探索以人為本的全方位實證方法：

多模態研究的幾個案例

卓牧融 | Mu-Jung ‘MJ’ Cho

mjcho@as.edu.tw

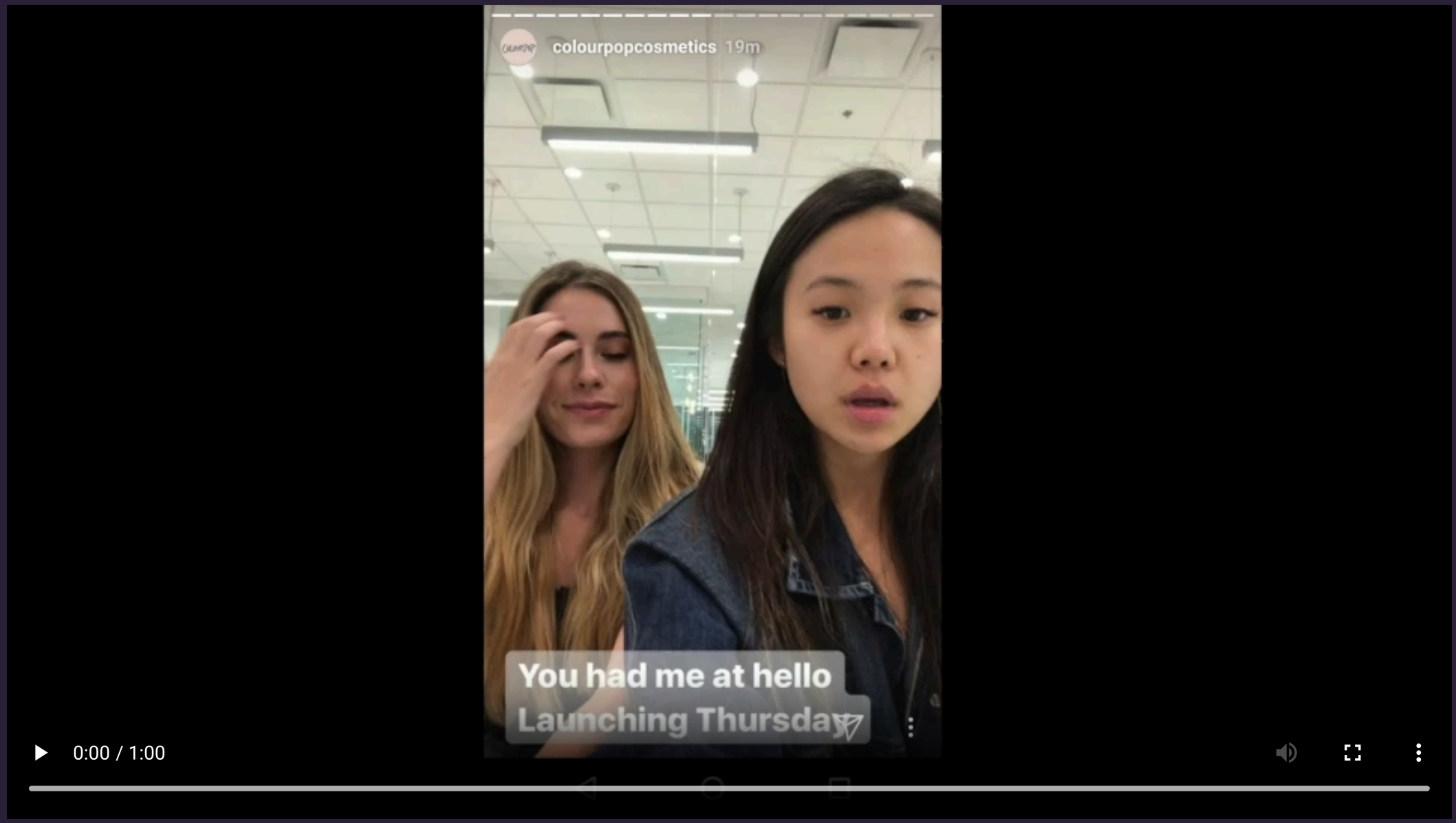
RCHSS, Academia Sinica

2025-11-20

Screenomics 3.0: A Framework for Visual Digital Trace Research

@ 中央研究院 / 人文社會科學研究中心 / 制度與行為研究專題中心
@ Center for Institution and Behavior Studies, RCHSS, Academia Sinica
2025.11.21

Why Study Real-World Digital Interactions?



Characteristics of Real-World Interactions

- Increasing reliance on digital media.
- Interactions are rapid and bursty across platforms.
- Fragmentation of content categories.
- Time domain issues: exposure over short vs. long intervals.
- Idiosyncrasy across individuals.

All of these challenge conventional social and behavioral research methods.

Challenges in Capturing Digital Trace Data

Screens as Digital Trace Data (DTD)

- DTD: “records of activity (trace data) undertaken through an online information system (thus, digital).” (Howison et al., 2011).
- Screens vs. Platform APIs & data donation:
  - Platform-specific vs. **user-specific** (Ohme et al., 2024).
  - Capture a **broader spectrum** of interactions.
  - **Multimodality**: images, text, interface elements, mixed content.
  - Flexible **unit of analysis** (screen, session, episode).
  - Ease of **passive data collection**.

The Stanford Human Screenome Project

Screenome: Capturing Real-World Interactions

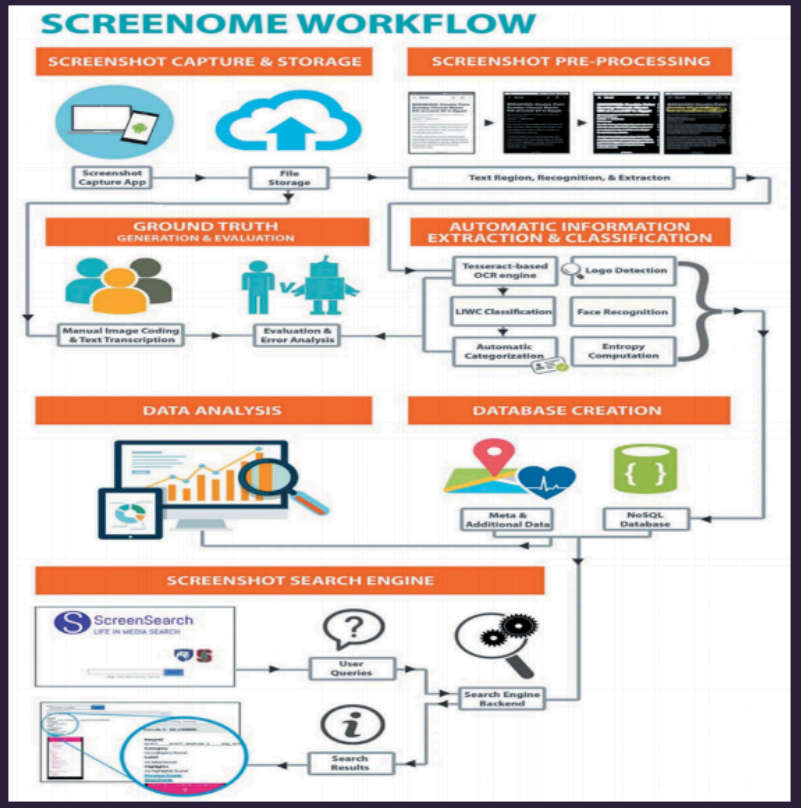
- Captures smartphone screens every **5 seconds (or less)**.
- **~500 million screens** from over **1,000 people** for up to **1 year**.
- Privacy, risk, and **data security** considerations.
- Linkage to periodic **surveys (health data)**.

Reeves et al. (2020). *Nature*; Reeves et al. (2020). *Human–Computer Interaction*.

Expanding on the Screenome Approach

Transition to Screenomics 3.0

1. **Screenome 1.0** – Research infrastructure & conventional ML-based measurements.
2. **Screenome 2.0** – Deep learning-based content analysis.
3. **Screenomics 3.0** – Multimodal encoders and large multimodal models (LMMs).

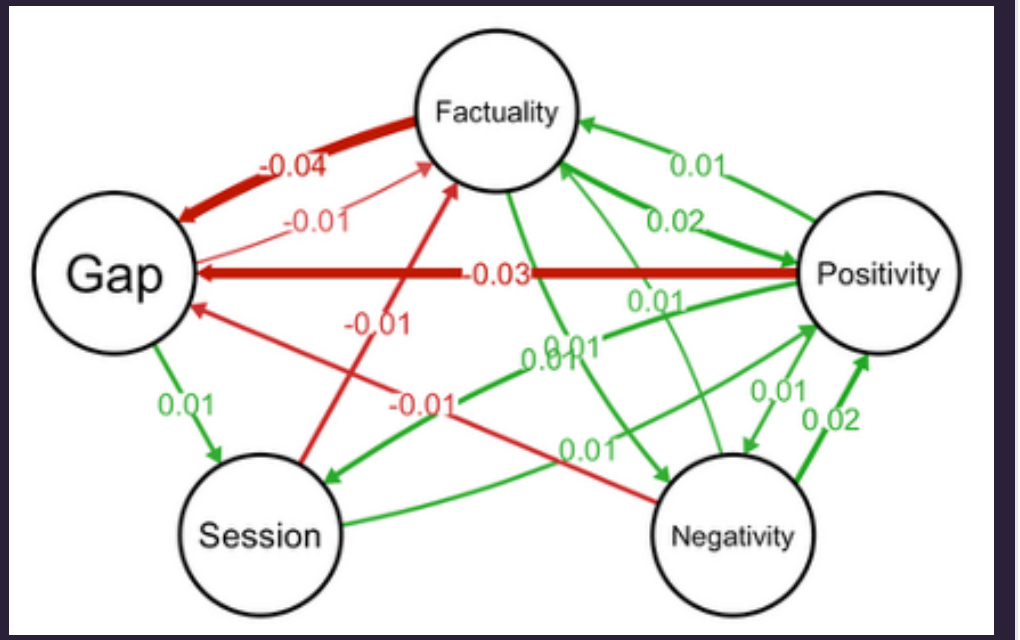


Cho et al. (in prep.).

