Multimodal Large Language Models for Understanding Human Self-Concept Evolution in Digital Social Environments

Milos J. Baker
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
miloswork@uc.edu

Casper J. Kennedy
Department of Computer Science, University of Houston, Houston, TX, USA.
casperkennedy960@uh.edu

Abstract

The digitally mediated social world is now a primary arena in which human self-concept is formed, revised, and expressed. The proliferation of multimodal platforms, from image-centric social networks to short-video ecosystems, has transformed identity construction into a continuous, cross-channel, and algorithmically shaped process. Understanding how individuals' self-representations evolve across these environments poses profound challenges for methods rooted in single-modality analysis or episodic self-report. This paper examines the potential of multimodal large language models (MLLMs) as systems instruments capable of capturing the textured, longitudinal dynamics of self-concept within digital social environments. We adopt a systems-level perspective that moves beyond model accuracy to interrogate the broader sociotechnical infrastructure required to deploy MLLMs responsibly. The discussion is organized around structural trade-offs, data governance, architectural choices, fairness considerations, robustness, sustainability, and policy implications. We argue that MLLMs, by virtue of their capacity to process and integrate text, image, video, and interaction metadata at scale, can serve as computational mirrors that reveal how platform affordances and algorithmic curation co-produce evolving self-views. However, harnessing this capacity demands rigorous attention to the inherent tensions between representational richness and privacy, between personalization and ecological validity, and between interpretability and model scale. The paper provides a conceptual analysis of these tensions, sketches design principles for a federated MLLM-based research infrastructure, and maps the regulatory landscape that must be navigated if such systems are to be deployed in ways that uphold human autonomy, fairness, and long-term trust.

Keywords

multimodal large language models, self-concept, digital identity, social media, system architecture, algorithmic fairness, data governance.

1. Introduction

The digital environment has graduated from being a supplementary context for social life to constituting a central infrastructure for identity formation. Early scholarship on internet identity emphasized textual self-presentation in chat rooms and online forums, treating the digital realm as a space of disembodied experimentation [1]. Contemporary platforms, by contrast, are saturated with multimodal signals: users curate visual profiles, share ephemeral

stories, livestream reactions, comment with animated stickers, and generate synthetic content through algorithmic templates. The self that emerges from this dense weave of practices is not a static entity but an ongoing negotiation shaped by platform affordances, community norms, recommendation algorithms, and real-time feedback loops [2]. Recognizing self-concept as a dynamic, multimodal phenomenon invites a re-examination of how large-scale computational systems can be designed to observe and interpret it without reducing it to mere digital exhaust.

Multimodal large language models represent a class of foundation models capable of ingesting and relating heterogeneous data types, including natural language, images, audio, and structured metadata [3, 4]. They extend the linguistic prowess of autoregressive transformers into perceptual domains, enabling cross-modal alignment that was previously possible only through bespoke pipelines. In principle, such models could be deployed to trace the evolution of an individual's self-concept across platforms by identifying recurrent narrative themes, visual self-symbols, affective tones, and relational patterns embedded in multimodal timelines. Yet the translation of this potential into a responsible research infrastructure is far from trivial. It requires grappling with systems questions that span data acquisition protocols, model architecture selection, longitudinal representation learning, fairness auditing across demographic intersections, the carbon footprint of continuous retraining, and the governance frameworks that determine whose self-concept is rendered legible and under what conditions.

This paper provides a systemic examination of these questions. We do not propose a finished implementation or a single optimal architecture. Instead, we develop a layered analysis that connects foundational theories of self-concept with the operational realities of deploying MLLMs in social computing contexts. Our contribution is structured as follows. After establishing the theoretical and technological backdrop, we analyze the architectural trade-offs inherent in building MLLM pipelines for longitudinal identity research. We then address the data governance and ethical framing that must surround any such undertaking, emphasizing the need to embed values of agency and contextual integrity at the infrastructure level. Subsequent sections explore fairness, robustness, sustainability, and policy implications, culminating in a forward-looking synthesis that identifies high-leverage nodes for intervention by researchers, platform operators, and regulators.

