AI-supported development of image-based tasks for assessment of event knowledge

University of Pittsburgh
Learning Research and Development Center

Chrabaszcz A., Lahner C., Baghel B.K., Olson H., Wang A., Lizar G., Caudle C., Li X.L., Dresang H.C., Dickey M.W., Warren T.

I. MOTIVATION

METHODOLOGICAL

- Image-based tasks are widely used in neuropsychological assessments because they can target different cognitive domains and minimize reliance on verbal expression/ comprehension.
- Manual creation of image stimuli is costly and time-consuming.
- AI text-to-image (T2I) models offer advantages for image generation: *speed, cost efficiency, accessibility, scalability, customizability*.
- There is a need to establish best practices for AI image **generation** and **evaluation** for research and clinical purposes.

THEORETICAL

- Conceptual knowledge is primarily assessed through judgement of isolated concepts: objects and actions
- Everyday life requires knowledge of more than discrete objects and actions → individuals must understand complex **events** involving participants, actions, objects, instruments, goals, context (e.g., time, location) and physical-contextual appropriateness.
- The world unfolds dynamically → individuals must continuously integrate discrete events into coherent, temporally ordered **event schemas** - the normal sequence of actions and expectations for roles / rules for a setting (e.g., what happens when you walk into a classroom).

Develop an AI-supported pipeline for flexible, customizable design of image-based tasks for assessment of event knowledge

II. APPROACH

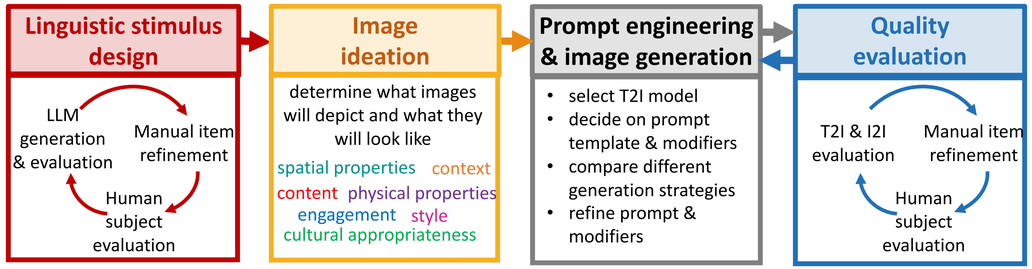
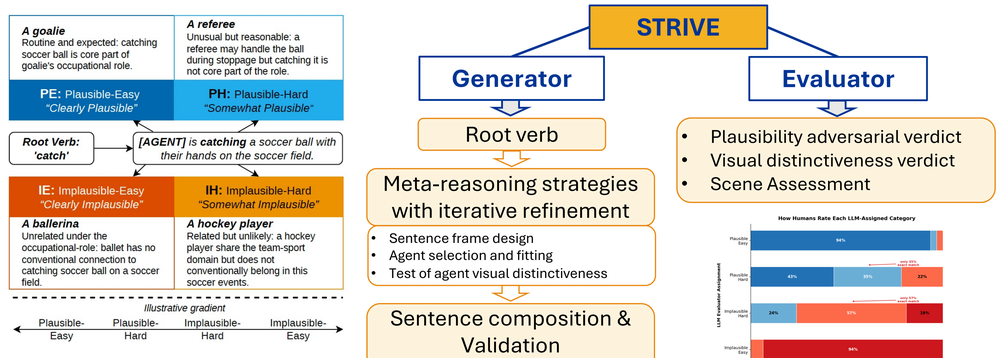


Fig 1. STRIVE Generation-Evaluation Framework for sentence stimuli with varying thematic role fillers (AGENT is used as an example). Given the verb, the generator produces sentences of structure s = (**AGENT**, VERB, PATIENT, INSTRUMENT, LOCATION), varying only the AGENT.



Model output

Model	Method	Verb	PE sentence	PH sentence	IH sentence	IE sentence
sonnet-4.6	RB	fish	A fly fisherman is fishing for a trout with a fishing rod on the riverbank.	A park ranger is fishing for a trout with a fishing rod on the riverbank.	A scuba diver is fishing for a trout with a fishing rod on the riverbank.	A ballet dancer is fishing for a trout with a fishing rod on the riverbank.

Fig 2. Three-Level Evaluation Framework for image series, combining T2I, I2I, and Narrative analyses. It uses Vision-Language Models (VLM) as a judge to flag misaligned content along a pre-defined rubric (e.g., incorrect object representation, spatial inconsistency, etc.).

