence retrieved by the Graph-RAG system. Two human coders independently annotated localness attributes for each target location, following the same schema used in LLM prompting. Prior to reconciliation, inter-rater reliability was assessed using Cohen's kappa ($\kappa = 0.72$), indicating substantial agreement between coders. They then reconciled their annotations through discussion to create a ground truth for evaluating the LLM's output.

We compared LLM-generated attributes against this ground truth using three metrics: (1) *Recall* — the proportion of human-annotated attributes identified by the LLM; (2) *Precision* — the semantic similarity between LLM and human attributes; and (3) *Hallucination Rate:* — the proportion of LLM-generated attributes that are neither present in the human annotations nor directly supported by retrieved evidence. This metric offers a lens into overgeneration behaviors and grounding failures. To complement these metrics, we conducted a content analysis of the generated attributes, reasoning, and evidence to better understand the quality, potential, and limitations of the approach.

3 Evaluating Localness Attributes Inference

We selected 3 geotags corresponding to different U.S. cities and collected a dataset comprising 1,033 TikTok videos. From these videos, we extracted 823 verified location mentions using our multimodal entity verification approach. We randomly sampled 28 locations appearing in 45 videos and created ground truth annotations of their localness attributes. By comparing ground-truth with generated attributes, we found that the integration of graph-enhanced retrieval and structured prompting proved effective in extracting nuanced localness attributes from TikTok videos. By combining human evaluation metrics — *coverage*, *semantic similarity*, and *hallucination rate* — with grounded content analysis, we demonstrate the strengths and limitations of our framework in modeling place representation through social video data.

✓**The Workflow Achieves Complete Coverage Across All Localness Components.** This indicates that our graph-enhanced RAG approach is able to activate the complete spectrum of place

Table 1: Domain components and evaluation metrics

Localness Component	**Description**	**Cov.**	**Halluc.**	**Sim.**	**Items**
Localness Domain: Cognitive					
Change Awareness	Recognition of how the place is evolving and developing over time.	1	0.3333	0.6667	12
Geographic Familiarity	Understanding spatial layout, recognizing landmarks, and knowing local boundaries and land features.	1	0.4103	0.5510	49
Food Culture	Understanding and appreciation of the place's culinary traditions.	1	0.5500	0.4929	30
Historical Knowledge	Understanding the place's past and how it has evolved over time.	1	0.1111	0.7787	18
Ecological Understanding	Awareness of seasonal patterns, environmental cycles, local watersheds, and environmental challenges.	1	0.1667	0.7287	17
Local Recommendations	Having detailed knowledge about local establishments and services.	1	0.0833	0.8123	25
Natural Environment	Knowledge of local flora and fauna, environmental literacy, and understanding natural systems and patterns.	1	0.0000	0.8134	8
Language and Dialect	Familiarity with the unique linguistic features of the place.	1	0.3810	0.6773	32
Local Customs and Norms	Understanding unwritten social rules and behaviors specific to the place area.	1	0.1538	0.7290	25
Localness Domain: Physical					
Formative Years	Having spent developmental years in the place, shaping one's worldview.	1	0.5000	0.5000	22
Geographic Familiarity	Practical navigation skills and spatial orientation in the local area.	1	0.4103	0.5510	49
Ecological Understanding	Physical connection and embodied familiarity with local ecosystems.	1	0.1427	0.7926	17
Environment Experience	Reflects a personal, physical connection to natural spaces and local landscapes.	1	0.2000	0.7000	12
Localness Domain: Relational					
Civic Engagement	Participating in local governance and political processes in the area.	1	0.5769	0.4615	35
Sense of Belonging	Feeling comfortable and accepted within the place.	1	0.1538	0.7900	37
Active Participation	Direct involvement in community activities and initiatives in the place.	1	0.1538	0.7409	23
Feeling of Home	Emotional attachment to the place that creates security and comfort.	1	0.1154	0.7949	47
Community Investment	Demonstrating care and commitment to the wellbeing of the area.	1	0.0417	0.8295	21
Identity Connection	Incorporating the place into one's self-concept and identity.	1	0.1923	0.7424	32

Note: Cov. = Coverage, Halluc. = Hallucination rate, Sim. = Similarity, Items = Total number of evaluated instances.

attributes identified in human annotations. The extracted attributes demonstrate the system's ability to represent place beyond physical descriptions alone, capturing the socio-cultural dimensions that define how communities experience places.

