

Before the Words Arrive: Trans-Modal Narrative Agency in Multilingual Children's AI Co-Creation

A Teacher-Inquiry Study

Xiao (Yuxiao) Yang

Abstract

Research on second language acquisition has long been criticized for casting learners as deficient communicators measured against an idealized native speaker (Firth & Wagner, 1997). Within that frame, the “silent period” of emergent bilinguals (Krashen, 1982) can look like a stage of waiting. This teacher-inquiry study looks at that period from another angle. I analyze sessions in which three trilingual children co-created stories with generative AI in my online teaching practice. The analysis draws on translanguaging theory (García & Kleifgen, 2020; García & Wei, 2014), sociocultural theory (Vygotsky, 1978), recent design research on child and AI co-creation (Fan et al., 2025), and recent work on generative AI for multilingual learners (Jeon et al., 2025). In these sessions, AI-generated images often came before the children's English and seemed to open a way into it. I propose the construct of *trans-modal narrative agency* to describe how the children began, shaped, and revised story meaning across images and words ahead of their second language fluency. I treat this agency as mediated, visible within the interaction of child, teacher, and tool. With three children, the study cannot claim generality. It offers a close account of what became possible in one practice and the questions that account raises.

1. Introduction

When six-year-old Heng, a Mandarin-speaking child living in Spain, looked at an AI-generated image of a dark cave, he did not wait for English he did not yet have. “*They are in a scary tunnel! The bats are laughing!*” The laughing bats were not in the story's text. Heng read them into the image. He brought the scene to life through an act of narrative imagination that his English could not yet have carried on its own.

Moments like this press on a view that Firth and Wagner (1997) criticized three decades ago, in which the second language learner appears mainly as someone who lacks the target language. For emergent bilinguals, the distance between what children can imagine and what they can say in a new language is often wide. Text-centered classrooms tend to read that distance as a deficit. The “silent period” becomes a stretch of time to be tolerated until productive speech appears (Krashen, 1982). This study asks what happens when a generative tool enters that distance. The tool does not supply the missing words. It offers a shared visual and narrative

surface where children's meaning can take shape alongside their second language, and at times ahead of it.

The question matters now. Multimodal and multilingual generative models have arrived at a moment when the field is looking for ways to widen access for linguistically diverse learners. These same models produce statistically probable output. They tend to reproduce dominant forms and to under-represent minoritized voices (Jeon et al., 2025). Any claim that generative AI serves multilingual children has to be held together with the risk that it flattens the differences it seems to welcome. This inquiry works inside that tension.

I approach the question from two positions. I am a trilingual online educator in Mandarin, English, and Spanish. I am also a co-developer of *StoryPainter*, a drawing-to-story application that responds to a child's own drawing instead of a typed prompt. The study is a teacher inquiry in the practitioner research tradition (Cochran-Smith & Lytle, 2009). Its site is my own practice, and its data are my own sessions. Across two rounds of one-on-one sessions with three trilingual children, I followed one guiding question: *When multilingual children co-create stories with generative AI, what forms of narrative agency appear, and what in the co-creative setting makes them possible?*

I argue that these children showed what I call **trans-modal narrative agency**. By this I mean the capacity to begin, negotiate, and revise story meaning across visual and verbal modes ahead of second language fluency. I treat it as a mediated agency. It belongs to the child, and it became visible within the interaction of child, teacher, and tool. In these sessions, the tool mattered less for the content it produced and more for the interpretive space it opened among the three.

2. Situating the Inquiry

Three strands of scholarship meet in this study. The first is design research on child and AI co-creation. The second is work on generative AI for multilingual learners. The third is the longer tradition of research on how emergent bilinguals make meaning across modes. We know a good deal about how children create with AI, and we know a good deal about how bilinguals make meaning across modes. We know much less about what happens when an emergent bilingual child creates alongside a generative tool in real time.