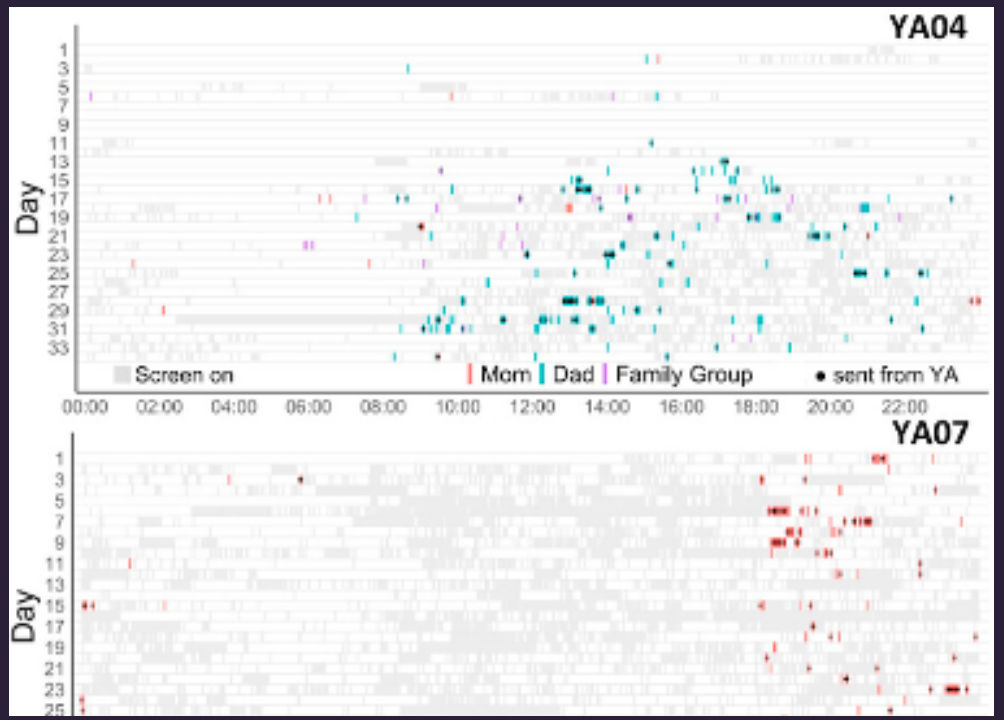
ML-Based Content Analysis & Mixed Methods

Media Sequencing and Family Dynamics

- Homeostatic mechanism of **media sequencing** in everyday life.
- Young adults' smartphone interactions with **family**.
- ML-based content analysis combined with **qualitative interviews** and diary methods.
- Focus on how digital interactions **sequenced with offline interactions** shape family dynamics.



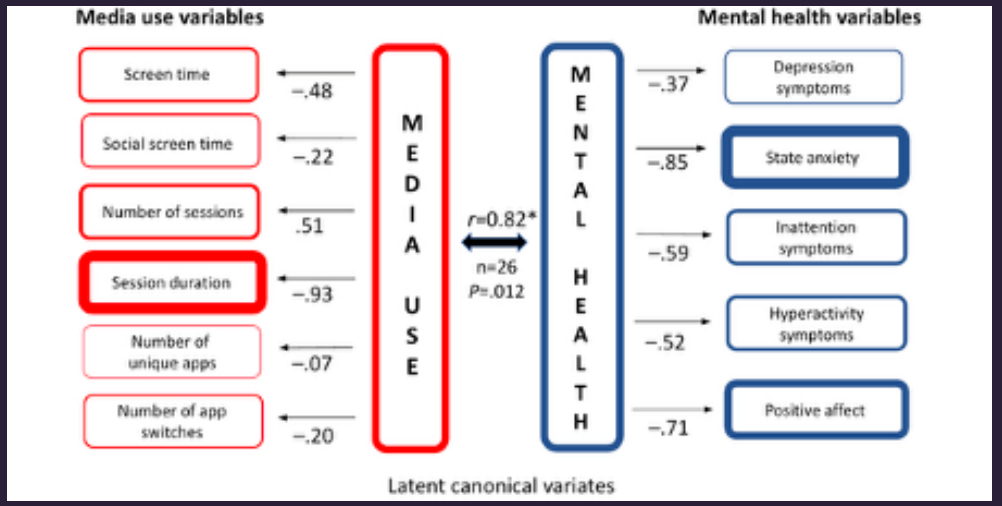
Cho et al. (2023). *Heliyon*.



ML-Based Content Analysis & Mixed Methods

Mental Health via Screenome

- Linking **survey measures** with **digital trace data (DTD)**:
  - Depression, State Anxiety, ADHD, Happiness.
  - Integrating **self-report scales** with DTD.
  - **Real-time, personalized detection** of mental health states.
  - Complementing **clinical and survey** indicators.

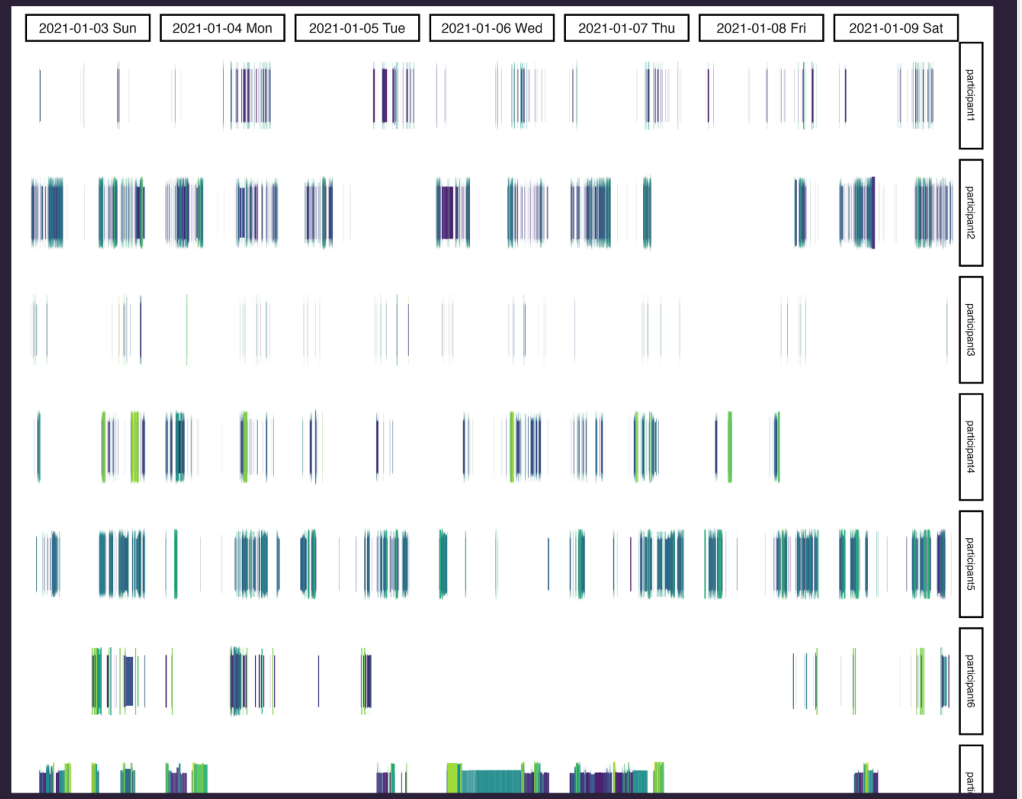


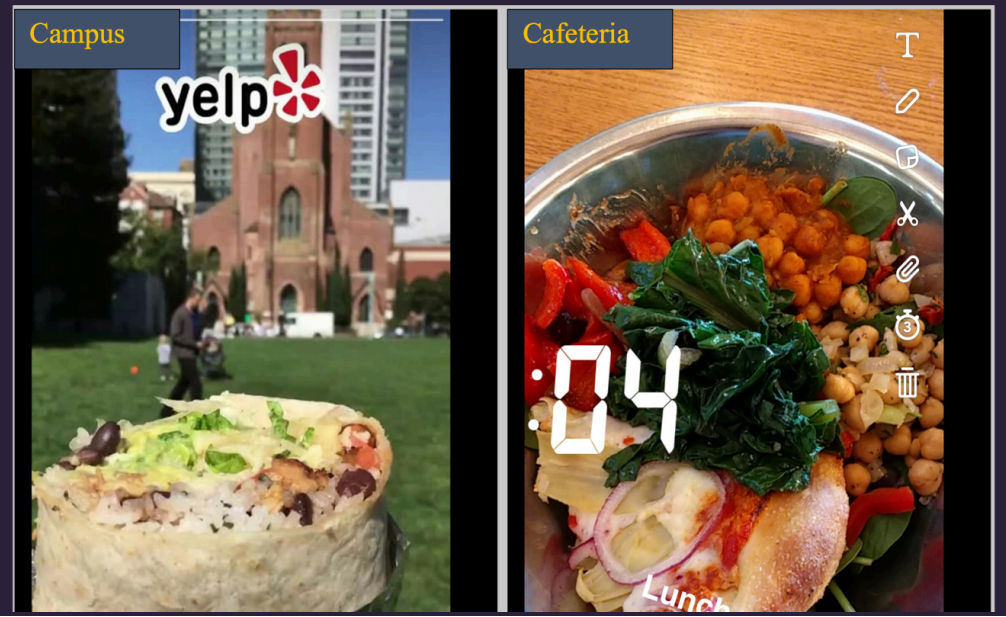
Cerit et al. (2025). *JMIR Formative Research*.

