Who Speaks for Public Data? Rethinking Meaningful Consent in Multimodal LLM Research

XIAO ZHAN*, VRAIN, Universitat Politècnica de València & University of Cambridge, Spain, United Kingdom
GUANGZHI SUN*, University of Cambridge, United Kingdom
JOSE SUCH, INGENIO (CSIC-Universitat Politècnica de València), Spain

In research practice, multimodal large language models are often trained on data treated as "public" (e.g., videos scraped from YouTube). However, these models may memorize and re-expose training data, and can introduce downstream risks that were not anticipated at the time of data upload. Such uses go beyond what individuals could reasonably understand or consent to when sharing content online. In light of these concerns, this paper reflects on data practices in multimodal LLM research and calls for renewed attention to how meaningful consent should be obtained and sustained in this context.

CCS Concepts: • **Security and privacy → Information accountability and usage control**;

Additional Key Words and Phrases: Meaningful consent, Public data reuse, Multimodal large language models, research ethics, privacy, audio-visual data

ACM Reference Format:
Xiao Zhan, Guangzhi Sun, and Jose Such. 2026. Who Speaks for Public Data? Rethinking Meaningful Consent in Multimodal LLM Research. In *CHI 2026 Workshop on Moving Beyond Clicks: Rethinking Consent and User Control in the Age of AI (CHI '26 Workshop), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 5 pages.

1 What Happened To Consent? Have I given consent?

The sources of training data for AI models have long been controversial, as advances in generative AI have exposed the extent to which model training relies on data that was never "meaningfully consented to" (i.e., consent given for a specific purpose, with a reasonable understanding of potential outcomes and the possibility of withdrawal [26]), but treated as permissible solely because it was publicly accessible [13, 14]. Much of the resulting public debate has framed this issue through a commercial lens [e.g., 2, 3, 15, 22], emphasizing questions of copyright infringement, financial compensation for creators, and the responsibilities of corporate actors who deploy large-scale models.

Yet this consent issue is significantly understudied in academic contexts. A common underlying assumption has long been that publicly available data can be used with the owner's consent for training. This assumption has long underpinned research practices involving large-scale datasets and is frequently invoked to justify ethical waivers in data collection and model development. However, with the rapid development of multimodal large language models (LLMs), their emergent capabilities, unpredictable usage, and the dense information in videos are forcing us to re-examine user

*Both authors contributed equally to this research.

Authors' Contact Information: Xiao Zhan, xzhan1@upv.es, VRAIN, Universitat Politècnica de València & University of Cambridge, Valencia, Spain, Cambridge, United Kingdom; Guangzhi Sun, gs534@cam.ac.uk, University of Cambridge, Cambridge, United Kingdom; Jose Such, jose.such@csic.es, INGENIO (CSIC-Universitat Politècnica de València), Valencia, Spain.

Manuscript submitted to ACM

consent. Ultimately, **as multimodal LLMs continue to blur the line between public availability and exploitation, addressing this "consent gap" will be critical to ensuring the responsible and sustainable evolution of the field.**

2 Rethinking Consent Through the Lens of LLM Security&Privacy Research

Findings from security and privacy research demonstrate that LLMs are capable of memorizing and reproducing elements of their training data, reorganizing information across contexts [8, 9, 33, 35], and re-exposing data through interactive model outputs [11, 23, 28, 30]. These capabilities transform public data into persistent, generative representations that enable reuse and redistribution beyond their original disclosure context [5].

Unlike text, audiovisual data encodes rich contextual and perceptual information [6, 24] that makes individuals more readily identifiable, even when they are not the intended subjects of the data [4]. Images, videos, and audio recordings often capture bystanders incidentally [19, 34], embedding information about individuals who neither created the content nor consented to its downstream use in model training.

Incorporating such data into multimodal LLMs introduces significant new risks. Cross-modal alignment enables models to associate visual, auditory, and textual signals, supporting emergent abilities, such as cross-modal reasoning and holistic inference with knowledge from all different resources [29, 31, 32]. These capabilities can unexpectedly enable such models or malicious users to expose sensitive attributes, reconstruct contexts of appearance, or even hallucinate misinformation that is absent from the original data [e.g., 10, 16–18, 20, 27].

While meaningful consent, which requires clarity about data use and a reasonable understanding of potential outcomes, is difficult to satisfy in multimodal LLM training, the transformative reuse of training data fundamentally undermines consent obtained at the point of original disclosure. Such downstream uses and risks were neither clearly articulated nor reasonably foreseeable when individuals shared content online.

3 Consent Gaps in Multimodal Research and What Comes Next?

We conducted a selective survey across multimodal LLM research papers in recent conferences, and found two major patterns of consent gaps. **The first gap lies in data reuse.** Many multimodal LLMs reuse datasets that contain videos originally collected through crowdsourcing for a clearly defined purpose, such as speech recognition or emotion classification. In these scenarios, participants agreed to make their data public for specific research use only. However, these datasets are often reused (sometimes unconsciously) to train LLMs for different purposes, or circulated across different research projects without re-collecting or re-examining the associated consent conditions. **Another gap arises from shifting platform norms.** For example, YouTube's 2024 policy update makes video uploads opt out of third-party model training by default [1, 25]. Yet videos that were previously scraped continue to exist as precompiled datasets hosted on platforms such as GitHub and Hugging Face. While consent is being reconsidered on the commercial side, its implications have yet to meaningfully carry over into research contexts, leaving many of these issues unresolved.

