POSITION PAPER · EVALEVAL @ ACL 2026

Evaluating Creative Generative AI with Conceptual Blending

Suzie Xi[1] · Zhongheng Cheng[2] · Deepak Kumar Soni[3] · Mark Turner[3]

[1]University of Notre Dame [2]Columbia University [3]Case Western Reserve University

email: wxi@nd.edu

ACL 2026 SAN DIEGO JULY 2–7

01 MOTIVATION AND POSITION

Evaluating creativity as the emergence of meaning.

As generative models enter real creative workflows, evaluation has lagged. Scores reduce creativity to surface-level novelty that is easy to count but weakly meaningful, leaving assessments unanchored, inconsistent, and hard to reproduce.

FOUR WAYS CREATIVITY EVALUATION FAILS

no shared standard	Creativity is scored differently across every model and benchmark.
surface novelty	Lexical variety and style stand in for genuine conceptual integration.
weak theory ground	Metrics lack interpretable mechanisms grounded in cognitive or linguistic science.
no unified protocol	Without shared prompting and evaluation, results do not reproduce across studies.

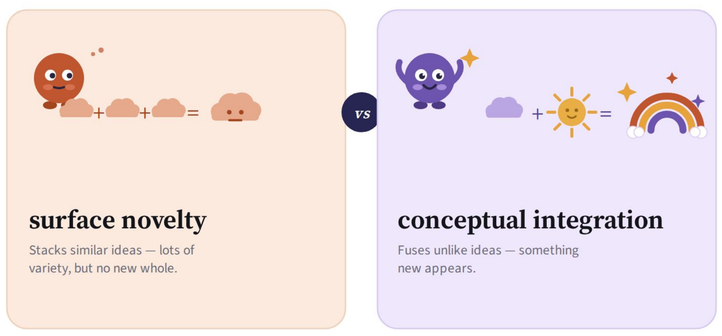
POSITION STATEMENT

> We propose a **FrameNet-grounded framework** for evaluating creativity as **conceptual integration** rather than surface novelty.
>
> A theory-grounded target for creative generation: interpretable, comparable, and reproducible evaluation.

02 We define creativity as

- **Creativity is not random novelty, but the ability to generate outputs that integrate distinct conceptual domains to produce emergent structure or meaning not reducible to either source alone.**
- **Distinct from temperature tuning, stylistic prompting, or scale-induced emergence — these raise surface novelty.**
- **Blending turns a fuzzy construct into a structured object can be specify, reproduce, and score.**

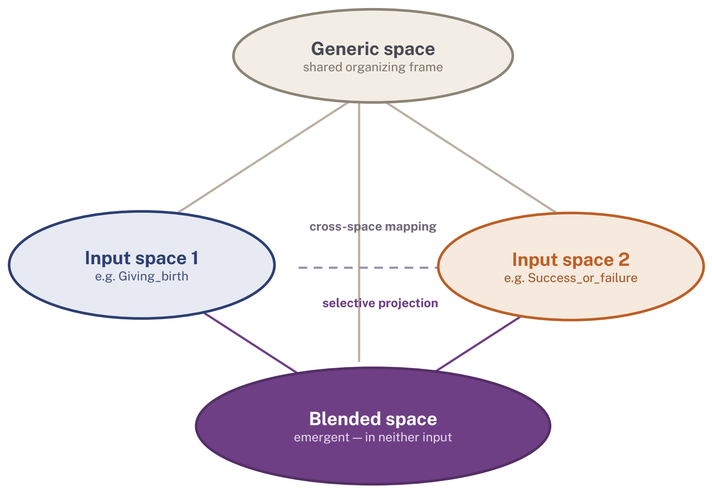
TWO MODES OF MAKING SOMETHING NEW



03 Theoretical roots

- **Bisociation of matrices** — creative insight begins when previously unrelated frames of thought are brought together (Koestler, 1964).
- **Conceptual blending** — input spaces are selectively projected into a blended space through shared structure (Fauconnier & Turner, 1998, 2002, 2003).
- **Emergent meaning** — the blend compresses complex relations into a new, usable idea: the basis of metaphor, analogy, and insight.
- **Why this matters for GenAI** — it gives creativity evaluation a theory-grounded target: measure conceptual integration, not just novelty.