✓**Physically Observable Attributes Are Captured with High Precision and Minimal Hallucination.** For example, *Natural Environment* scored perfectly with 0% hallucination and a high similarity score of 81.3%, suggesting that environmental features are easier to ground using visual and narrative content. This is consistent with observations from locations such as *springs_recreation_area_A*:

> "The *springs_recreation_area_A* features a diverse ecological system with winter landscapes that include snow-covered paths and a frozen lake."

The attribute is grounded in direct visual evidence from videos showing snowy landscapes with people walking on frozen lakes.

✓**Emotional and Relational Dimensions Are Effectively Extracted from Social Media Cues.** Attributes tied to emotional connection, such as *Feeling of Home*, *Sense of Belonging*, and *Identity Connection*, demonstrated both high similarity scores (> 74%) and low hallucination rates (11.5%, 15.4%, and 19.2%, respectively). This suggests that emotional cues, especially when expressed through hashtags, enthusiastic captions, and visual event participation senarios, are particularly amenable to LLM-based inference. In the case of *block_party_B*, the model accurately captured emotional resonance and collective rituals through phrases like:

> "The location fosters a sense of community among its residents, as evidenced by social interactions during events."

This attribute reflects how specific social rituals like the *block_party_B*, serve as anchors for community identity and belonging, both detectable through repeated social media themes.

✓**Temporal and Relational Context Provides Foundation for Advanced Place Reasoning.** Certain components, like *relational context* and *temporal dimensions* stood out for their balance of low hallucination and high similarity. *Community Investment* (hallucination: 4.2%, similarity: 82.9%), *Historical Knowledge* (11.1%, 77.9%), and *Local Recommendations* (8.3%, 81.2%) exemplify categories where local knowledge is often stated directly or implied with strong supporting evidence. For example, in the restaurant context:

> "*restaurant_C* was a predecessor to the current *restaurant_D*, indicating its historical significance in the area."

The system not only identifies individual places but situates them within historical narratives and evolving relationships. Similarly, the system's sensitivity to temporal dimensions is evident in its ability to represent how places transform across seasons:

> "The *natural_landscape_E* experiences significant seasonal changes, particularly in winter when the area is popular for outdoor activities like walking on the ice and ice fishing."

These attributes capture the dynamic nature of place experience across time — something traditional place representations often miss. The relatively strong performance metrics for temporally-oriented categories validate the graph-enhanced approach's effectiveness in capturing these crucial dimensions of place.

? **The Framework Overgeneralizes Cultural Patterns from Sparse Cues.** Despite prompt-level instructions to avoid using prior knowledge, the system frequently overgeneralizes cultural attributes from minimal evidence. *Food Culture* and *Language and Dialect* showed high hallucination rates (55% and 38.1%, respectively), often extrapolating broad cultural claims from isolated mentions. For example, a single dish — cheese curds — led to the attribution:

"This place is known for its dairy-based cuisine."

These culturally plausible yet unsupported inferences suggest the LLM fills contextual gaps using learned priors when retrieved content is sparse. This limitation is especially salient for culturally nuanced components requiring deeper, grounded understanding of local practices.

? **The System Hallucinates Social and Civic Attributes in the Absence of Evidence.** *Civic Engagement*, in particular, exhibited the highest hallucination rate (57.7%) and the lowest semantic similarity score (46.2%). Unlike cultural overgeneralization, which arises from minimal evidence, these hallucinations occur in the complete absence of supporting data. For instance, in videos from a *student_neighborhood_F*, the model inferred community activism and political engagement, despite no visual or textual indicators of such behavior. These hallucinations occurred even though prompts explicitly instructed the LLM to skip attributes without grounded evidence. This suggests a deeper limitation: the model defaults to filling gaps with plausible social narratives, especially for dimensions of localness that are abstract or difficult to observe visually.