2. Self-Concept in Algorithmically Mediated Environments

Human self-concept has been theorized in the psychological sciences as a multifaceted cognitive structure comprising self-representations, personal narratives, and relational schemas that organize experience and guide behavior. From a cultural psychological standpoint, the self is not an insulated monad but is profoundly shaped by social practices, symbolic resources, and institutional arrangements [4]. In digital societies, these arrangements increasingly include the algorithmic curation of social feedback. Platforms do not simply mirror pre-existing selves; they actively structure which facets of a person are amplified, which audiences are assembled, and what metrics of social worth are made salient. This dynamic is at the heart of self-concept evolution, as feedback loops between self-presentation and audience reaction can recalibrate an individual's sense of who they are, sometimes stabilizing a coherent identity and other times fragmenting it across incompatible contexts [5].

A particularly salient mechanism is the interplay between possible selves and algorithmic recommendation. Individuals project aspirational self-images through content creation, consuming media that evoke desired future identities. Recommendation systems, optimizing for engagement, may systematically expose users to narrowly calibrated exemplars that

polarize these possible selves or anchor them to unrealistic benchmarks [6]. Over time, repeated exposure modulates self-discrepancies, the perceived gaps between actual, ideal, and ought selves that are central to self-regulatory processes [5]. Meanwhile, the filter bubble phenomenon, originally described in primarily textual terms, now extends across modalities; curated image feeds, short-video cascades, and influencer economies create self-referential semiotic systems that can homogenize expressions of identity even as they appear personalized [7, 8].

Narrative identity theory further illuminates the diachronic dimension of self-concept, emphasizing that people construct storied accounts of their lives to confer meaning and continuity [9]. Digital platforms host these stories in fragmented and interactive forms: a series of Instagram posts documenting a transformation, a TikTok narrative arc spanning months, or a Reddit thread chronicling a personal journey. Capturing the narrative spine beneath such fragmented data is a formidable analytical challenge. An empirical study that examined self-perception in digital interaction contexts demonstrated that individuals' conscious self-reports and behavioral trace data often diverge, with platform metrics exerting an independent influence on how users evaluate themselves [10]. This divergence signals the need for computational approaches that can integrate explicit self-report with the implicit, multimodal traces that platforms archive, an intersection at which MLLMs become especially relevant.

3. Multimodal Large Language Models: Enabling Cross-Modal Interpretation of Identity Signals

MLLMs represent a synthesis of large-scale language pretraining with visual and auditory encoders, sometimes augmented by additional modality-specific towers, trained to produce unified representations that support tasks ranging from visual question answering to video-grounded dialogue. Architectures such as Flamingo, BLIP-2, and LLaVA demonstrate that frozen language models can be effectively conditioned on visual features through lightweight adapter modules, preserving the rich world knowledge and reasoning patterns acquired during text-only pretraining while injecting perceptual grounding [11, 12, 13]. More monolithic efforts, including GPT-4 and Gemini, incorporate multimodal inputs natively into decoder-only or mixture-of-experts designs, achieving fluid interchange between modalities within a single generative process [14, 15].

Underlying these models is the principle of contrastive or generative alignment between representations of different modalities, a lineage that includes foundational work such as CLIP, which demonstrated that image-text pairs at web scale can yield robust zero-shot transfer capabilities [16]. In the context of self-concept analysis, such alignment offers a pathway to operationalize constructs that are inherently cross-modal. For example, a visual self-symbol such as a profile picture depicting a particular aesthetic can be linked to the linguistic self-descriptions that accompany it, while temporal sequences of such symbols can be embedded in a trajectory space that reveals stylistic shifts correlated with life transitions. Because MLLMs can ingest both the content and the metadata of social interactions, including timestamps, engagement signals, and platform-specific framing, they hold the promise of modeling self-concept evolution as a multimodal dynamical system rather than as a series of decoupled textual snapshots.