04 Conceptual blending



Two input spaces are linked by a **cross-space mapping** and selectively projected into a **blended space**, organized by a shared generic space. The blend carries emergent structure present in neither input (Fauconnier & Turner, 1998, 2002, 2003).

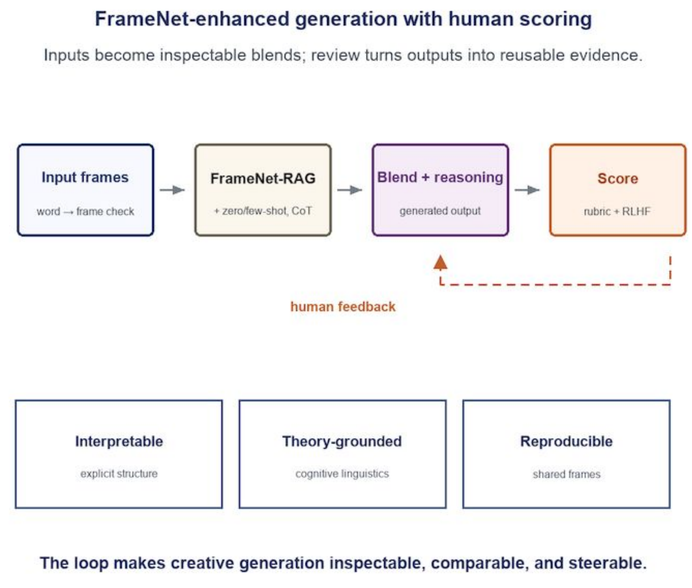
05 Conceptual vocabulary of creativity

06 Evaluation

Completeness	Both frames clearly represented and integrated.
Clarity	Fluent, clear, and concise output.
Relevance	Core roles and relations from both frames preserved.
Understanding Depth	Non-trivial integration with genuinely emergent meaning.
Coherence	Logically consistent and well integrated.
Execution Time	Efficient, minimal intervention, reproducible.

Understanding Depth and **Coherence** target Mini-C and Little-C creativity — everyday, meaningful integration, not eminent originality (Kaufman & Beghetto 2009).

07 Human-in-the-loop pipeline



08 FrameNet & frame blends

We instantiate blends over **FrameNet** semantic frames: 1,000+ frames and 10,000+ lexical units give structured inputs that make conceptual integration explicit, inspectable, and measurable.

Giving_birth ⊕ Success_or_failure →

"The successful delivery of her baby was like hitting a financial jackpot… every moment of labor felt like a valuable investment in her future."

Childbirth mapped onto financial investment — meaning that lives in neither source frame alone.

Emotion ⊕ Kinship →

"Emotion is like kinship — worth investing in, but easily squandered if not managed properly."

Emotions recast as relational resources; the blend supplies the emergent "investment" logic.

09 Conclusion

Conclusion makes creativity **interpretable, theory-grounded, and reproducible** — targeting output homogenization, weak steering, and atheoretical novelty metrics.

- A working definition of creativity as cross-domain conceptual integration.
- A FrameNet-based substrate that makes blends explicit and measurable.
- A multidimensional rubric and a human-in-the-loop generation pipeline.
- An invitation: build shared annotation protocols, test inter-rater reliability, and extend to multimodal creativity.

10 References

[1] Fauconnier, Gilles, and Mark Turner. 1998. Conceptual integration networks. Cognitive Science, 22(2):133-187.

[2] Fauconnier, Gilles, and Mark Turner. 2002. The Way We Think: Conceptual Blending and the Mind's Hidden Complexities. Basic Books, New York.

[3] Fauconnier, Gilles, and Mark Turner. 2003. Conceptual blending, form and meaning. Recherches en Communication, 19:57-86.

[4] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics.

[5] James C. Kaufman and Ronald A. Beghetto. 2009. Beyond Big and Little: The Four C Model of Creativity. Review of General Psychology, 13(1):1-12.

[6] Arthur Koestler. 1964. The Act of Creation. Hutchinson.