4 Discussion

4.1 Graph-RAG Successfully Captures Rich Place Representations While Enabling Transparent Reasoning

4.1.1 ***Graph-RAG enables structured, situated representations of localness.*** Our findings confirm that the Graph-RAG framework is a powerful approach for generating comprehensive, fine-grained representations of place. By structuring TikTok-derived information into a place knowledge graph, the system achieved full coverage across all localness components we examined, including physical, cognitive, and relational domains of place. In particular, it excelled in categories reflecting lived experience, such as identity, belonging, and community connection, which are often hard to quantify. This suggests our framework successfully extracts the emotional, cultural, and social layers of locality that conventional location-based services tend to overlook. The ability to surface insider perspectives (e.g. local expressions, traditions, or norms) has practical implications for systems supporting newcomers and community members [18, 33]. In short, graph-guided retrieval allows an LLM to paint a more human-centric portrait of place, preserving the contextual nuances essential to place-based reasoning [8].

4.1.2 Graph grounding improves transparency and supports community trust. Unlike end-to-end LLMs, our graph-enhanced pipeline allows users to trace where specific inferences originated (e.g., which videos or captions supported an emotional attribute). Interpretability of model outputs is critical in civic and community-facing systems [35]. By linking outputs to evidence nodes, the system fosters transparency and enables community members to validate, contest, or augment representations of place. Future work could incorporate interactive maps or visualizations to make this transparency user-facing, similar to place graph tools explored in urban informatics [6].

4.1.3 Component analysis can isolate contributions and failure points. Our current evaluation confirms the effectiveness of the full pipeline, but does not disentangle which components (e.g., graph retrieval, semantic similarity, visual tagging) contribute most to success. Ablation studies, e.g., comparing outputs with vs. without graph retrieval, could clarify the added value of structured memory. Prior work in RAG architectures shows that relational retrieval can significantly reduce hallucinations in knowledge-poor settings [22]. Likewise, a modality-level error breakdown (e.g., visual vs. auditory sources) could identify where most grounding failures originate and guide improvements.

4.2 Technical Limitations Reveal Systematic Challenges in Grounding and Bias Mitigation

4.2.1 Hallucinations reflect sparsity and overgeneralization in social video. Despite strong coverage, our results revealed notable limitations in cultural and social reasoning. High hallucination rates in categories like *Civic Engagement* and *Food Culture* underscore the challenge of generating grounded localness attributes from sparse or implicit evidence. As Quercia et al. [33] warned, even multimodal data can yield impoverished representations when deeper context is absent. Although our prompts instructed the model to skip unsupported claims, it frequently defaulted to plausible but ungrounded inferences, especially for abstract or institutional attributes. Future work should target two key improvements. First, better cross-referencing of evidence across multiple videos and sources can improve grounding, particularly for abstract or underrepresented dimensions of localness. Second, refining prompt strategies may help reduce hallucinations in socially sensitive categories, such as political participation or community dynamics. Expanding to include video data from multiple platforms (e.g., Instagram, YouTube) will also address the challenge of fragmented representations across digital ecosystems, offering a more holistic view of place experience.

4.2.2 Biases and upstream errors may compound through the pipeline. Like other social media-based systems, our model is exposed to selection biases in content creation. TikTok content skews toward younger, visually expressive users, and trends often over represent commercial or aesthetic places over utilitarian or marginalized spaces [3, 27, 38]. These biases risk reinforcing dominant narratives while excluding overlooked communities. Moreover, minor upstream errors, like OCR misreads or object misclassification, may propagate through the graph and amplify in final inferences. Such cascading failure is a known risk in long pipelines. Addressing this in future works requires diverse input sampling, debiasing modules, and post-hoc verification using population-level data (e.g., census records or local surveys).

4.3 Generalizability and Ethical Responsibility

4.3.1 Cultural generalization requires adaptation and multilingual grounding. Our study focused on English-language TikTok data from three U.S. cities, but notions of localness are culturally situated and may not transfer directly. Place attachment and civic participation, for instance, manifest differently across geographies and governance models. Additionally, digital footprints are unequally distributed: many communities are "data poor" or invisible online

[17, 26, 28]. To generalize globally, future works will require multilingual support, culture-aware prompt tuning, and collaborative refinement of the localness taxonomy through cross-regional studies.