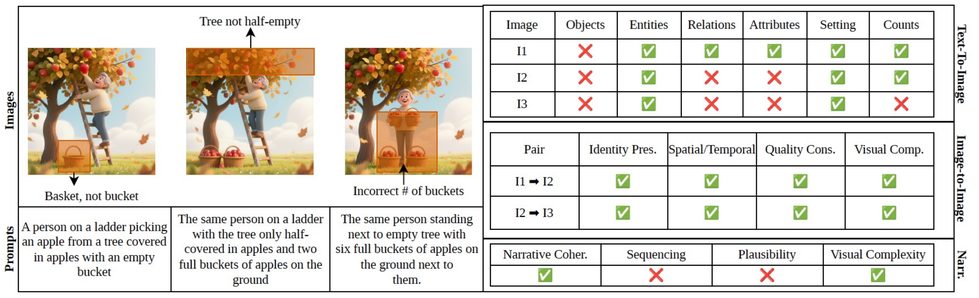


Image	Objects	Entities	Relations	Attributes	Setting	Counts
I1	✗	✓	✓	✓	✓	✓
I2	✗	✓	✗	✗	✓	✓
I3	✗	✓	✗	✗	✓	✗

Pair	Identity Pres.	Spatial/Temporal	Quality Cons.	Visual Comp.
I1 ➡ I2	✓	✓	✓	✓
I2 ➡ I3	✓	✓	✓	✓

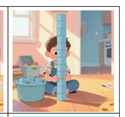
Narrative Coher.	Sequencing	Plausibility	Visual Complexity
✓	✗	✗	✓

III. ITEM SETS

Fig 3. Event Task Stimuli (n = 64), varying in plausibility, difficulty and role fillers (AGENT, PATIENT, INSTRUMENT, LOCATION).

PE: Plausible Easy (n = 16)		Typical events with frequent role fillers	The child is pulling a toy car with a rope at the playground
PH: Plausible Hard (n = 16)		Typical events with infrequent role fillers	The referee is catching the ball with his hands in the field
IE: Implausible Easy (n = 16)		Atypical events with implausible role fillers	The woman is peeling a tennis ball with a knife in the kitchen
IH: Implausible Hard (n = 16)		Atypical events with implausible role fillers that are semantic associates with one of the sentence constituents	The man is lighting the cigarette with a flashlight in the street

Fig 4. Event Schema Task Stimuli (n=60), varying in difficulty & N of steps (2, 3, 4).

C1	Clear causal ordering with visually trackable objects(s)		
C2	Clear causal ordering without visually trackable objects(s) or where object(s) change shape		
C3	Causally coherent, more complex activities often comprised of subevents in different locations		
C4	Culturally or conventionally ordered sequences without inherent causal links		

IV. PILOT DATA

5 participants with chronic stroke and history of aphasia completed 2 administrations of each task

subj	WAB-AQ	aphasia-class	event 1	event 2	schema 1	schema 2	PPT	KDT	ACOM
1003	51.2	wernicke	0.97	0.94	0.89	0.90	0.96	0.92	57.44
1010	97.4	latent	0.95	0.97	0.88	0.92	0.98	0.96	55.41
1011	97.6	latent	0.89	0.94	0.96	0.97	0.96	0.98	60.42
1012	73.9	broca	0.97	0.92	0.89	0.89	0.96	0.98	57.30
1016	72.6	broca	0.89	0.94	0.73	0.74	0.85	0.83	49.00

WAB-AQ = Western Aphasia Battery-Aphasia Quotient and Classification. Event 1 and 2 = % accuracy in 1st and 2nd administrations of Event Task. Schema 1 and 2 = % accuracy in 1st and 2nd administrations of Event Schema Task. PPT = Pyramids and Palm Trees. KDT = Kissing and Dancing Test. ACOM = Aphasia Communication Outcome Measure (t score).

For **Event Task**, participants indicated whether a given image was plausible or implausible.

- High accuracy rates were observed, with near ceiling performance on implausible items; RTs were also faster for implausible than plausible items (β = -0.73, p = .033).
- Within plausible items, hard items produced more errors than easy ones (β = 2.50, p = .003); RT differences were not significant.

For **Event Schema Task**, participants rearranged printed image sets in the correct temporal order.

- Accuracy was not sensitive to difficulty category, but decreased as sequence-length increased (β = -0.11, p = .028).

V. CONCLUSIONS

- AI-generated event tasks are **feasible** for use in clinical research settings, as supported by pilot data.
- AI tools show promise for **efficient** stimulus construction and task design, reducing development time.
- **Human oversight** remains essential: AI-generated materials require careful validation against established empirical and clinical standards.
- Best practice: **hybrid approach** (combine AI-driven generation and evaluation with expert refinement).
- Future work should focus on **systematic validation** and **iterative improvement** of AI-developed materials.