2.1 Child and AI Co-Creation of Stories

Design research has begun to examine how children create stories with generative systems. StoryDrawer paired children aged six to ten with an AI partner that drew alongside them as they told a story (Zhang et al., 2022). More recently, Fan et al. (2025) developed StoryPrompt, a system in which children co-create stories and comics with generated text and images. In a mixed-methods evaluation with 40 children in Grades 2 through 6, StoryPrompt improved the creativity and richness of children's stories compared with a storyboard method. The authors also observed that children made strategic use of AI-generated words and images as they planned.

These studies do not center children who are composing in a language they are still acquiring. That case is where this inquiry sits.

2.2 Generative AI and Multilingual Learners

A second body of work weighs the promise of generative AI for linguistically diverse learners against its risks. Jeon et al. (2025) describe these systems as flexible receivers of input across languages and modes. They also argue that the systems return standardized, statistically probable output shaped by their training data. From a translanguaging perspective, they name three dilemmas in this position: creativity and standardization, inclusion and fixity, and meaning and form. Sun (2026) examines how generative AI shapes writing in English as an additional language and frames translingual practice as a way to resist homogenization.

The concern is also empirical. Henry et al. (2025) used ChatGPT-4o to revise five published bilingual children's books under several prompting strategies. The model standardized translanguaging, standardized structure, and homogenized culture. The authors warn that such tools may reinforce deficit views of multilingual writers. For a classroom study, the lesson is that researchers need to look at what a tool removes as closely as they look at what it enables. In this study I treat generative AI as an artifact that needs ongoing pedagogical mediation.

2.3 Emergent Bilinguals and Multimodal Meaning-Making

Long before generative AI, research on emergent bilinguals established that meaning-making is multimodal. Image, gesture, and more than one language work together as resources (García & Kleifgen, 2020). This work invites a second look at the "silent period" (Krashen, 1982). What appears to be silence in the dominant language may be active

meaning-making in other modes. This inquiry extends that line of work into a generative, co-creative setting. I ask whether such a setting can make non-verbal meaning-making visible as authorship.

3. Theoretical Framework

Three frameworks guide the analysis. I use each one to read the data, and I also note where the data seem to reach past it.

3.1 Translanguaging and the Meaning-Making Repertoire

Translanguaging views a bilingual person as drawing on one integrated repertoire for making meaning (García & Wei, 2014). García and Kleifgen (2020) extend that repertoire beyond speech to include image and gesture. This view changes what counts as success. When Heng read laughing bats into an AI image, he was succeeding at something. He was using visual and linguistic resources together to make meaning.

Translanguaging theory was developed around communication among people. It says less about what happens when a system trained on dominant language forms enters the exchange. Jeon et al. (2025) raise this question in theory, and this study offers a small empirical look at it.

This raises a fair question about why a new term is needed. Trans-modal narrative agency does not replace multimodal translanguaging. It sits inside that framework and names something narrower in three ways. First, it concerns narrative authorship, which is only one of the things a translanguaging repertoire does. Second, it attends to direction and sequence. The prefix follows Hawkins's (2018) account of transmodalities, which describes meaning as moving across semiotic resources. Hawkins works at the scale of transnational encounters among young people. My focus is narrower. In these sessions meaning often moved from an image into English, and the prefix marks that movement over time. Third, the images in these sessions had another author. A generative model drew them. Part of the children's agency lay in how they responded to an image that someone else had made. They could accept it, or they could contest it. The translanguaging literature has not yet had to address that situation.

3.2 Sociocultural Theory and a Distributed Zone of Proximal Development

Vygotsky's (1978) account of socially mediated learning provides the second frame. The zone of proximal development usually places a more capable person, such as a teacher or a peer,

in the role of support. In these sessions the support seemed to be shared. The AI offered images and story lines beyond what the child could produce alone. I offered confirmation, questions, and invitations to say more. The zone of proximal development was built by a teacher and a tool together. This arrangement was productive, and it was also fragile. For this reason I describe the agency in this study as mediated. The children brought the capacity, and the teacher and the tool created the conditions in which it could be seen.