4.3.2 Ethical safeguards and participatory input are essential for responsible use. Even publicly shared content requires careful ethical handling. TikTok videos often feature identifiable people, routines, or communities, raising concerns about privacy and misrepresentation. We adopted basic anonymization and sensitive content filtering, but future work should integrate participatory annotation, consent-aware data governance, and adversarial red teaming [9, 31]. Drawing from contextual integrity theory [30], we argue that content must be interpreted and used in ways aligned with its original audience expectations. Working with community moderators and documenting dataset construction (e.g., via datasheets) can enhance both ethical rigor and public trust [12, 14].

4.4 Beyond Place Representation: A Generalizable Approach to Fragmented, Multimodal Content

More broadly, this workflow offers a generalizable method for synthesizing meaning from fragmented, multimodal social content. While developed for place representation, the graph-enhanced RAG approach can extend to other CSCW domains, such as crisis informatics, online health communities, or migration studies, where understanding depends on piecing together narrative, visual, and audio signals across noisy media [25]. As social content continues to diversify in form and platform, methods that integrate structure and grounding will be increasingly important for making sense of complex human experiences.

References

[1] Andrea Ballatore and Stefano De Sabbata. 2020. Los Angeles as a digital place: The geographies of user-generated content. *Transactions in GIS* 24, 4 (2020), 880–902.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*. Springer, 213–229.

[3] Pew Research Center. 2024. How Americans Use Social Media. Pew Research Center Internet & Technology. https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/ Accessed: 2025-07-09.

[4] Hao Chen, Min Wang, and Zhen Zhang. 2022. Research on rural landscape preference based on TikTok short video content and user comments. *International Journal of Environmental Research and Public Health* 19, 16 (2022), 10115.

[5] Justin Cranshaw, Raz Schwartz, Jason Hong, and Norman Sadeh. 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the international AAAI conference on web and social media*, Vol. 6. 58–65.

[6] Justin B Cranshaw, Kurt Luther, Patrick Gage Kelley, and Norman Sadeh. 2014. Curated city: capturing individual city guides through social curation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3249–3258.

[7] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143* (2020).

[8] Shuangyan Du and Cecilia Yin Mei Cheong. 2025. Beyond the scenic view: a multimodal discourse analysis of sustainable tourism imaginaries on TikTok in Anhui, China. *Humanities and Social Sciences Communications* 12, 1 (2025), 1–14.

[9] Casey Fiesler, Michael Zimmer, Nicholas Proferes, Sarah Gilbert, and Naiyan Jones. 2024. Remember the human: A systematic review of ethical considerations in reddit research. *Proceedings of the ACM on Human-Computer Interaction* 8, GROUP (2024), 1–33.

[10] Zihan Gao, Justin Cranshaw, and Jacob Thebault-Spieker. 2024. Journeying Through Sense of Place with Mental Maps: Characterizing Changing Spatial Understanding and Sense of Place During Migration for Work. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–31.

[11] Zihan Gao, Cranshaw Justin, and Thebault-Spieker Jacob. 2025. A Turing Test for "Localness": Conceptualizing, Defining, and Recognizing Localness in People and Machines. *arXiv preprint arXiv:2505.07282* (2025).

[12] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[13] Steve Harrison and Paul Dourish. 1996. Re-place-ing space: the roles of place and space in collaborative systems. In *Proceedings of the 1996 ACM conference on Computer supported cooperative work*. 67–76.

[14] Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems* 35 (2022), 29217–29234.

[15] M Carmen Hidalgo and Bernardo Hernandez. 2001. Place attachment: Conceptual and empirical questions. *Journal of environmental psychology* 21, 3 (2001), 273–281.

[16] Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. 2024. Grag: Graph retrieval-augmented generation. *arXiv preprint arXiv:2405.16506* (2024).

[17] Kee Moon Jang, Junda Chen, Yuhao Kang, Junghwan Kim, Jinhyung Lee, Fabio Duarte, and Carlo Ratti. 2024. Place identity: a generative AI's perspective. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–16.