Nevertheless, the application of MLLMs to identity research must contend with fundamental tensions. The pre-training corpora of these models are dominated by publicly available web data, which embeds their own representational biases and Western-centric visual cultures that

may distort interpretations of self-expression in communities with different aesthetic norms [17]. Moreover, the generative capacities of MLLMs blur the line between analysis and synthesis; a model trained to summarize self-concept trajectories could subtly rewrite narratives in ways that align with its training distribution rather than reflecting the participant's authentic sense-making. Addressing these risks requires system-level safeguards that go beyond fine-tuning, encompassing provenance tracking, human-in-the-loop validation, and interpretability mechanisms that expose the evidential basis for model-generated inferences.

4. System Architecture for Longitudinal Self-Concept Analysis

Designing an MLLM-based system for understanding self-concept evolution demands a layered architecture that separates data ingestion, feature representation, temporal modeling, inference, and feedback interfaces. At the ingestion layer, the system must handle multimodal data streams from multiple platforms while enforcing granular consent and data minimization principles. This suggests a federated design in which raw data remains on user-controlled devices or platform-specific vaults and only model updates, summary statistics, or differentially private representations travel to a central analysis node. Federated pretraining and fine-tuning paradigms, while still maturing for multimodal model components, offer a blueprint for decoupling computational scale from data centralization [18].

The representation layer must generate temporally aligned embeddings that capture user identity signals across modalities. Here, design choices include whether to rely on late-fusion architectures, where modality-specific encoders produce vectors that are concatenated before being fed to a temporal model, or to adopt early-fusion strategies that interleave modality tokens at the input level. Late fusion simplifies incremental deployment and allows modality-specific privacy budgets, but it may miss cross-modal interaction effects that are critical for understanding phenomena like ironic self-presentation, where a cheerful caption accompanies a melancholic image. Early fusion, while potentially richer, increases the attack surface for model inversion and requires heavier computational resources. Hybrid schemes that learn cross-modal attention patterns through a dedicated gating mechanism offer a pragmatic compromise, provided they are auditable and do not become opaque black boxes.

Temporal modeling of self-concept requires tracking both gradual drift and abrupt phase transitions. Transformer-based architectures with relative positional embeddings can capture long-range dependencies in unevenly sampled time series of multimodal posts. Self-supervised objectives such as masked trajectory modeling, where the model is trained to predict missing segments of a user's identity trajectory based on surrounding context, can encourage representations that encode the narrative continuity individuals themselves experience. However, an overemphasis on continuity risks smoothing over legitimate discontinuities, such as an identity reboot after a major life event. Architectures must therefore incorporate mechanism for detecting and respecting change points, possibly through auxiliary change-point detection modules that condition the trajectory encoder.

Inference and interpretability constitute the apex of the architecture. Given the sensitive nature of self-concept data, any system output must be accompanied by evidence tracing and confidence quantification. Techniques such as perturbation-based attribution maps, cross-modal attention visualizations, and contrastive explanations can help human analysts understand which signals drove a particular inference [19]. Incorporating a human-in-the-loop review stage before any individual-level findings are reported is not optional; it is an

architectural necessity required by both ethical commitments and emerging regulatory frameworks.

5. Data Governance, Infrastructure, and Ethical Framing

The deployment of MLLMs for self-concept research engages a lattice of data protection regimes, including the GDPR in Europe and sectoral privacy laws in other jurisdictions. Self-concept data are intrinsically intimate, often revealing mental health status, political orientation, or marginalized identity dimensions. Even when data are pseudonymized, the high dimensionality of multimodal embeddings renders re-identification risks substantial, particularly when cross-referenced with public social media corpora [20]. A strict reading of data minimization would suggest that only derived representations, and never raw content, should be stored for research purposes. Yet the very richness that makes MLLMs appealing also depends on preserving enough perceptual detail to distinguish subtle identity signals, creating a tension that cannot be resolved through technical means alone.