Deep Learning-Based Content Analysis

Visual Emotions, Scene Recognition, Food Detection

- CNN-based **scene recognition** to categorize environments.
- CNN-based **food detection** for identifying food-related content.
- Visual **emotion recognition** (valence and arousal) from screen images.
- Weekly trends: link **visual emotion trajectories** to well-being.

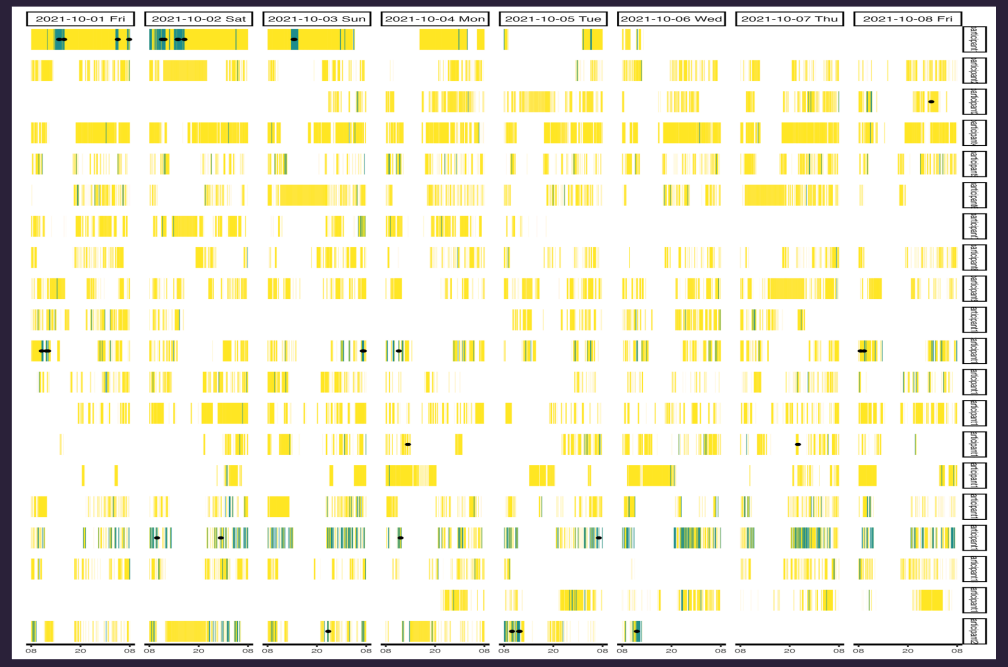




Multimodal Encoders and LMMs

Adolescents' Food-Related Content Exposure

- **20 adolescents**, 1-week observation of smartphone screens.
- **3%** of all exposure is food-related; **0.6%** branded.
- Demonstrates how a specific **content category** can be mined from large screen sets.
- Shows feasibility of **longitudinal content tracking** at screen level.

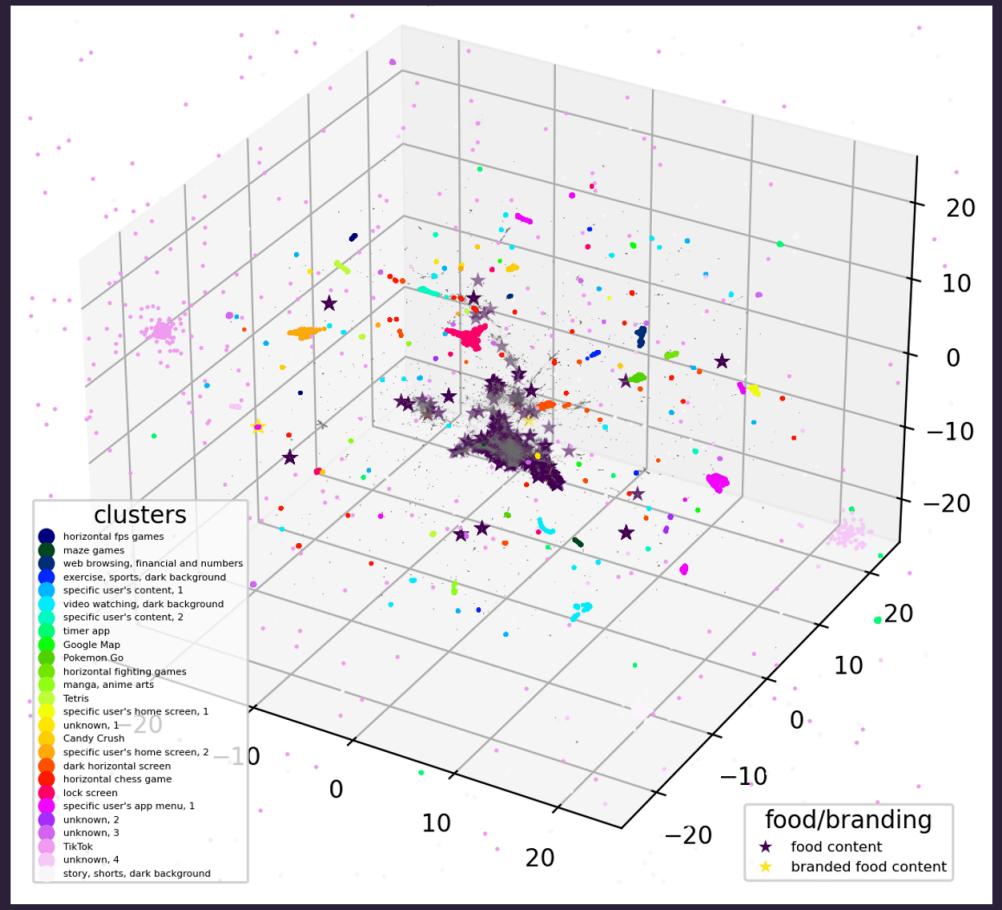


Cho et al. (in prep.).

Multimodal Encoders and LMMs

Spatial + Behavioral Clusters with CLIP & OCR

- **Screenotype**: unique screenomes associated with behaviors and experiences.
- Used **HDBSCAN + UMAP** on 320K data points for cluster analysis.
- **26 distinct clusters**; ~19% non-noise.
- Within-app **CLIP variance > between-app variance** → rich within-app heterogeneity.

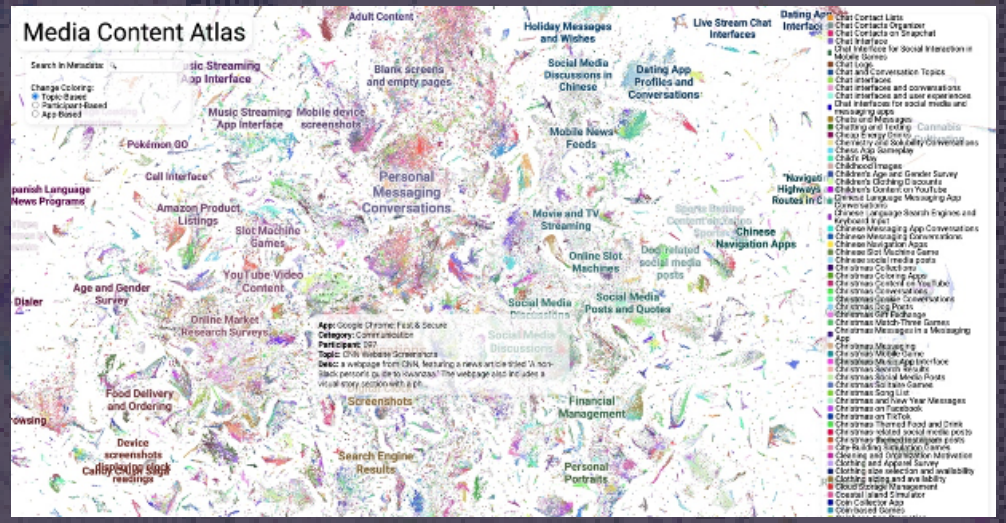


Cho et al. (in prep.).

Multimodal Encoders and LMMs

Mapping Digital Screen Content

- **Media Content Atlas** for open-ended exploration of digital media interactions.
- Content mapping with **HDBSCAN + UMAP** on **1.12M** data points from **112 participants**.
- Supports hypothesis generation about **media repertoires** and **use patterns**.

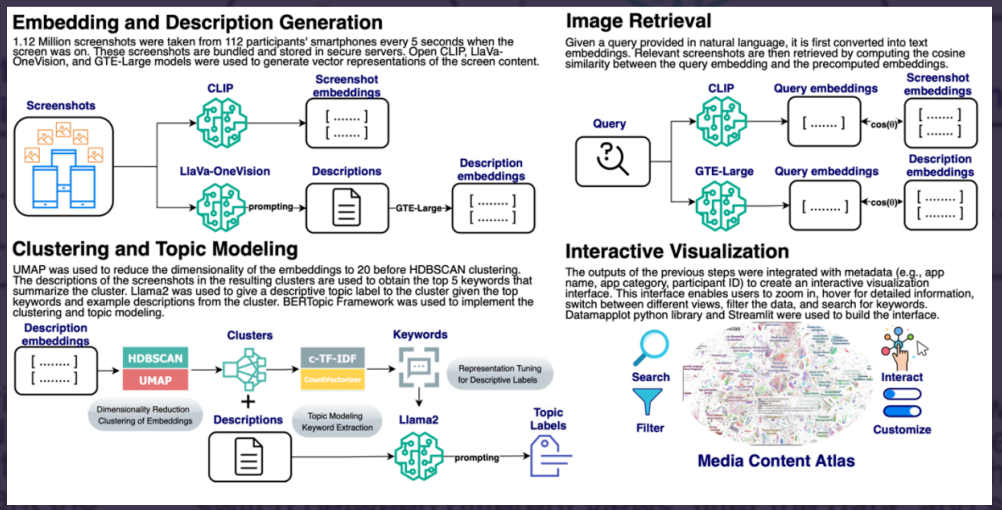


Cerit et al. (2025). CHI EA.