[18] Kee Moon Jang and Youngchul Kim. 2019. Crowd-sourced cognitive mapping: A new way of displaying people's cognitive perception of urban space. *Plos one* 14, 6 (2019), e0218590.

[19] Bradley S Jorgensen and Richard C Stedman. 2011. Measuring the spatial component of sense of place: a methodology for research on the spatial dynamics of psychological experiences of places. *Environment and Planning B: Planning and Design* 38, 5 (2011), 795–813.

[20] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2880–2894.

[21] Charis Lengen and Thomas Kistemann. 2012. Sense of place and place identity: Review of neuroscientific evidence. *Health & place* 18, 5 (2012), 1162–1171.

[22] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[23] Fangzhou Liu, Lin Xu, Riya Patel, Qi Zhao, and Yiming Sun. 2025. GRAG: Graph Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics.

[24] Jiaying "Lizzy" Liu, Yiheng Su, and Praneel Seth. 2025. Can Large Language Models Grasp Abstract Visual Concepts in Videos? A Case Study on YouTube Shorts about Depression. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 127, 11 pages. doi:10.1145/3706599.3719821

[25] Jiaying (Lizzy) Liu, Yunlong Wang, Yao Lyu, Yiheng Su, Shuo Niu, Xuhai "Orson" Xu, and Yan Zhang. 2024. Harnessing LLMs for Automated Video Content Analysis: An Exploratory Workflow of Short Videos on Depression. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing* (San Jose, Costa Rica) *(CSCW Companion '24)*. Association for Computing Machinery, New York, NY, USA, 190–196. doi:10.1145/3678884.3681850

[26] Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680* (2024).

[27] Darragh McCashin and Colette M Murphy. 2023. Using TikTok for public and youth mental health–A systematic review and content analysis. *Clinical child psychology and psychiatry* 28, 1 (2023), 279–306.

[28] Mazda Moayeri, Elham Tabassi, and Soheil Feizi. 2024. Worldbench: Quantifying geographic disparities in llm factual recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1211–1228.

[29] Stein Monteiro. 2024. Searching for settlement information on Reddit. *International Migration* 62, 3 (2024), 100–119.

[30] Helen Nissenbaum. 2011. A contextual approach to privacy online. *Daedalus* 140, 4 (2011), 32–48.

[31] Will Orr and Kate Crawford. 2024. Building better datasets: Seven recommendations for responsible design from dataset creators. *arXiv preprint arXiv:2409.00252* (2024).

[32] Sangkeun Park, Yongsung Kim, Uichin Lee, and Mark Ackerman. 2014. Understanding localness of knowledge sharing: a study of Naver KiN'here'. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*. 13–22.

[33] Daniele Quercia, Neil Keith O'Hare, and Henriette Cramer. 2014. Aesthetic capital: what makes London look beautiful, quiet, and happy?. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 945–955.

[34] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.

[35] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[36] Peter T Suzuki. 1985. Vernacular cabs: Jitneys and gypsies in five cities. *Transportation research part A: general* 19, 4 (1985), 337–347.

[37] Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Toward a geographic understanding of the sharing economy: Systemic biases in UberX and TaskRabbit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 3 (2017), 1–40.

[38] Gianna Williams. 2023. Black Content Creators' Responses and Resistance Strategies on TikTok. *arXiv preprint arXiv:2312.12727* (2023).

[39] Yuchen Zhang, Wei Li, Hao Chen, Ming Wang, and Yang Liu. 2024. Graph Retrieval-Augmented Generation: A Survey. *arXiv preprint arXiv:2408.00309* (aug 2024).

[40] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*. Springer, 1–21.

[41] Diana Zulli and David James Zulli. 2022. Extending the Internet meme: Conceptualizing technological mimesis and imitation publics on the TikTok platform. *New media & society* 24, 8 (2022), 1872–1890.