Governance models must therefore embed procedural safeguards that go beyond consent forms. Multi-stakeholder data trusts, in which participants retain collective agency over how their identity data are used and for what purposes, have been proposed as vehicles for aligning research infrastructure with community values [21]. In practice, this would require the system architecture to support dynamic consent management, allowing participants to withdraw not merely from a study but from specific model components, and to audit how their contributions influenced learned representations. Such granularity challenges current machine learning paradigms, which treat training data as a monolithic corpus to be consumed once. It may necessitate advances in machine unlearning, continual learning under data removal, and segmentation of model parameters corresponding to individual data subjects.

Ethical framing also must contend with the power asymmetries inherent in algorithmically mediated identity research. Platforms currently control both the apparatus of self-expression and the instrumentation for measuring its effects. An MLLM-based research infrastructure, if centralized, could replicate this asymmetry by consolidating interpretative authority in the hands of a few research institutions or corporate laboratories. Counterbalancing this tendency requires investment in open, community-governed model versions trained on consented data, paired with interpretability toolkits that are accessible to social scientists, educators, and affected communities. The goal should be to equip individuals and collectives with the capacity to reflect on their own digital identity trajectories, not to produce proprietary psychological profiles.

6. Fairness, Robustness, and Socio-Technical Trade-offs

Fairness in the context of self-concept analysis extends beyond the familiar taxonomies of demographic parity and equalized odds to encompass representational harm, the risk that a model systematically distorts or erases certain ways of being. MLLMs trained predominantly on data from commercial platforms may overrepresent the identity aesthetics valued by engagement-maximizing algorithms, such as highly polished visual self-presentation, while underrepresenting subcultural, non-normative, or privacy-oriented modes of self-expression [17]. The consequence is a measurement apparatus that renders some people's self-concept evolution hyper-visible while rendering others' illegible. Mitigating these harms requires deliberate curation of training and fine-tuning corpora that include counter-stereotypical and culturally diverse identity data, combined with evaluation benchmarks that measure not only

aggregate performance but also worst-group fairness across intersectional demographic categories.

Robustness concerns arise from the dynamic and adversarial nature of digital social environments. Identity signals can be deliberately manipulated through satire, pseudonymity, or collective trolling, making straightforward literalism a liability for MLLM-based interpretation. A model that cannot distinguish between a sincere autobiographical post and an ironic meme risks constructing distorted identity trajectories. Robustness can be improved by training on adversarially augmented data and by incorporating uncertainty modeling into the architecture, but there is an irreducible limit to what can be automatically disambiguated without contextual knowledge of the discursive community in which a post is embedded. This points to the indispensability of qualitative human expertise, not as a temporary crutch but as a permanent component of any sociotechnical system that aspires to interpret human self-concept faithfully.

A further trade-off concerns the tension between model scale and environmental sustainability. The largest MLLMs incur substantial energy costs during both training and frequent fine-tuning cycles that a longitudinal identity monitoring infrastructure might demand [22]. Efforts to reduce this footprint through sparse activation, knowledge distillation, and once-for-all training regimes must be balanced against the risk that compressed models lose the representational nuance needed to capture marginal identity expressions. A systems-level sustainability strategy might involve scheduling retraining during periods of low-carbon grid intensity, investing in shared regional compute nodes that serve multiple ethical identity research projects, and developing frugal, context-adaptive inference paths that activate full multimodal capacity only when a post's complexity warrants it.

7. Policy Implications and Sustainable Deployment

Policymakers are increasingly attentive to the psychosocial impacts of algorithmic systems, as evidenced by legislative initiatives targeting addictive design, transparency obligations for recommender systems, and algorithmic impact assessments. An MLLM-based infrastructure for self-concept research could inform these regulatory efforts by providing evidence on how specific platform features, such as quantified social metrics or autoplay mechanisms, correlate with identity volatility and well-being indicators over time. However, instrumentalizing such research for platform regulation raises the specter of dual use: the same models capable of illuminating harmful dynamics could be deployed by platforms to further optimize engagement by tuning the presentation of self-relevant content in ever more psychologically targeted ways.