Multimodal Encoders and LMMs

Mapping Digital Screen Content

- **Image + text embeddings** combined.
  - Large multimodal model (LMM) descriptions.
- **Topic label generation** for clusters of screens.
- **Information retrieval**: querying screens and descriptions for:
  - Content categories.
  - usage contexts, etc.

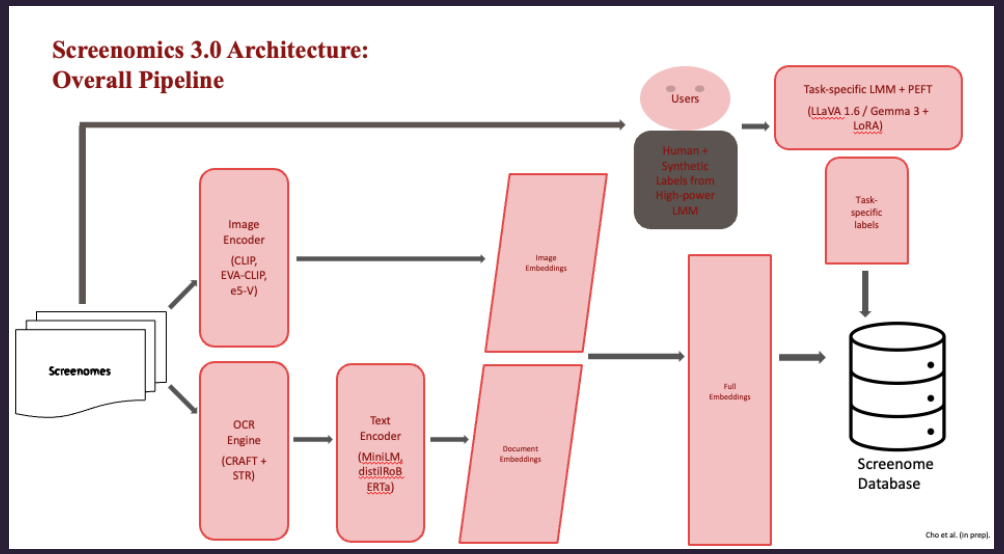


Cerit et al. (2025). CHI EA.

Screenomics 3.0 Architecture

Overall Pipeline

- **Screenomes**: longitudinal smartphone screenshots.
- **Image Encoder** (CLIP, EVA-CLIP, e5-V) → **image embeddings**.
- **OCR Engine** (e.g., CRAFT + STR) → text from screens.
- **Text Encoder** (MiniLM, distilRoBERTa) → **document embeddings**.
- **Task-specific labels** with LMM + PEFT (Gemma 3 + LoRA).
- Users provide **human labels**; high-power LMMs supply **synthetic labels**.

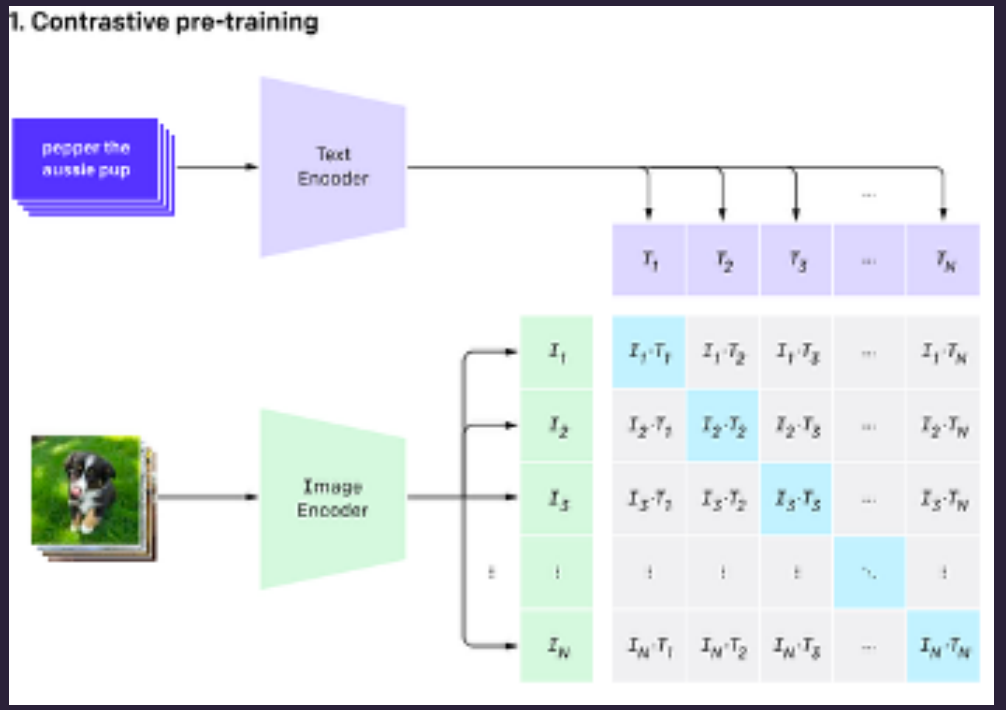


Cho et al. (in prep.).

Screenomics 3.0 Architecture

Image Encoder: CLIP

- CLIP (Contrastive Language–Image Pre-training).
- Jointly learns **image–text representations**.
- Surpasses many fully supervised baselines on zero-shot tasks.
- Core to many **state-of-the-art multimodal architectures**.

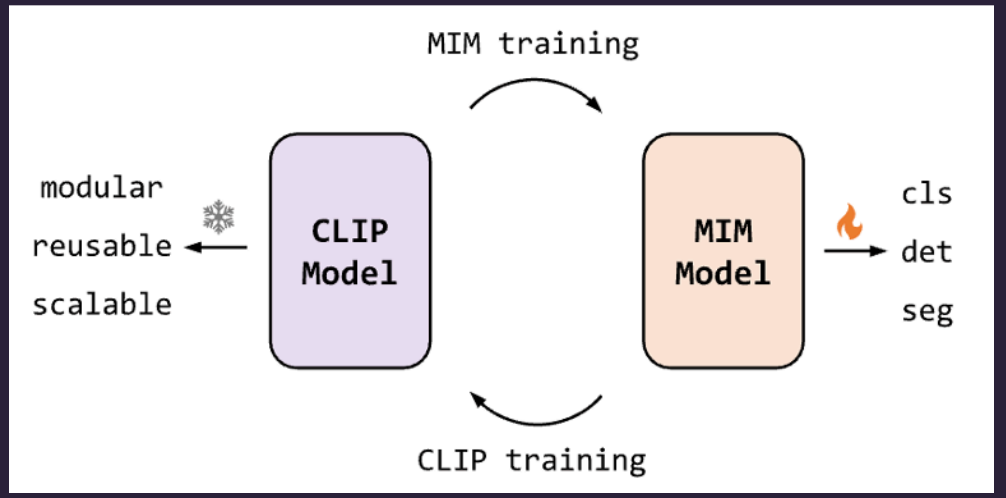


Radford et al. (2021).

Screenomics 3.0 Architecture

Image Encoder: EVA-CLIP

- EVA (Explore the limits of Visual representation at scAle).
  - Pretrains **ViTs** using masked image modeling (MIM).
  - Helps ViTs learn fine-grained image structure before contrastive learning.
- EVA-02-CLIP: outperforms similar-sized models on many tasks, including **image retrieval**.

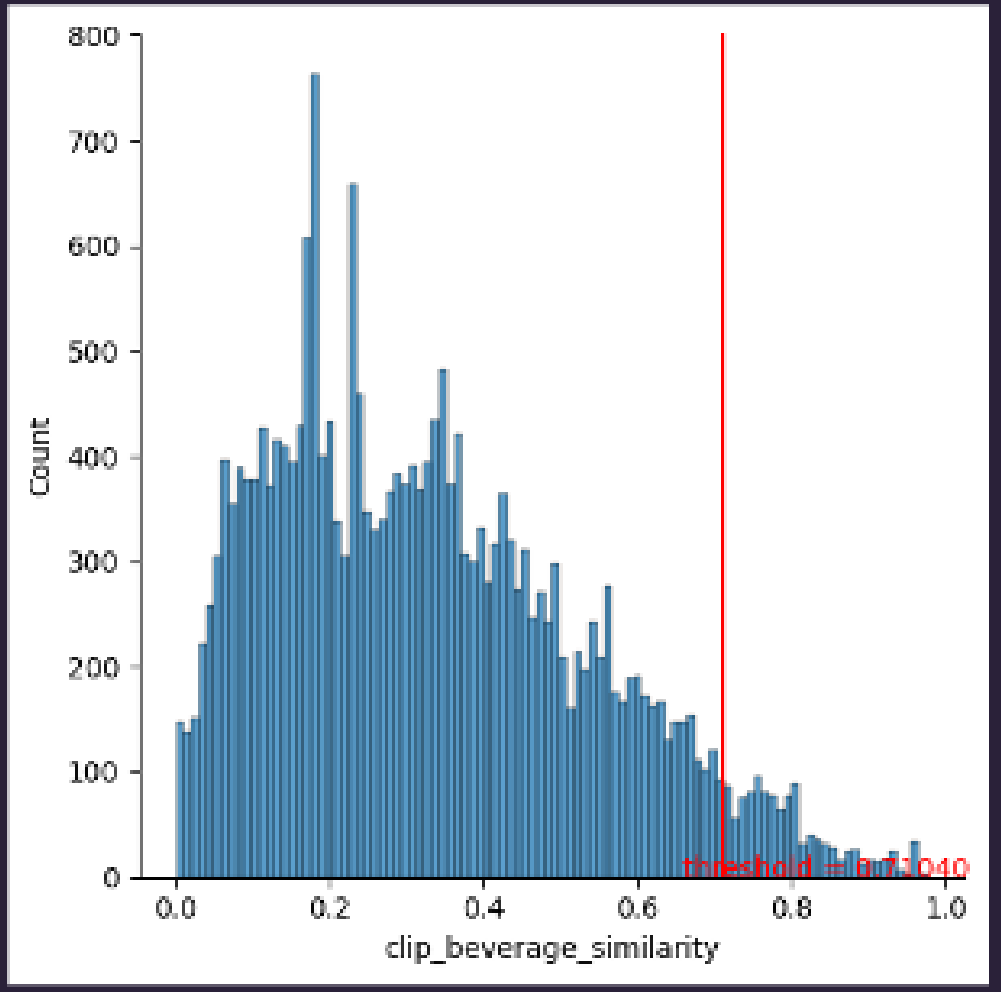


Fang et al. (2023).