This further substantiates another challenge of meaningful consent: once audio-visual data has propagated across benchmarks and been absorbed into trained models, there is often no realistic way for creators or data subjects to withdraw it. In practice, this means that consent – if it was ever given – becomes effectively irreversible. This situation suggests that current research practices place greater weight on public availability than on consent itself, even as this assumption becomes increasingly difficult to defend in the context of multimodal LLMs.

Therefore, in this position paper, we outline several directions that warrant sustained research attention to stimulate discussion at the workshop and inform future work.

- What constitutes meaningful consent in multimodal LLM research, and how should it be operationalized in practice?
  - How should researchers inform creators or data subjects about potential uses, risks, and impacts in ways that enable genuinely informed decisions – especially when many audio-visual creators (e.g., YouTubers) view data use primarily through the lens of compensation [12] rather than privacy?
- Many papers claim exemption from formal research ethics committees (REC) approval in their ethics statements because the data is publicly available. We argue that public availability alone should not justify exemption and that such research ought to undergo ethics review. This raises a further question: when review is required, what responsibilities should REC assume in auditing claims that meaningful consent has been obtained for the data used in research?
  - Without enforceable intervention from existing ethics review bodies, the academic community is unlikely to take this issue seriously, and the problem is likely to worsen in both scale (more datasets, more models) and in that these practices are becoming standard and unquestioned. However, key questions remain unresolved, including how standards for assessing meaningful consent should be specified and what criteria could be used to determine whether such standards are met. Research ethics committees might consider whether consent evaluation could be supported or delegated to automated or AI-based tools. However, potential newly introduced risks should also be carefully considered.
- What technical solutions might help bridge the gap between the limitations of traditional consent and the need for meaningful individual control over personal data?
  - To what extent can emerging techniques such as *machine unlearning* or targeted data removal [7, 21] support post-hoc withdrawal of consent in multimodal LLMs? How should the limitations and failure modes of these technical approaches be communicated, so that they are not treated as substitutes for consent but as partial safeguards within broader ethical frameworks?
- How should research practices account for individuals who appear incidentally in data as bystanders, but have neither created the content nor consented to its reuse? Recent work has begun to address bystander privacy in specific modalities, e.g. in audio LLMs [34]. However, broader questions remain about how to allocate control over bystander data across different modalities and contexts:
  - When consent cannot reasonably be obtained, how should control over the use and inference of bystanders' data be allocated, and who should be responsible for exercising that control in multimodal LLM research?

4 Author Biography

Xiao Zhan is a postdoctoral researcher at VRAIN, Universitat Politècnica de València, and a visiting research scholar at the University of Cambridge. Her research interests lie at the intersection of AI and cybersecurity, with a particular focus on human-centered privacy and security, as well as LLM–powered conversational AI.

Guangzhi Sun is a junior research fellow at Trinity College, University of Cambridge. His research focuses on controllable and reliable multimodal conversational AI with LLMs, including multimodal contextual knowledge integration, hallucination reduction, and contextualised AI safety.

Jose Such is a Research Professor at the Spanish National Research Council (CSIC). His research lies at the intersection of artificial intelligence, cybersecurity, and human–computer interaction, adopting a human-centered approach to understanding online harms and to designing secure, privacy-respecting systems.

References

[1] 2024. YouTube policies. Retrieved Feb 01, 2026 from https://support.google.com/youtube/answer/15509945?hl=en-GB

[2] 2026. AI training and copyright – the debate continues. Retrieved Feb 01, 2026 from https://bto.co.uk/blog/ai-training-and-copyright-the-debate-continues/

[3] Katie Baker. 2024. Is Google Stealing YouTube Videos to Train AI? Retrieved Janurary 30, 2026 from https://em360tech.com/tech-articles/google-stealing-youtube-videos-train-ai

[4] Solon Barocas and Helen Nissenbaum. 2014. *Big Data's End Run around Anonymity and Consent.* Cambridge University Press, 44–75.

[5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, 610–623. doi:10.1145/3442188.3445922

[6] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1536–1546.

[7] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*. IEEE, 141–159.

[8] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying Memorization Across Neural Language Models. In *Proc. ICLR*.

[9] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*. 2633–2650.

[10] Tiejin Chen, Pingzhi Li, Kaixiong Zhou, Tianlong Chen, and Hua Wei. 2025. Unveiling privacy risks in multi-modal large language models: Task-specific vulnerabilities and mitigation challenges. In *Findings of the Association for Computational Linguistics: ACL 2025*. 4573–4586.

[11] Yang Chen, Ethan Mendes, Sauvik Das, Wei Xu, and Alan Ritter. 2023. Can language models be instructed to protect personal information? *arXiv preprint arXiv:2310.02224* (2023).

