Bridging the Cognitive Gap: Cross-Modality Perception in AI-Generated Content

Can Liu
School of Creative Media
City University of Hong Kong
Hong Kong, China
canliu@cityu.edu.hk

Wengxi Li
School of Creative Media
City University of Hong Kong
Hong Kong, China
wengxili@cityu.edu.hk

Abstract

This workshop paper explores the challenges and opportunities in cross-modality perception for AI-generated content, with a focus on bridging the cognitive gap between textual and graphical representations. We propose a roadmap of fundamental research questions aimed at understanding how humans perceive and interact with content as it transitions across modalities. Drawing from empirical studies of recent tools, we identify key challenges in semantic space representation, dynamic visualization, and cognitive alignment. This paper contributes to the discourse on human-AI interaction by proposing research directions to enhance the human experience with generative content, emphasizing reduced cognitive load, semantic fidelity, and cross-modality coherence.

CCS Concepts

• **Human-centered computing** → *Empirical studies in HCI*.

Keywords

Semantic Space, Cross-modality, AI-generated Content, Cognitive Process

ACM Reference Format:
Can Liu and Wengxi Li. 2025. Bridging the Cognitive Gap: Cross-Modality Perception in AI-Generated Content. In *Proceedings of CHI 2025 Workshop - Tools for Thought*. ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The rapid advancement of AI-generated content has transformed how humans create, interact with, and perceive information. From text and images to 3D models and interactive media, generative AI systems are increasingly capable of producing rich, multimodal outputs. However, as these systems grow more sophisticated, a critical challenge emerges: the cognitive gap between how humans conceptualize content and how machines represent it across different modalities.

The motivation for this work stems from our observation in several recent and ongoing works, where we noticed misalignment in users' cognitive processes across content representations. For instance, we introduced an AI-supported dictation tool Rambler [5], which allows users to manipulate pieces of text at a semantic level while AI performs the verbatim edits. Although the interface showed promises, it still relied on textual representations, which can be cumbersome to parse during dictation. Similarly, AI systems that generate graphical content from textual prompts often produce outputs that feel disconnected from the user's intent, creating friction in creative workflows. These challenges highlight a broader need for tools and models that bridge the gap between modalities, enabling users to interact with content in ways that feel natural, intuitive, and aligned with their mental models.

This workshop paper discusses the challenges and opportunities in cross-modality perception for AI-generated content. The rest of this paper introduces the background and motivation of this discussion from our recent works, from which we propose a list of fundamental research questions in cross-modality perception. We also discuss the broader implications of addressing these challenges.

2 Rambler: A Case Study in Semantic-Level Text Manipulation

Rambler, a dictation tool, allows users to dictate content and manipulate it at a semantic level rather than verbatim. By helping users merge and split spoken content into coherent pieces, Rambler enables users to perform coarse-grained manipulations with AI support. This approach has shown promise in improving text input and manipulation efficiency. However, a critical limitation remains: the semantic space is still represented textually, which makes it hard to parse, especially during dictation. We provided both keyword highlights and LLM-generated summaries in the tool, but it was unclear how effective they were.

To further investigate this, we conducted an eye tracking study comparing different speech-to-text interfaces designed to display spoken content with varying representations: (1) raw transcripts, (2) cleaned transcripts (basic error and punctuation correction), (3) transcripts with real-time keyword highlighting (two strategies), (4) summaries with AI-rephrased wording. Using eye-tracking metrics to measure reading difficulty, we found that *AI-generated summaries with altered wording* unexpectedly reduced reading difficulty most effectively, despite introducing unfamiliar phrasing. We termed this phenomenon **desirable unfamiliarity**.

This finding underscores the importance of designing tools that **prioritize semantic fidelity over verbatim accuracy**. Users' mental models of their dictated content are inherently abstract [3]

CHI 2025 Workshop - Tools for Thought, Yokohama, Japan

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

and tied to intent rather than specific word choices. Thus, intermediate checkpoints in tools like Rambler shall ideally visualize semantic structures, not just textual fragments, to align with users' cognitive processes.

3 Challenges in Representing Semantic Space

The challenge in representing the semantic space is a broad topic. The specific challenges we have encountered in our work so far include the following three aspects.

3.1 Semantic Zoom and Continuity

One of the challenges in representing semantic space is enabling users to navigate across different levels of semantic zoom. Existing solutions allow users to drill down into detailed text or zoom out for an overview, but the reading experience navigating these levels is often disconnected. Hayatapur et al. explored how using semantic zoom to visualize program behaviors at different levels of scope and abstraction [2]. Suh et al. leveraged semantic zoom to enable users to easily overview large amounts of text responses generated by LLMs across three different levels of abstraction [8]. Their later work took one step further and push it to five abstraction levels [7]. However, users must "start over" when transitioning between levels, disrupting their cognitive flow. A key research question this leads to is: How can we create a continuous perception across semantic zoom levels?