Screenomics 3.0 Architecture

Text-to-Image Retrieval with CLIP

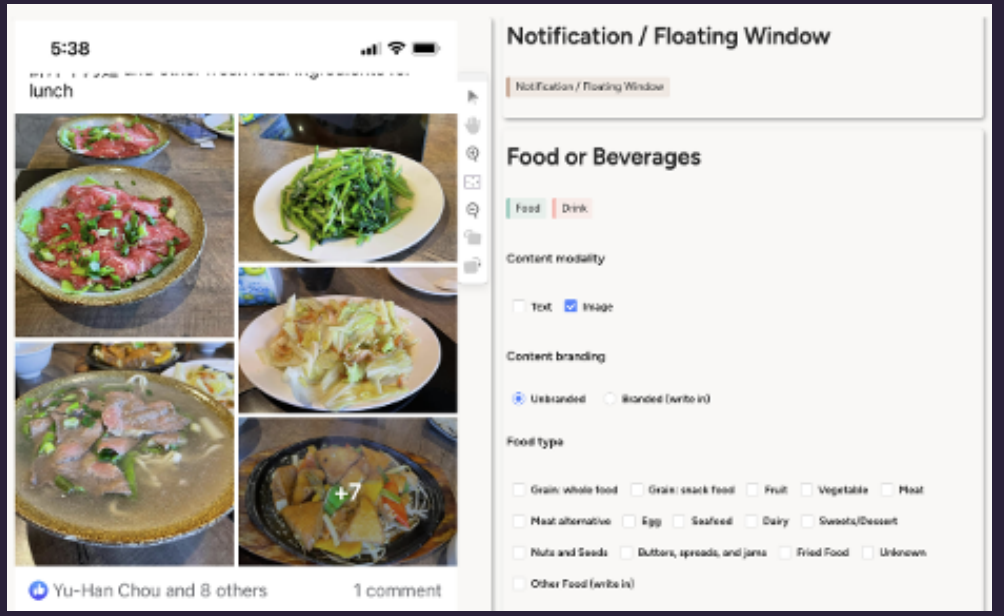
- Compute **cosine similarity** between text queries and image embeddings.
- Efficiently find **relevant (and irrelevant)** images.
- Supports rapid creation of **balanced labeled datasets**.
- Enables exploratory analyses of **usage contexts** and **media patterns**.



Screenomics 3.0 Architecture

Human Labels with Label Studio

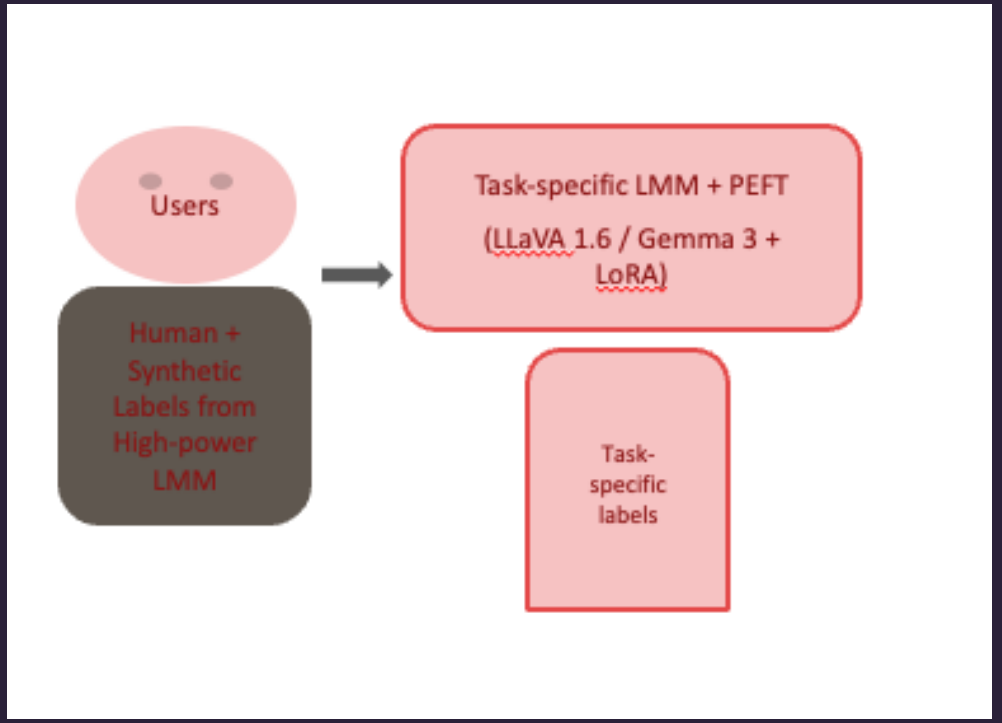
- **Label Studio** for customized labeling interfaces across projects.
- Supports **multiple annotation tasks** (e.g., food presence, valence, task type).
- Reliability is a key challenge:
  - need for consistent instructions,
  - quality control and adjudication.



Screenomics 3.0 Architecture

Task-Specific Labels with Large Multimodal Models

- **LLaVA** (Large Language and Vision Assistant): open-source LMM.
- Use cases in this project:
  - **VQA** over smartphone screens,
  - **PEFT** for task-specific labeling,
  - Large-N inference over millions of screens.
- Competitive families of models:
  - Gemma 3 (12B, 8-bit quantization).
  - Qwen2.5-VL (7B, bf16).
  - Llama 3.2-V (11B, 8-bit, split across GPUs).



Liu et al. (2023).

Screenomics 3.0 Architecture

PEFT with LoRA

- LoRA (Low-Rank Adaptation of Large Language Models).
  - Replace full-rank weight matrix **W (d × d)** with low-rank **A (d × r)** and **B (r × d)**.
  - Freeze **W**, train only **B · A**.
- Makes large-model fine-tuning **feasible** on modest hardware.
  - Pretraining → SFT (supervised fine-tuning) → RLHF.
  - Reduces trainable parameters and memory footprint substantially.
- Example performance (food-detection use case):
  - Accuracy ≈ .96, macro F1 ≈ .93, Kappa ≈ .85.

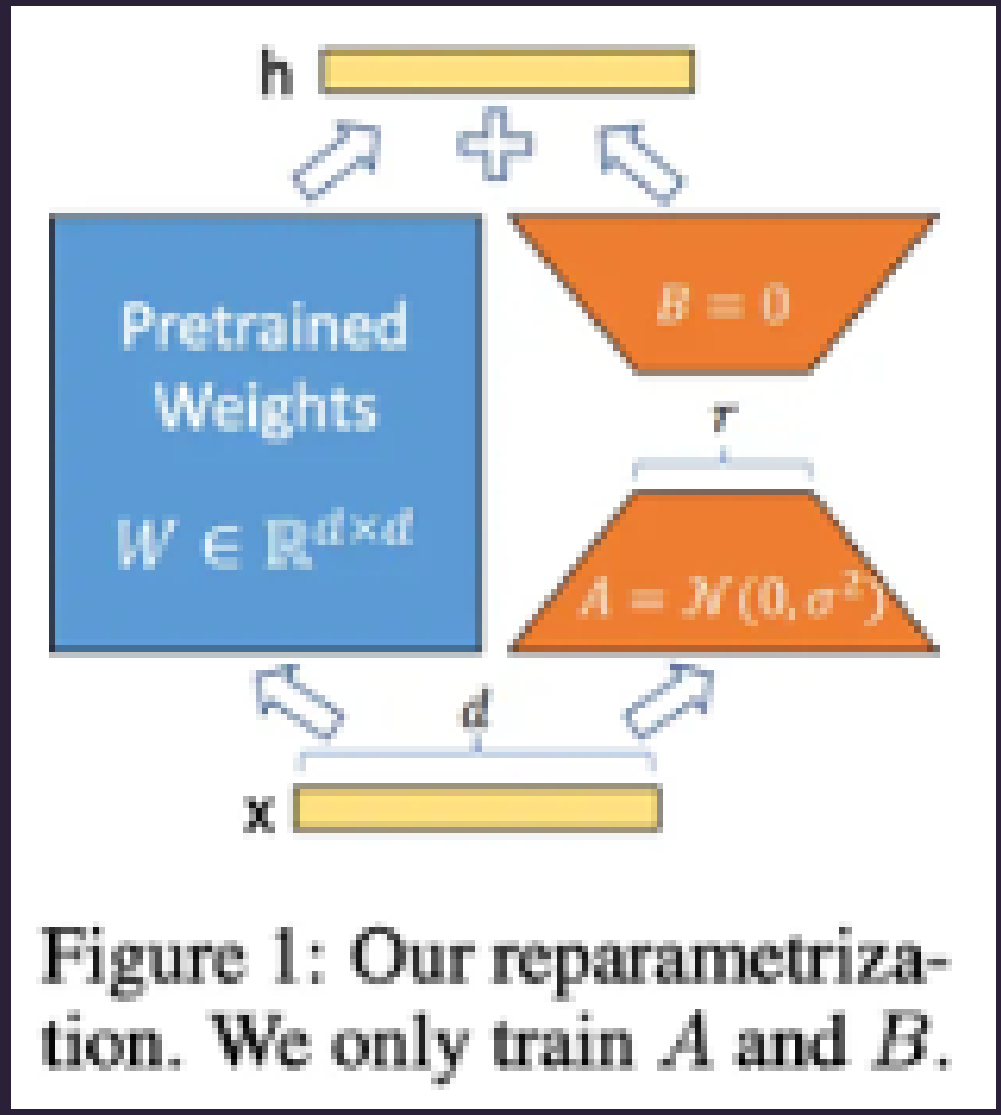


Figure 1: Our reparametrization. We only train A and B.