Sustainable deployment therefore demands a clear separation between the research layer, governed by independent ethics boards and data trustees, and the operational optimization layer that platforms control. Interoperability standards that permit researchers to query platform-side MLLM embeddings through privacy-preserving APIs, without exposing raw user content, could strike a productive balance between epistemic access and data protection. Such an approach aligns with the broader movement toward platform-to-researcher data sharing that has gained traction in misinformation and public health research, though it would require extending existing frameworks to accommodate multimodal data types and the specific vulnerabilities associated with identity research [23].

International coordination will be necessary to prevent regulatory arbitrage, where identity research migrates to jurisdictions with the weakest protections. An international accord on

digital identity research ethics, analogous to the Oviedo Convention in biomedicine, could establish baseline prohibitions against non-consensual profiling and mandate model explainability for any system that makes inferences about a person's self-concept. Within such a framework, certification schemes for MLLM-based identity research systems could provide assurance that architectural and governance safeguards meet specified thresholds, while leaving room for contextual adaptation across cultural settings.

8. Conclusion

The use of multimodal large language models to understand human self-concept evolution in digital social environments sits at the intersection of computational systems, psychology, and platform governance. This paper has argued that realizing the promise of such an endeavor requires a thoroughgoing systems orientation, one that treats the architectural, data governance, fairness, robustness, and policy dimensions not as afterthoughts but as constitutive elements of the research infrastructure. We have traced the conceptual lineage from the psychological literature on self-concept through the algorithmic mediation of identity expression to the specific capabilities and limitations of contemporary MLLMs. In doing so, we have emphasized the structural trade-offs that any real-world deployment will inevitably confront, including the tension between representational richness and privacy, between early-fusion integration and auditability, and between model scale and sustainability.

The analytical trajectory developed here suggests that the path forward lies not in chasing ever larger, more opaque models but in designing federated, interpretable, and community-governed systems that embed self-concept research within a rights-respecting envelope. By treating the mirror that MLLMs hold up to digital society as an object of co-design rather than a neutral instrument, researchers can contribute to a public understanding of how platforms shape who we become, while simultaneously empowering individuals and communities to reclaim narrative agency over their digitally mediated lives.

References

1. Turkle, S. (1995). Life on the screen: Identity in the age of the Internet. Simon & Schuster.
2. boyd, d. (2014). It's complicated: The social lives of networked teens. Yale University Press.
3. Goffman, E. (1959). The presentation of self in everyday life. Doubleday.
4. Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. Psychological Review, 98(2), 224–253.
5. Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. Psychological Review, 94(3), 319–340.
6. Oyserman, D., Elmore, K., & Smith, G. (2012). Self, self-concept, and identity. In M. R. Leary & J. P. Tangney (Eds.), Handbook of self and identity (2nd ed., pp. 69–104). Guilford Press.
7. Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin Press.
8. Sunstein, C. R. (2017). #Republic: Divided democracy in the age of social media. Princeton University Press.
9. McAdams, D. P. (2001). The psychology of life stories. Review of General Psychology, 5(2), 100–122.

10. Solanki, D., Hsu, H. M., Zhao, O., Zhang, R., Bi, W., & Kannan, R. (2020, July). The way we think about ourselves. In International Conference on Human-Computer Interaction (pp. 276-285). Cham: Springer International Publishing.

11. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., ... Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35, 23716-23736.

12. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International Conference on Machine Learning (pp. 19730-19742). PMLR.

13. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. Advances in Neural Information Processing Systems, 36.

14. OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.

15. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Petrov, S., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., ... Vinyals, O. (2023). Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.

16. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748-8763). PMLR.

17. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.

18. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623).

19. Zuboff, S. (2019). The age of surveillance capitalism. PublicAffairs.

20. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679.

21. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).

22. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59-68).

23. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645-3650).

24. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P.

(2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

25. Zhou, D. (2026, May). A Code Visualization Graph-Based Method for Vulnerability Severity Assessment Using a Multi-Scale Feature Fusion Network. In 2026 3rd International Conference on Image Processing and Artificial Intelligence (ICIPAI) (pp. 306-309). IEEE.