3.3 Multimedia Learning, Re-Anchored

Mayer's (2009) theory of multimedia learning holds that people learn more deeply from words and pictures together than from words alone. Most discussions of the theory treat the picture as support for a verbal message. For these emergent bilinguals, the order often seemed to run the other way. The image did not illustrate a sentence the child already had in English. It came first, and the English grew from it. I do not claim that this overturns Mayer's framework. I suggest that for emergent bilinguals, the framework may need to be re-anchored on a visual point of entry. This re-anchoring is the theoretical center of the study.

4. The Inquiry

4.1 Stance, Positionality, and Ethics

This practitioner inquiry (Cochran-Smith & Lytle, 2009) took place in my online teaching practice during the second year of my master's program. I held three roles at once. I was the teacher, a co-developer of the tool, and the researcher. I name this openly. It gave me close access to the children's learning, and it also created a real risk that I would see what I hoped to see.

I took three steps to reduce that risk. First, I coded disconfirming moments, such as times when the tool misread a drawing or a child lost interest. Second, I built sessions around the children's own drawings. The child set the direction of the story, which gave the child less reason to guess what I wanted to hear. Third, I limit my claims to what a small practitioner study can support.

Caregivers gave informed consent for all sessions. All names are pseudonyms, and identifying details have been altered.

4.2 Site and Tool

Data come from a series of 50-minute, one-on-one online sessions held between September and November 2025. The main tool was *StoryPainter*. Most text-to-image tools require a typed prompt. StoryPainter takes the child's drawing as its main input and returns an interpreted image, a short narrative, and a read-aloud. A few sessions also used a common text-to-story tool, and the analysis below focuses on the StoryPainter sessions.

The difference between the two designs shaped my thinking before the analysis began. A text-driven tool gives the first creative move to whoever can type a fluent prompt. With young emergent bilinguals, that person is often the teacher. A drawing-driven tool is designed to give the first move to the child. I offer this as a design rationale and not as a finding, since this study does not compare the two tools systematically.

4.3 Participants

Three trilingual children from my teaching practice took part. Each lives in a different transnational and linguistic arrangement.

- **Ashley** (age 12) was born in China, spent part of her early childhood in Mexico, and now lives in the United States. She moves among Mandarin, Spanish, and English.
- **Elena** (age 12) speaks Russian at home and has grown up in Germany with German. She uses both at a level typical for her age, and she is learning English as an additional language.
- **Heng** (age 6) speaks Mandarin at home and has grown up in Spain with Spanish. He uses both at a level typical for his age. He was the youngest participant and the least fluent in English.

4.4 Data and Analysis

The primary data are video recordings of the sessions, about five hours in total. The recordings capture the children's speech, gestures, facial expressions, and actions on the interface, and I transcribed them in full. Secondary data include the children's digital story artifacts and my reflective memos on my own teaching moves.

Analysis followed an iterative process of thematic coding (Braun & Clarke, 2006). I began with open coding that stayed close to the children's multimodal actions. I then grouped recurring codes into a working codebook and refined it over several passes as categories merged or split. I compared across the three children to find shared patterns and meaningful differences.

When an exchange was ambiguous, I checked it against the story artifacts. In a few cases I asked parents to confirm what a child had meant. This is a limited form of member checking suited to practitioner inquiry.

One boundary of that analysis belongs here. My codebook followed movement between image and English, which was the focus of these sessions and of my teaching. I did not code systematically for the children's home languages, and the excerpts in Section 5 present their English utterances. What follows is an account of one path through a translanguaging repertoire and not a full picture of that repertoire.

Following Dyson and Genishi (2005), I make no claim of statistical generalization from three cases. I aim instead for analytic generalization. The goal is to describe the mechanisms and conditions under which trans-modal narrative agency appeared in these sessions.

5. Findings

I organize the findings as four levels of authorial control. The levels are ordered by how much control the child took over the story. They are not ordered by time, and I do not claim they are developmental stages. Heng, the youngest child, showed all four. Across the levels, generative AI seemed to work less as a content generator and more as a semiotic partner.

- **Level 1. Decoding (5.1).** The child reads meaning into an AI image and produces a first utterance.
- **Level 2. Naming (5.2).** The child finds and tests new vocabulary anchored to the image.
- **Level 3. Interpreting (5.3).** The child reasons about a character's emotions and state of mind.
- **Level 4. Authoring (5.4).** The child revises or overturns the story the AI offers and claims it as their own.