Hu et al. (2021).

Multimodal Encoders and LMMs

High-Level Behavioral Constructs

- LMM-based **user activity** and **intention measurement**:
  - Combine **commercial solutions** with prompt engineering.
  - Measure high-level behavioral constructs (e.g., consuming media, functional activities).
  - In one application, Krippendorff's Alpha = 0.877, precision = 0.92, recall = 0.91.
  - Context-based inference: identifying **user intention** behind each screen.

Fig. 1. Figure (a) illustrates the process of sending phone screenshots along with app information to GPT-4o for activity inference and reasoning. We used OpenAI's API to connect to GPT-4o. Each request included screenshot data, app information, and the prompt provided in the Appendix. The request also contained parameters such as temperature, top_p, and seed. The model returned its response in a JSON format. Figures (b) and (c) present an example of a request and its corresponding response.

Chang et al. (under review).

Pitfalls and Future Directions

Challenges, Limitations, & Future Possibilities

- Over-reliance on models can lead to **subtle biases** or overconfidence in results.
- Need for **interpretability tools** to better understand model decisions.
- Future expansions:
  - Efficient model training and inference,
  - Real-time analytics,
  - User feedback and participatory design.
- Importance of **interdisciplinary collaborations** for domain-specific content analysis.

Some related ideas…

The Language of Biology

Digital Biology = Biology Becomes Computable

- Integration of heterogeneous biological data (genes → cells → organs → behavior)
- Multi-scale modeling that links molecular processes to whole-organism function
- Networked, collaborative science powered by AI and high-performance computing
- Simulation + experimentation loop accelerates biological discovery
- Biology shifts from descriptive to predictive and engineering-compatible

The Language of Biology

AI as the Engine of Digital Biology

- Foundation models enable pattern discovery in massive biological datasets
- Generative models simulate biological systems and guide interventions
- AI accelerates modeling from molecules → cells → circuits → behavior
- Computational frameworks unify biological processes across scales
- AI turns biology into something legible, learnable, and designable

From Digital Trace Data to Communication Styles

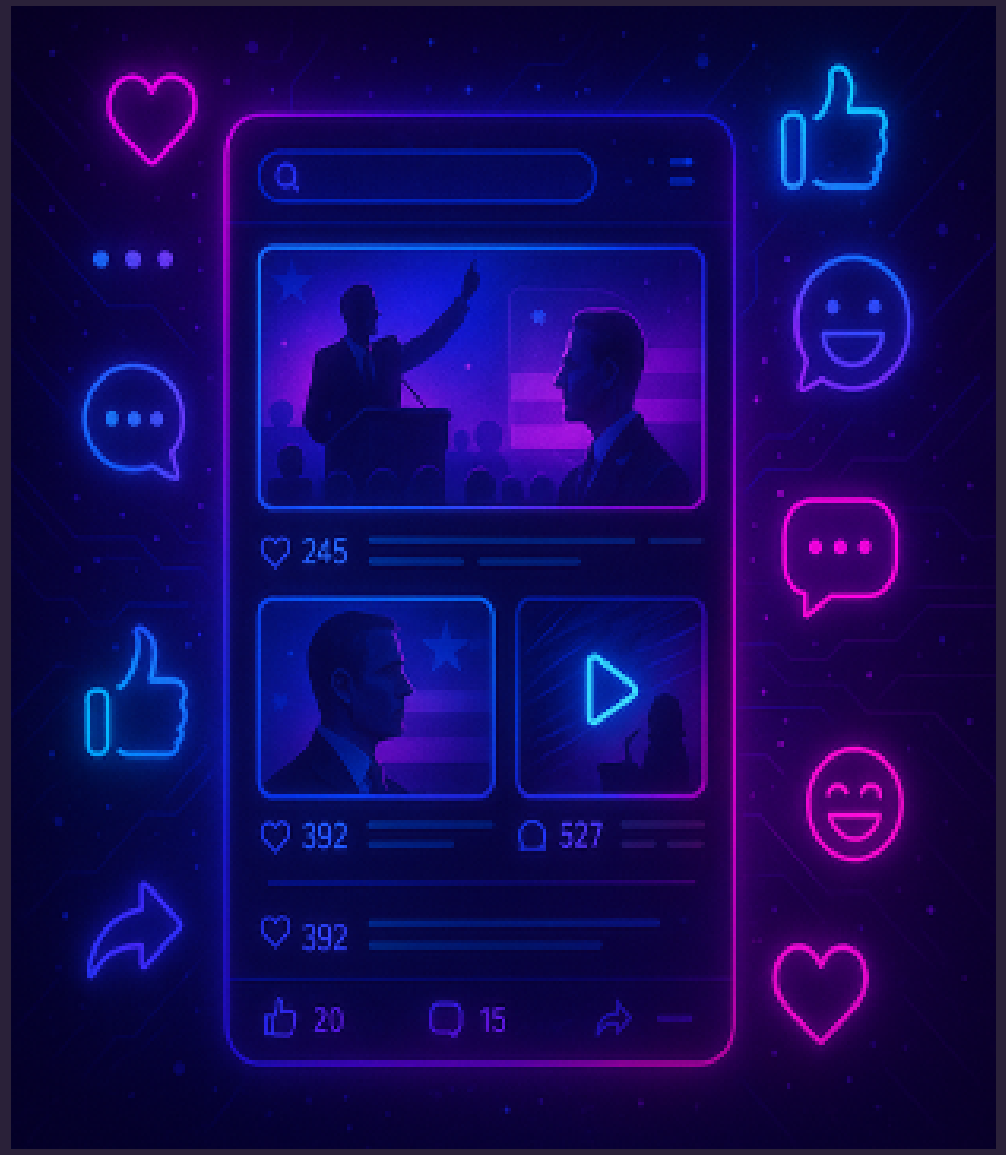
A Framework for Analyzing and Linking Multimodal Social Media Content

- MJ Cho, Chingching Chang, Yuan Hsiao, Hen-Hsen Huang
- RCHSS, Academia Sinica

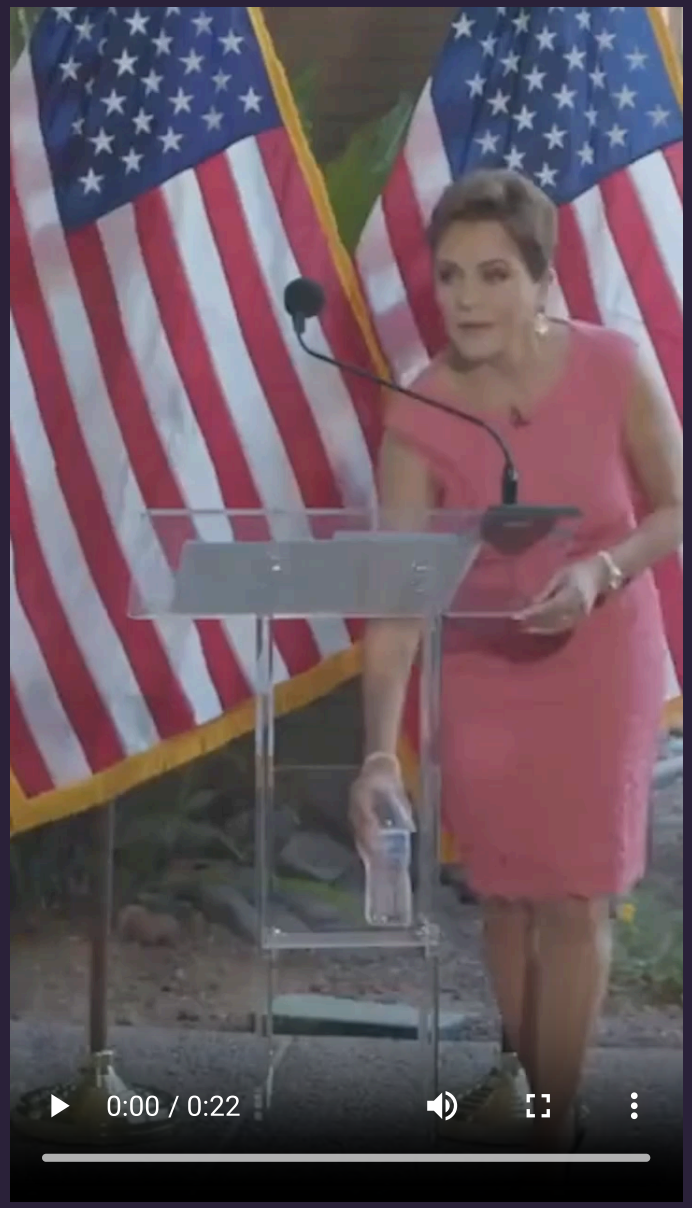
The Challenge: A New Era for Media Effects Research

Why We Need New Approaches

- The field is shifting from **quantity of media use** to **content and its effects** (Pouwels et al., 2024).
- **Nature Research Intelligence**: Multimodal communication as a frontier.
- Traditional methods are insufficient for **personalized** and **fragmented** media environments (Ohme et al., 2024; Otto et al., 2024).
- The **video problem**:
  - Dominant form of social media content.
  - Multimodal and challenging to study (Kroon et al., 2024).
  - Audio channel remains an **under-studied dimension** of communication.



