Pencils to Pixels: A Systematic Study of Creative Drawings across Children, Adults and AI

Surabhi S Nath[1,2,3], **Guiomar del Cuvillo y Schröder**[4], **Claire E. Stevenson**[4]
[1]Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[2]Max Planck School of Cognition, Leipzig, Germany
[3]University of Tübingen, Tübingen, Germany
[4]University of Amsterdam, Amsterdam, Netherlands

Abstract

Can we derive computational metrics to quantify visual creativity in drawings across intelligent agents, while accounting for inherent differences in technical skill and style? To answer this, we curate a novel dataset consisting of 1338 drawings by children, adults and AI on a creative drawing task. We characterize two aspects of the drawings—(1) style and (2) content. For style, we define measures of ink density, ink distribution and number of elements. For content, we use expert-annotated categories to study conceptual diversity, and image and text embeddings to compute distance measures. We compare the style, content and creativity of children, adults and AI drawings and build simple models to predict expert and automated creativity scores. We find significant differences in style and content in the groups—children's drawings had more components, AI drawings had greater ink density, and adult drawings revealed maximum conceptual diversity. Notably, we highlight a misalignment between creativity judgments obtained through expert and automated ratings and discuss its implications. Through these efforts, our work provides, to the best of our knowledge, the first framework for studying human and artificial creativity beyond the textual modality, and attempts to arrive at the domain-agnostic principles underlying creativity. Our data and scripts are available on GitHub*.

Keywords: visual creativity; drawings; children; adults; Dall-e; content; style; computational measures

Introduction

Human visual creative expression emerges early on—children start drawing before they can write (Levin & Bus, 2003), and cave paintings predate the written word (Ardila, 2004). However, empirical research in creativity, especially in the context of AI, has focused more on verbal than visual creativity. This discrepancy arises in part due to the complexities of producing and evaluating visual outputs. Evaluating visual creativity involves subjective perceptual judgments, for example aesthetic considerations and challenges of separating creativity from technical skill (Chan & Zhao, 2010). For this reason, visual creativity is commonly assessed using simple shape completion creative drawing tasks (e.g., TTCT Picture Completion, (Torrance, 1966); TCT-DP, (Urban, 2005); MTCI, (Barbot, 2018)). Unlike other forms of visual expression (e.g., paintings, digital art), creating such drawings requires limited technical or artistic expertise, without compromising on creative potential (Barbot & Tinio, 2015) and serves as a useful cognitive tool (Fan, Bainbridge, Chamberlain, & Wammes, 2023).

However, evaluating such drawings are still challenging due to the lack of a well-formalized framework for assessment, especially across intelligent agents, a crucial comparison in the present era where human and machine creativity increasingly intersect (Acar, 2023; O'Toole & Horvát, 2024; Marr, 2023). For example, how can we meaningfully compare the creativity of a child's pencil drawing to that of a 1024×1024 pixel image generated by an AI model? To bridge this gap, we make two key contributions: (1) we curate a novel dataset of creative drawings and evaluations spanning different intelligent agents, (2) we develop a computational framework that quantifies two core aspects of the drawings—content, and style, and use them to study drawing creativity.

Dataset. In order to create a diverse dataset, we take note of the vast literature studying individual differences in drawing abilities (Chan & Zhao, 2010), particularly across development (Lowenfeld, 1957; Philippsen, Tsuji, & Nagai, 2022; Heard, 1988; Hart et al., 2022; Narvaez, Polsley, & Hammond, 2024). Further, we note that in AI text-to-image models, prompting can lead to significant diversity in outputs (Oppenlaender, Linder, & Silvennoinen, 2024). Therefore, we curate drawings from children in two age groups, adults, and AI models prompted using a collection of prompts.

For each drawing, we obtain human ratings from two experts raters and get automated scores from two recently released tools for automated assessment of drawing creativity, AuDrA (Patterson, Barbot, Lloyd-Cox, & Beaty, 2024) and OSC-figural (Acar, Organisciak, & Dumas, 2023). Research suggests that AI-based evaluation methods favor AI-generated responses, whereas human evaluators prefer human-created outputs (Laurito et al., 2024). Therefore, by incorporating both expert (Kaufman & Baer, 2012) and automated evaluation (Cropley & Marrone, 2022), we obtain a balanced perspective on creativity assessment.