This agency had two sides. The tool opened possibilities beyond what the children could produce alone in a second language. Its statistical design also pulled content toward generic and culturally narrow defaults. At the higher levels, the children's clearest acts of authorship came when they pushed back against those defaults.

5.1 Level 1. Decoding: The Image as First Utterance

AI-generated images served as prompts for building a story. The children read details in the images and turned them into narrative claims, often in directions I had not expected. In one session with Heng, the tool produced an image of a dark cave.

Teacher: *What do you think is happening in this picture?*

Heng: *They are in a scary tunnel! The bats are laughing!*

The laughing bats did not appear in the text. Heng inferred them from the image. His reading carried a literary move, personification, that his English syntax could not yet have produced on its own. In Mayer's (2009) terms, the image did not support an English sentence Heng already had. It came before the sentence and helped make it possible. Through a translanguaging lens (García & Kleifgen, 2020), Heng was thinking in the visual part of an integrated repertoire and expressing meaning that his English could not yet carry.

A similar moment came with Ashley. Shown an AI image of a storm gathering over a village, she paused. Then she offered a causal reading the text had not stated.

Ashley: *The people are running because the sky is angry. Something bad will happen.*

Teacher: *The sky is angry! I love that. What tells you it is angry?*

Ashley: *The clouds are dark and they push together, like a face.*

Ashley's reading is again personification, and again it came from the image. My response mattered here. When I treated these readings as authorship and did not correct them toward the text, both children seemed more willing to take risks in English. For each child, the image offered a first utterance that English alone might have delayed.

5.2 Level 2. Naming: Vocabulary as Situated Discovery

In the co-creative setting, new words tended to arrive as parts of a story the child was building. In one session with Ashley, the tool produced a scene with a princess in a white gown.

Ashley: *The princess is elegant. She has a white dress.*

Teacher: *You used our new word. How do you know she is elegant?*

Ashley: *Because she stands so nice, and the dress is shiny.*

The word *elegant* sat just beyond what Ashley could produce on her own. The read-aloud, the on-screen text, and the image arrived together, and together they seemed to bring the word within her zone of proximal development (Vygotsky, 1978). Ashley did not translate

elegant into Mandarin. She tied it to what she could see, which was the way the princess stood and the shine of the dress. I call this *concept-to-L2 mapping*, a direct link between a new word and a concept anchored in an image. It fits Mayer's (2009) account of how words and pictures together support deeper learning.

Heng showed the same process at a simpler level with the word *glowing*.

Heng: *Glowing means the light comes out, like this, shiny and warm?*

Teacher: *Yes! How did you figure that out?*

Heng: *Because the lantern is yellow and the dark goes away around it.*

Heng built his guess from the contrast between light and dark in the picture. He formed a hypothesis about the word and tested it against the image. Fan et al. (2025) observed children making strategic use of AI-generated words and images as they planned their stories. Heng's move may be an early form of that strategy.

This level came with a caution. The concepts the model offered, such as princesses and glowing lanterns, reflect what the model finds probable, and these images lean toward dominant cultural defaults. The children gained new words, and those words came inside standard images that the children would later need to question.

5.3 Level 3. Interpreting: The AI Character as a Safe Object for Emotion

AI-rendered characters gave the children a third party through which to explore emotion at a safe distance. In one session with Heng, the tool drew a frightened character.

Heng: *He looks scared! His mouth is open and his eyes are shaking!*

Teacher: *Yes. And why do you think he feels that way?*

Heng: *Because he hears the ghost coming, but he doesn't know it's just Kiki!*

Heng attributed a false belief to the character. The character thinks a ghost is coming, and Heng knows it is only Kiki. This is theory of mind reasoning, which means thinking about a mind other than one's own. The clear drawing of fear seemed to lighten the work of naming the emotion, and that left more of Heng's English for explaining its cause. The character worked as a kind of transitional object. It was human enough to invite empathy and artificial enough to examine without personal risk.

Elena used the same distance to reach for a finer emotional distinction.