Multimodality and Media Psychology



Multimodality and Media Psychology

Why Multimodal Thinking Matters

- Online media and platform–operator (PO) data are inherently **multimodal**.
- **Format / schema** shape psychological effects.
- Media theory is also multimodal:
  - visual framing,
  - vocal tone,
  - textual content.
- Psychological effects of **expression and tone**.
- Social meaning of **objects and scenes**.
- Our goal:
  - Measurement of **styles**,
  - Their **effects**,
  - A coherent analytical framework.

A Roadmap from the Literature & Our Contribution

Three-Step Approach (Pouwels et al., 2024)

1. **Collect digital trace data** (DTD, e.g., via APIs, tracking).
2. Perform **automated content analysis** (text and visuals).
3. Conduct **linkage analysis** to study effects on outcomes.

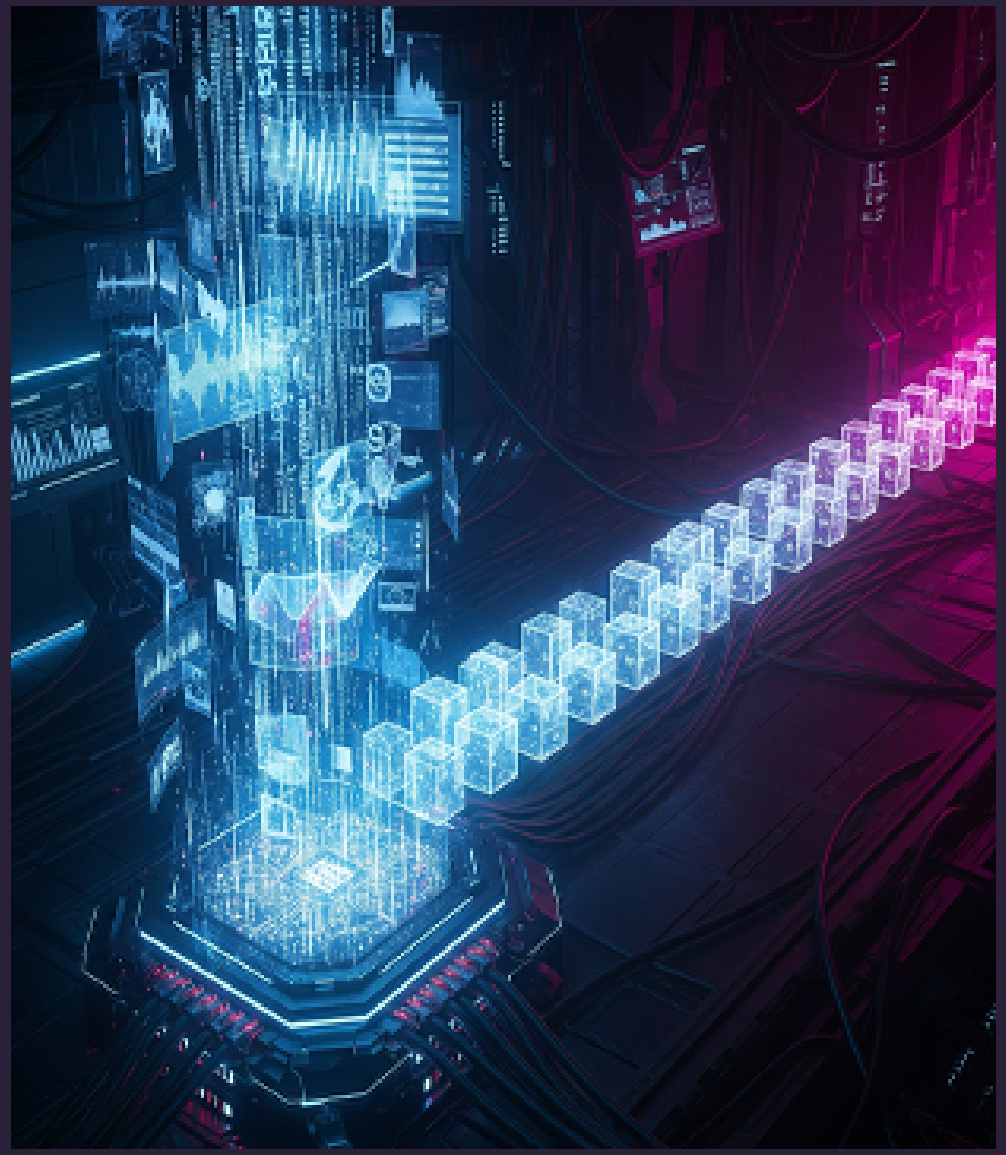
Our Contribution

- An **end-to-end framework** that operationalizes this approach for **video**.
- Goes beyond isolated text or image analysis.
- Incorporates the crucial **audio modality**.

Our Framework: An Overview

From Raw Videos to Communication Styles

- **Goal**: A replicable pipeline to transform raw videos into theoretically meaningful **communication styles**.
- **Data**:
  - 398 Instagram videos from 12 U.S. Senate candidates during the 2024 election.
- **Three core stages**:
  1. **Preprocessing** – Raw video files → analysis-ready data streams.
  2. **Multimodal Feature Extraction** – Cleaned data streams → comprehensive behavioral features.
  3. **Style Identification & Linking** – Feature sets → interpretable styles for linkage analysis.



Stage 1: From Raw Video to Analysis-Ready Streams

Preprocessing and Filtering

- **Goal**: Preprocess and transform noisy social media video into **isolated communication sources**.
- **Visual filtering**:
  - Isolate candidate presence using **face detection** (MediaPipe) and **recognition** (DeepFace).
- **Audio filtering**:
  - Isolate candidate's voice using **denoising** (Demucs) and **speaker diarization** (PyAnnote).
- **Text generation**:
  - Transcribe only the filtered candidate audio with **ASR (Whisper)**.
- **Result**: A validated set of synchronized **video, audio, and text** data — only the **target communicator**.

Stage 2: Quantifying Sight, Sound, and Speech

Multimodal Feature Extraction

- **Visual features (nonverbal performance)**:
  - Facial Action Units (AUs) → emotional expression.
  - Head pose → engagement cues.
  - Valence & arousal (EmoNet) → affective state.
- **Audio features (vocal performance)**:
  - Prosodic cues → pitch (F0), intensity (MFCC0), speech rate.
  - Vocal emotion → valence & arousal (wav2vec2).
- **Textual features (verbal content)**:
  - Topic modeling (BERTopic) → substantive themes.

Stage 3: Identifying Communication Styles

Hybrid Quantitative–Qualitative Method

Stage 3: Tri-Modality Communication Styles

Styles from Visual, Audio, and Text Combined

Cluster	Style Label	Key Traits	Campaign Strategy
0	**Formal Low-Energy Neutral**	↓ Clout, ↓ Certitude, ↓ Tone, ↓ MFCC0, ↓ Facial Action Units (AUs), ↑ Neutral emotion	Delivers content in a flat, factual manner, avoids emotion or confrontation
1	**Casual Expressive Happy**	↑ Pronouns, ↑ Filler, ↑ MFCC0, ↑ AU06 (smile), ↑ Happiness, ↓ Negative Tone	Friendly and warm, aims to build rapport with casual delivery
2	**Analytic Calm Neutral**	↑ Analytic, ↑ Word Count, ↓ Arousal (text/audio), ↑ Neutral Face & Tone	Provides informative, calm delivery to establish expertise
3	**Emotive Dynamic Angry**	↑ Risk, ↑ Emo_anger, ↑ MFCC0, ↑ AU04/20 (tension), ↑ Anger facial/emotional scores	Uses emotionally intense delivery, likely to provoke or mobilize
4	**Empathic Warm Positive**	↑ Social/Family words, ↑ Emo_pos, ↑ AU12 (smile), ↑ Tone, ↓ Fatigue	Builds trust via warmth and positivity, targeting shared values