3.2 Visualizing Evolving Semantic Space

Building on the "desirable unfamiliarity" phenomenon mentioned above, we argue that dynamic visualizations must balance abstraction with recognizability. For example, in our study, summaries succeeded because they retained semantic intent while smoothing verbatim wording. Similarly, visual representations of evolving semantic spaces could use *adaptive abstraction* to preserve core themes while simplifying surface details. One good example of existing work for visualizing speech uses Topic-Space views to track the movement of speakers across the thematic landscape in multi-party conversation [1]. They implemented an animated view that shows the speakers' talk topics and the transitions over time. This visual sedimentation metaphor enables users to track subtle changes in the flow of a conversation over time while keeping an overview of all past speaker turns. More future interventions are needed to advance this field.

3.3 Cross-Modality Representation

To bridge the gap between textual and graphical representations, we explore how to map semantic dimensions (e.g., themes, relationships, and hierarchies) to visual dimensions (e.g., color, spatial position, motion, and direction). For instance, color could represent thematic categories, while spatial position could indicate hierarchical relationships.

If we move beyond basic visual dimensions, icons, sketches or various graphs could also be considered. Several works have explored augmenting speech with visual content in real-time settings. Liao et al. demonstrated a speech-driven system that can display relevant images and icons during live presentations based on spoken keywords [4]. Similarly, Liu et al. developed an approach to augment video conferencing by automatically suggesting contextual visuals based on real-time speech recognition [6]. These works show the potential of speech-triggered visual augmentation while highlighting key challenges in semantic matching and timing. In a preliminary study, we tested icon-based representations of keywords from STT-transcribed external speech, and found that users struggled to map icons to their semantically matched text in real-time. This suggests that while textual representation can leverage abstraction effectively, graphical representations require semantic anchoring. We find this cognitive gap an interesting problem space for better understanding and future solutions.

4 Emerging Research Questions

This paper identifies several fundamental research questions to guide future work:

- Semantic Continuity: How can we create a continuous perception across semantic zoom levels?
- Dynamic Visualization: What visual representations best capture the evolving semantic space of spoken content?
- Customization: How can users define mappings between semantic and visual dimensions to reflect their unique perspectives?
- Cognitive Impact: How do cross-modality AI models affect perception and cognitive load?
- Abstraction vs. Familiarity: How can tools balance semantic abstraction with familiarity to support cognition?

5 Implications and Impact

Addressing the challenges outlined in this paper has the following potential implications.

Tools that bridge textual and graphical modalities could enable users to interact with content at the level of **ideas** rather than words or pixels. For example, a designer dictating a storyboard could manipulate visual metaphors directly, with AI translating their intent into both text and imagery. A teacher dictating a lesson plan might use a semantic map to reorganize ideas visually, with AI handling the syntactic details. This shifts the focus from technical skill (e.g., prompt engineering) to intentional expression.

By prioritizing semantic fidelity over verbatim accuracy—as demonstrated in our "desirable unfamiliarity", study—tools could reduce the mental effort required to parse dense text or ambiguous visuals. This is particularly critical for users with cognitive disabilities or those working in high-pressure environments (e.g., emergency responders using voice-to-visual tools).

6 Conclusion

This workshop paper proposes a roadmap for addressing the challenges of cross-modality perception in AI-generated content. By exploring the semantic space, dynamic visualization, and cognitive alignment, we aim to enhance the human experience with generative tools. The research questions outlined here provide a starting point for discussing future work with other attendees, with the ultimate goal of bridging the cognitive gap between textual, graphical, and even other emerging modalities like 3D.

References

[1] Mennatallah El-Assady, Valentin Gold, Carmela Acevedo, Christopher Collins, and Daniel Keim. 2016. ConToVi: Multi-Party Conversation Exploration using Topic-Space Views. *Comput. Graph. Forum* 35, 3 (June 2016), 431–440.

[2] Devamardeep Hayatpur, Daniel Wigdor, and Haijun Xia. 2023. CrossCode: Multi-level Visualization of Program Execution. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 593, 13 pages. doi:10.1145/3544548.3581390

[3] David Kirsh. 2010. Thinking with External Representations. *AI and Society* 25, 4 (2010), 441–454. doi:10.1007/s00146-010-0272-8

[4] Jian Liao, Adnan Karim, Shivesh Singh Jadon, Rubaiat Habib Kazi, and Ryo Suzuki. 2022. RealityTalk: Real-Time Speech-Driven Augmented Presentation for AR Live Storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 17, 12 pages. doi:10.1145/3526113.3545702

[5] Susan Lin, Jeremy Warner, J.D. Zamfirescu-Pereira, Matthew G Lee, Sauhard Jain, Shanqing Cai, Piyawat Lertvittayakumjorn, Michael Xuelin Huang, Shumin Zhai, Bjoern Hartmann, and Can Liu. 2024. Rambler: Supporting Writing With Speech via LLM-Assisted Gist Manipulation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1043, 19 pages. doi:10.1145/3613904.3642217

[6] Xingyu "Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, and Ruofei Du. 2023. Visual Captions: Augmenting Verbal Communication with On-the-fly Visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 108, 20 pages. doi:10.1145/3544548.3581566

[7] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 644, 26 pages. doi:10.1145/3613904.3642400

[8] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 1, 18 pages. doi:10.1145/3586183.3606756

Received 20 February 2025; revised 10 April 2025