Elena: *She is not just sad. She is disappointed. She wanted it and it did not happen.*

Teacher: *That is a real difference. What in the picture shows disappointed and not only sad?*

Elena: *She looks down, away from the gift. Sad would cry, but she just turns away.*

Elena's distinction rests on posture and on the direction of the character's gaze. With a teacher's questions around it, the AI character became an occasion for social and emotional talk. The technology did not isolate her. It gave us something to talk about together.

5.4 Level 4. Authoring: Taking the Story Back

The most important pattern was that the children came to treat AI output as a draft to revise. Each child did this in a different way. Elena claimed ownership of a character's development.

Elena: *At first she was sad, but I made her brave.*

Heng set aside an ending the tool proposed and wrote his own.

Teacher: *The story ends with the robot resting under the trees. What do you think happens next?*

Heng: *Maybe the robot wakes up and finds some animal friends! They help him find the way home.*

Ashley went further and questioned the premise of the story.

Ashley: *The story says she is a bad witch, but I don't want that. She is just lonely, so people are scared of her.*

Teacher: *So you are changing who she is. What happens to the story if she is lonely and not bad?*

Ashley: *Then someone can be her friend, and the ending is happy instead of scary.*

Ashley's refusal of the bad witch is the clearest moment in the data. The tool reached for a familiar and statistically probable figure. Jeon et al. (2025) and Henry et al. (2025) warn that defaults of this kind can homogenize the stories of minoritized learners. Ashley did not accept the default. She replaced it with a reading of her own.

This suggests that narrative agency appeared most clearly when a child pushed back against what the tool offered. The systems studied by Zhang et al. (2022) and Fan et al. (2025) were built around a collaborative loop between child and machine. In these sessions, that loop became a place where the children did editorial and even ideological work. They decided what a sentence said, and they also decided who a character was.

6. Discussion: Toward Trans-Modal Narrative Agency

Across the four levels, these emergent bilinguals did authorial work before their English syntax could fully carry it. They personified, inferred causes, reasoned about other minds, and rewrote a story's premise. Text-centered classrooms often do not see these moves.

I define **trans-modal narrative agency** as a learner's capacity to begin, negotiate, and revise narrative meaning across visual and verbal modes ahead of fluency in the target language. The capacity belongs to the child, and in this study it appeared only within mediated interaction. The construct offers another way to read the "silent period" (Krashen, 1982). In these sessions it looked like a period of active authorship, carried out in modes that verbal assessment does not capture.

In this account, the key element is the interpretive space that mediation opens, and the tool's content matters less. Each exchange turned on a question from the teacher, such as "What tells you it is angry?" or "Why do you think he feels that way?" The AI did not replace the teacher. Together, the teacher and the tool formed a distributed zone of proximal development (Vygotsky, 1978). The tool produced objects worth talking about, and the talk did the developmental work. Without mediation, a generative system is more likely to hand a finished story to a passive viewer. StoryPainter's drawing-first design was meant to avoid that kind of consumption.

The construct also carries a warning. Models built on statistical probability tend to favor dominant and standardized forms (Jeon et al., 2025). In practice, that pull can standardize the language and culture of multilingual children's texts (Henry et al., 2025). Trans-modal narrative agency, then, is something children achieve with the help of mediation that invites them to question the tool's defaults. The tool alone does not produce it. Translanguaging (García & Wei, 2014) works here as both theory and practice. Part of a teacher's task is to protect each child's linguistic and cultural specificity against a system that tends to smooth it away.

7. Implications, Limitations, and Next Steps

7.1 *Implications*

For bilingual pedagogy, these sessions suggest widening what counts as authorship. If children can personify, infer, reason about emotion, and revise a story before they can write

fluent English, then curricula that wait for written production may undercount what emergent bilinguals can do. Multimodal co-creative tools, used with a teacher, offer one way to let that competence show.

For tool design, the findings point toward revisability and legibility. Realism mattered less. The most useful outputs were the ones a child could treat as a shared, editable object.