A Modality Alignment Details

- **Visual Keyframe Extraction:** We apply Shot Boundary Detection (SBD) using PySceneDetect[1] to identify major visual transitions. Representative frames are then extracted at these boundaries using FFmpeg, ensuring essential visual content is retained for further analysis. We recorded each Keyframe's timestamp for the following key frame time-aligned transcript segment extraction.
- **On-screen Text Extraction:** We apply OCR to extracted keyframes using PyTesseract[2], capturing textual elements embedded in visuals — such as signs, hashtags, and subtitles—which often carry crucial contextual information.
- **Visual Object Detection:** Using DETR (DEtection TRansformer) [2], we identify and localize salient objects within each video keyframe. DETR's transformer-based architecture enables robust detection even in cluttered or dynamic environments common in user-generated videos. Following object detection, we generate descriptive scene descriptions using the GPT-4o vision-language model. This model ingests both raw keyframes and previously detected objects, producing semantic captions that help contextualize the scene beyond surface-level object recognition.
- **Landmark Recognition:** To resolve place-specific semantics, we apply the Google Landmark Recognition API[3] to identify notable geographical or cultural landmarks present in keyframes. Detected locations are further verified by comparing them with entities extracted from other data, the detailed verification process is in Section 2.1.2. To further enhance semantic depth, we integrate structured knowledge from Wikidata. These resources provide structured metadata (e.g., location type, cultural tags, region hierarchy), enriching both the visual and textual representations with authoritative external information.
- **Audio Transcription:** We extract textual information from both spoken and embedded sources. Audio segments are transcribed using OpenAI's Whisper model [34]. Whisper is particularly well-suited due to its multilingual capabilities and robustness to real-world noise, a common feature in user-generated TikTok videos. we also extract both semantic and paralinguistic features. We employ pre-trained PANNs [20] and YAMNet models to tag environmental sounds and audio events, such as ambient noise, music genres, or human activity. Additionally, SpeechBrain's ECAPA-TDNN model [7] is used to generate speaker embeddings that reflect accents, voice timbre, and other sociolinguistic cues potentially indicative of local or regional identity.
- **Temporal Alignment:** Temporal patterns are incorporated through ByteTrack [40], which tracks detected objects across frames. This adds continuity to spatial dynamics and captures context-specific activities—such as people walking in front of a monument or traffic near a city square—which can provide additional cues about location types and usage. Using timestamp information, we synchronize the extracted keyframes, transcribed text, OCR outputs, and audio features to form a cohesive, time-aligned representation of each video.

B Knowledge Graph building

Nodes and Their Attributes.

- **Location Nodes:** Represent geographic entities identified in video content. Each node includes geographic coordinates, standard and vernacular names, types (e.g., park, neighborhood), and engagement metrics such as average frame duration, frequency of appearance, and narrative significance. Semantic embeddings from object co-occurrence, activities, and spatial context vectors are also attached to these nodes.
- **Video Nodes:** Encode metadata about the TikTok video including its textual content, interaction metrics (likes, shares, comments), hashtag use, and narrative summary.
- **Keyframe Nodes:** Contain visual snapshots of the video with associated scene descriptions and temporal positioning.
- **Object Nodes:** Represent recognized visual elements within frames, enriched with labels, confidence scores, bounding box coordinates, and linked Wikidata entries.
- **Narrative Segment Nodes:** Capture story-level divisions within videos, including segment descriptions, main locations, and narrative purposes.
- **Relationship Nodes:** Define types of semantic and spatial connections (e.g., "nearby," "transitioned to," "co-occurred with") and are used to annotate edges for downstream reasoning.

Edges and Relationship Semantics.

- **Location–Location:** Derived from the *location_relationships* table. These capture co-occurrence, transitions, or cultural links between places, enabling multi-hop spatial reasoning.

[1] https://github.com/Breakthrough/PySceneDetect

[2] https://github.com/madmaze/pytesseract

[3] https://developers.google.com/maps/documentation/landmark

- **Video–Location:** Denote that a location is mentioned or shown in a video. These edges carry source evidence (transcripts, hashtags, OCR, etc.) and primary/secondary status flags.
- **Keyframe–Location:** Link visual frames with specific geographic entities based on visual detection and timestamp-aligned textual mentions.
- **Video–Keyframe / Keyframe–Object:** Encode the video structure and visual content hierarchy, preserving which frames contain which objects and when.
- **Narrative Flow and Transition Edges:** Capture the movement or thematic transition between frames or places within a narrative, annotated with transition types and confidence scores.