[12] DACS. 2025. Our joint statement calling for transparency, fairness and respect for creators' rights in the age of AI. Retrieved Feb 01, 2026 from https://www.dacs.org.uk/news-events/joint-statement-on-creators-rights-in-age-of-ai?

[13] Anna Desmarais. 2024. Elon Musk's X quietly changes default settings to allow it to train AI model Grok with your posts. Retrieved Janurary 30, 2026 from https://www.euronews.com/next/2024/07/27/elon-musks-x-quietly-changes-default-settings-to-allow-it-to-train-ai-model-grok-with-your

[14] Annie Gilbertson and Alex Reisner. 2024. Apple, Nvidia, Anthropic Used Thousands of Swiped YouTube Videos to Train AI. Retrieved Janurary 30, 2026 from https://www.proofnews.org/apple-nvidia-anthropic-used-thousands-of-swiped-youtube-videos-to-train-ai/#:~:text=Apple%2C%20Nvidia%2C%20Anthropic%20Used%20Thousands,YouTube%20Videos%20to%20Train%20AI

[15] Benjamin Hiorns. 2025. AI's Great (Training) Robbery: How Small Creators Can Fight Back. Retrieved Janurary 30, 2026 from https://creativepool.com/magazine/features/ais-great-training-robbery-how-small-creators-can-fight-back.33406

[16] Yuke Hu, Zheng Li, Zhihao Liu, Yang Zhang, Zhan Qin, Kui Ren, and Chun Chen. 2025. Membership inference attacks against vision-language models. In *34th USENIX security symposium (USENIX Security 25)*. 17 pages.

[17] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. doi:10.1145/3703155

[18] Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. 2024. Membership inference attacks against large vision-language models. *Advances in Neural Information Processing Systems* 37 (2024), 98645–98674.

[19] Zhaojiang Lin, Yong Xu, Kai Sun, Jing Zheng, Yin Huang, Surya Teja Appini, Krish Narang, Renjie Tao, Ishan Kapil Jain, Siddhant Arora, et al. 2026. WearVox: An Egocentric Multichannel Voice Assistant Benchmark for Wearables. In *Proc. ICLR*.

[20] Feiran Liu, Yuzhe Zhang, Xinyi Huang, Yinan Peng, Xinfeng Li, Lixu Wang, Yutong Shen, Ranjie Duan, Simeng Qin, Xiaojun Jia, et al. 2025. The eye of sherlock holmes: Uncovering user private attribute profiling via vision-language model agentic framework. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 4875–4883.

[21] Tyler Lizzo and Larry Heck. 2025. UNLEARN efficient removal of knowledge in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*. 7257–7268.

[22] Robin Mansell. 2025. A capitalist contest: the AI industry v. the creative industries. *Journal of the British Academy* 13, 3 (2025).

[23] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2025. Scalable extraction of training data from (production) language models. In *Proc. ICLR*.

[24] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021).

[25] Sarah Perez. 2024. YouTube will now let creators opt in to third-party AI training. Retrieved Feb 01, 2026 from https://techcrunch.com/2024/12/16/youtube-will-let-creators-opt-out-into-third-party-ai-training/

[26] Giada Pistilli and Bruna Trevelin. 2025. Can AI be Consentful? *arXiv preprint arXiv:2507.01051* (2025).

[27] Jakub Proboszcz, Paweł Kochanski, Karol Korszun, Donato Crisostomi, Giorgio Strano, Emanuele Rodolà, Kamil Deja, and Jan Dubinski. 2025. Membership and Dataset Inference Attacks on Large Audio Generative Models. In *Proceedings of the NeurIPS 2025 Workshop on Generative and*

Protective AI for Content Creation.

[28] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Beyond memorization: Violating privacy via inference with large language models. In *Proc. ICLR*.

[29] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. 2024. video-salmonn: Speech-enhanced audio-visual large language models. In *Proc. ICML*.

[30] Shang Wang, Tianqing Zhu, Bo Liu, Ming Ding, Dayong Ye, Wanlei Zhou, and Philip Yu. 2025. Unique Security and Privacy Threats of Large Language Models: A Comprehensive Survey. *ACM Comput. Surv.* 58, 4, Article 83 (Oct. 2025), 36 pages. doi:10.1145/3764113

[31] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-Omni Technical Report. *arXiv:2503.20215* (2025).

[32] Yudong Yang, Xuezhen Zhang, Zhifeng Han, Siyin Wang, Jimin Zhuang, Zengrui Jin, Jing Shao, Guangzhi Sun, and Chao Zhang. 2025. Speech-Audio Compositional Attacks on Multimodal LLMs and Their Mitigation with SALMONN-Guard. *arXiv:2511.10222* (2025).

[33] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing* 4, 2 (2024), 100211.

[34] Xiao Zhan, Guangzhi Sun, Jose Such, and Phil Woodland. 2025. Protecting Bystander Privacy via Selective Hearing in Audio LLMs. *arXiv preprint arXiv:2512.06380* (2025).

[35] Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. Counterfactual memorization in neural language models. *Advances in Neural Information Processing Systems* 36 (2023), 39321–39362.