These sessions also raise a question about interface design. In every session a teacher was present to ask a child what in the picture made her say that. Most children who use such tools at home will not have that teacher. It is worth asking whether part of that work can sit in the interface itself, for example in a short question that invites a child to explain an image before the story moves on. The present data cannot answer that question. They can only show that the question is worth asking.

7.2 Limitations

The study has five main limits. First, my three roles as teacher, co-developer, and researcher gave me close access and also a stake in the outcome. The study should be read as situated practitioner knowledge. Second, a teacher was present and asking questions in every session. These data cannot separate what the tool contributed from what my questions contributed, and they cannot show whether the same agency would appear if a child worked with the tool alone. Third, three children over two rounds of sessions cannot show whether trans-modal narrative agency lasts, whether it develops over time, or whether it transfers to growth in the second language. Fourth, I noticed cultural standardization in the tool's output, and this needs systematic study beyond the observations I could offer here. Fifth, my analysis followed the children into English and did not track their home languages. Mandarin, Spanish, Russian, and German were part of what these children brought to the screen. A design built to follow those resources would see more of the repertoire than this one did.

7.3 Next Steps

These limits point to the research I hope to do next. This inquiry gave me a phenomenon and a construct. It did not give me a design that can test them. I want to study trans-modal narrative agency over longer periods and with more varied multilingual communities. I also want to compare sessions with and without a teacher present, so that the contribution of each part of the triad can be seen. I want a design that tracks every language a child brings to the screen and

not English alone. Finally, I want to identify the design conditions under which generative tools reproduce cultural defaults less often, and to describe the teacher moves that keep the co-creative loop open.

The children in this study suggest that authorship can arrive before the words do. The work ahead is to build tools and teaching practices that take that order seriously.

References

- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Cochran-Smith, M., & Lytle, S. L. (2009). *Inquiry as stance: Practitioner research for the next generation*. Teachers College Press.
- Dyson, A. H., & Genishi, C. (2005). *On the case: Approaches to language and literacy research*. Teachers College Press.
- Fan, M., Cui, X., Ma, W., Li, H., Tong, X., Yang, L., & Wang, Y. (2025). From words to wonder: Designing and evaluating an AI-empowered creative storytelling system for elementary children. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). Association for Computing Machinery. <https://doi.org/10.1145/3706598.3713478>
- Firth, A., & Wagner, J. (1997). On discourse, communication, and (some) fundamental concepts in SLA research. *The Modern Language Journal*, 81(3), 285–300. <https://doi.org/10.1111/j.1540-4781.1997.tb05480.x>
- García, O., & Kleifgen, J. A. (2020). Translanguaging and literacies. *Reading Research Quarterly*, 55(4), 553–571. <https://doi.org/10.1002/rrq.286>
- García, O., & Wei, L. (2014). *Translanguaging: Language, bilingualism and education*. Palgrave Macmillan.
- Hawkins, M. R. (2018). Transmodalities and transnational encounters: Fostering critical cosmopolitan relations. *Applied Linguistics*, 39(1), 55–77. <https://doi.org/10.1093/applin/amx048>
- Henry, V. K., Vitullo, A., Fahami, D. I., Seyfried, P., & Ghiso, M. P. (2025). Writing “right” with multilingual learners: Translanguaging, hegemonic writing instruction, and the possibility of not. *Reading Research Quarterly*, 60(4), Article e70038. <https://doi.org/10.1002/rrq.70038>
- Jeon, J., Wei, L., Tai, K. W. H., & Lee, S. (2025). Generative AI and its dilemmas: Exploring AI from a translanguaging perspective. *Applied Linguistics*, 46(4), 709–717. <https://doi.org/10.1093/applin/amaf049>
- Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon Press.
- Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press.
- Sun, Y. (2026). Resisting homogenization in EAL writing education: Translingual practices in the rise of generative AI. *TESOL Quarterly*. Advance online publication. <https://doi.org/10.1002/tesq.70063>

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Zhang, C., Yao, C., Wu, J., Lin, W., Liu, L., Yan, G., & Ying, F. (2022). StoryDrawer: A child–AI collaborative drawing system to support children’s creative visual storytelling. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. <https://doi.org/10.1145/3491102.3501914>