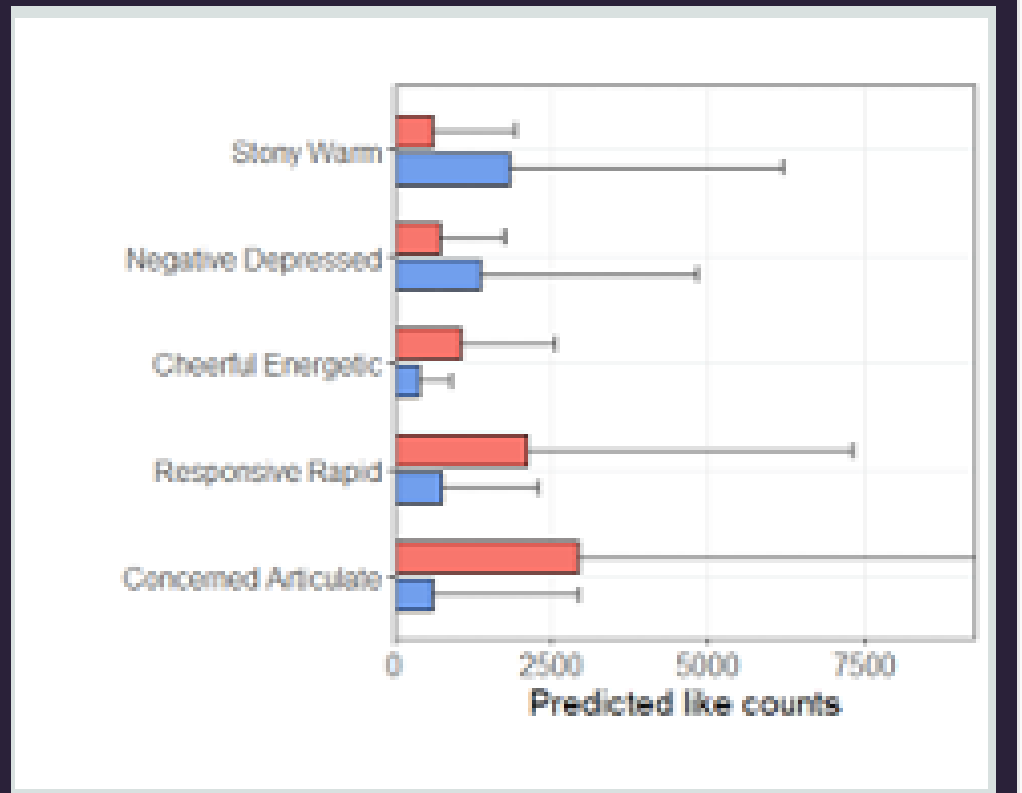
Validating the Styles

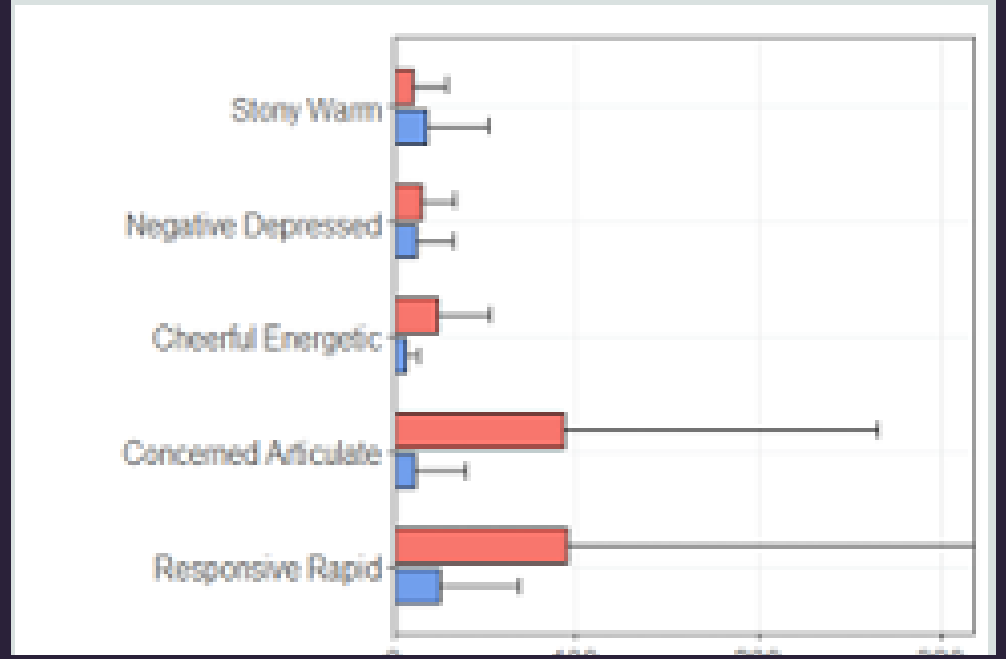
Distribution Across Parties

Style Effectiveness and Party Differences

How Style Effectiveness Varies by Party

- The **effectiveness** of a communication style often depends on **party**.
- Examples:
  - "Neutral Calm" audiovisual style generates **high engagement** for Republicans but **very little** for Democrats.
  - Democrats benefit more from a "Cheerful Energetic" approach, which is less effective for Republicans.
- Implication: Style is not universally good or bad; its impact is **conditional on identity and context**.





Contributions & Implications

What This Framework Enables

- Direct, practical response to the **CMM special issue** call for computational social media effects research.
- **Advances step 1**: Pipeline for creating valid, analysis-ready DTD streams from raw video.
- **Innovates step 2**: Multimodal method that integrates **sight, sound, and speech**.
- **Enables step 3**: Generates validated, predictive variables for **theory-driven linkage analyses**.
- Opens new research avenues:
  - Move beyond **what is said** to quantify **how it is said**.
  - Paves the way for new theories of **multimodal political communication** and media effects.

Why All These Matters?

A tale of scientific evolution

- When Biology Becomes Computable, Mind Becomes Modelable
  - Cognition emerges from biological information processing
  - Multi-scale biological models create mechanistic foundations for psychological theory
  - Integrated biological + digital-trace data enables predictive models of behavior
  - Synthetic agents & personas become valid testbeds for psychological mechanisms
  - Psychology shifts from correlational to mechanistic and intervention-driven

網路輿情的人智互動：即時輔助社會議題理解

研究團隊與背景

- 人社中心跨領域團隊：卓牧融、張卿卿、蕭遠、黃瀚萱、張永儒
- 目標：提升社會議題理解、促進公共理性對話
- 基礎：台灣規模最大的網路輿情資料庫

社會爭議的根源：理解不足

認知偏差 + 新媒介生態 = 理解鴻溝

- 認知協調、選擇性暴露
- 演算法濾泡與迴聲室效應持續存在
- 新媒介型態造成資訊碎片化
- 認知權威分散與不穩定 → 難以形成公共共識

以人與 AI 的互動提升議題理解

多模態輿情資料 × RAG × Multi-agent System

- 台灣最完整的網路輿情資料庫
- 檢索增強生成（RAG）提升資訊可得性
- 多代理人系統（multi-agent）協作產生深度理解
- 利用互動的框架化來引導認知過程

以人為本的多模態實證研究

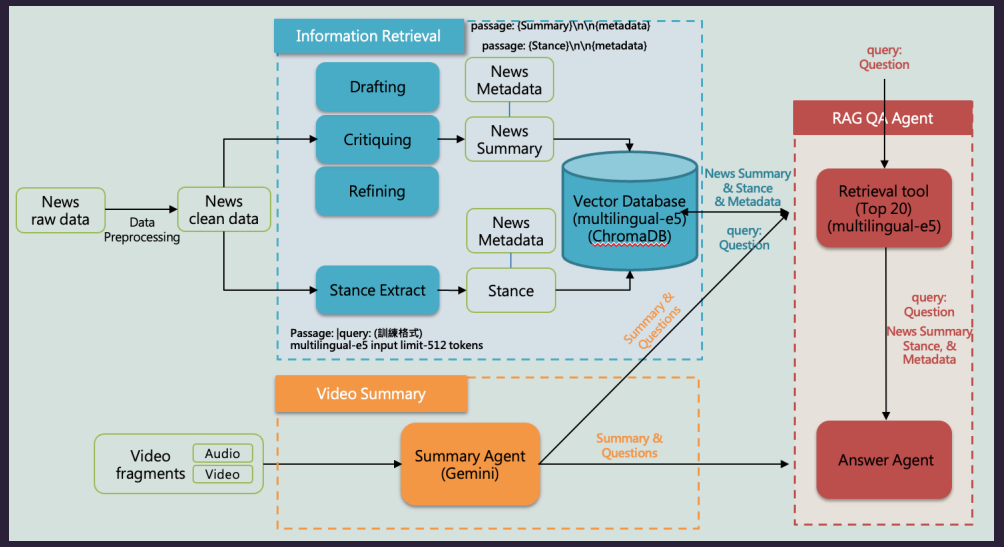
全方位方法：從社會科學到人機互動

- 多模態資料在社會科學中的應用迅速成長
- 跨模態分析（文本、影像、社群動態）
- 結合人智互動以提升議題理解深度

Multi-agent 系統架構

RAG × 向量資料庫 × 多代理協同

- 多階段流程：Drafting → Critiquing → Refining
- 向量資料庫（multilingual-e5 + ChromaDB）
- RAG Q&A Agent、Summary Agent（Gemini）
- 支援逐段精煉、立場抽取、影片摘要

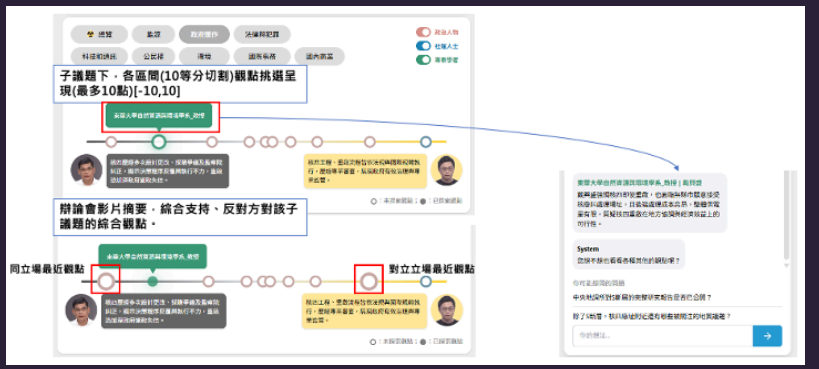


人智互動引導：觀點光譜與 QA

人智互動引導：觀點光譜與 QA

幫助使用者理解多元觀點，降低二元對立

- 系統協助呈現多元立場與論點光譜
- 引導式 QA 幫助理解議題的關鍵爭點
- 促進更成熟、更細膩的公共討論能力



Why Do All These Matters?

A tale of scientific evolution

- A New Phase for Social Science
  - Behavior becomes measurable across biological, cognitive, and digital layers
  - Emotion/thought contagion, media effects, social connectedness become model-ready
  - Computational social science gains grounding in biological and cognitive models
  - Interventions can be designed across biological, psychological, and social scales
  - Social science enters its own “digital biology moment” — a designable science of behavior

Thank You

Questions?

Mu-Jung 'MJ' Cho | 卓牧融

mjcho@as.edu.tw