Framework. To analyze this diverse dataset, we begin by distinguishing two key aspects of creative drawings: content (*what* is depicted) and style (*how* it is rendered). This distinction between content and style is studied in vision research, for example in the context of art perception (Augustin, Leder, Hutzler, & Carbon, 2008; Augustin, Defranceschi, Fuchs, Carbon, & Hutzler, 2011) and in generative artificial intelligence (Kotovenko, Sanakoyeu, Lang, & Ommer, 2019; Zhang, Zhang, & Cai, 2018).

We explore various computational metrics for characteriz-

*https://github.com/surabhisnath/pencils_to_pixels

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

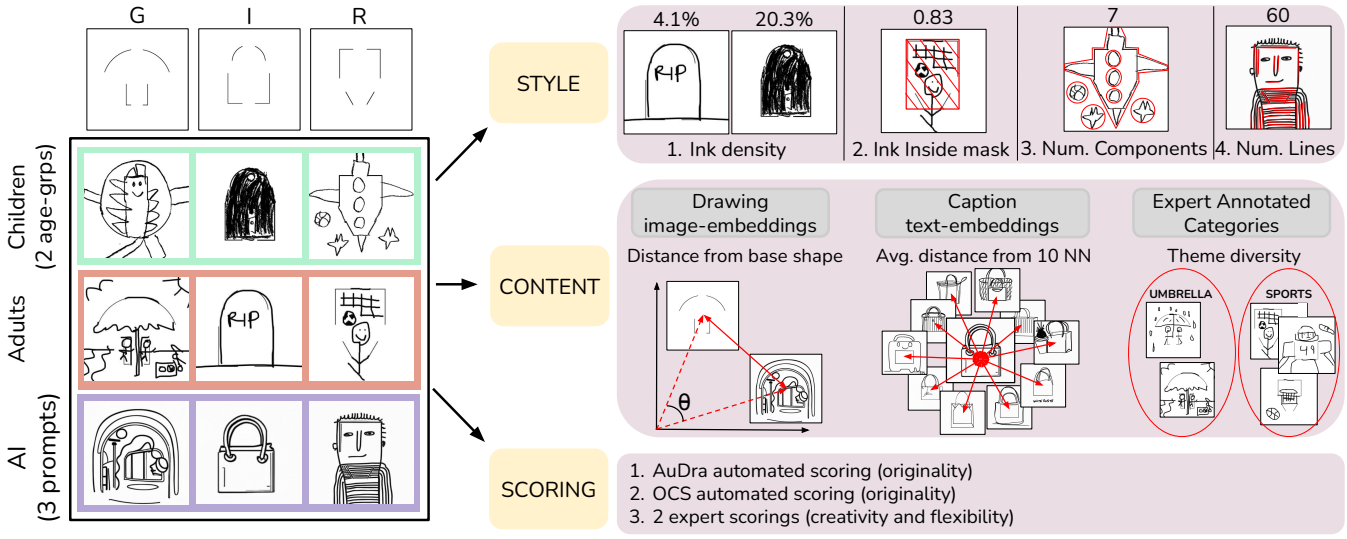

Figure 1: Overview of task and framework for studying creative drawing. The three MTCI stimuli G, I, R are shown on the top left below which example drawings by children, adults and AI are shown. Our style measures are shown on the top right panel. The middle right panel shows our content measures. The bottom right panel lists the different creativity scoring methods.

ing content and style in creative drawings. For style, the number, length, distribution or smoothness of strokes are popular choices (Thomas, Powell, Polsley, Ray, & Hammond, 2022). Other measures related to order or complexity from aesthetics research can be useful candidates (Nath, Brändle, Schulz, Dayan, & Brielmann, 2024; Van Geert & Wagemans, 2020).

For content, we propose a multimodal approach that utilizes the drawing itself, captions generated for the drawing, or expert annotations. For this, one could begin by extending the measures developed for textual responses (Ismayilzada, Stevenson, & van der Plas, 2024; Ramesh & Sanampudi, 2022). For example, *SemDis*, measuring the distance between a response and the task stimuli (Beaty & Johnson, 2021), or inverse frequency, quantifying the uniqueness of a response (Weaver, Caldwell, & Sheafer, 2019), or process flexibility, tracking jumps in the responses sequence (Nath, Dayan, & Stevenson, 2024) can be adapted for drawings.

Together, with this dataset and framework, we are the first to systematically characterize differences in creative expression across intelligent systems, and arrive at core computations underlying visual creativity, and creativity generally.

Methods

Data

Our dataset consists of 1338 drawings by children, adults and AI on stimuli from the Multi-trial Creative Ideation (MTCI) Task (Barbot, 2018) (Fig 1 left panel).

Children Data comprises of 444 drawings from 148 children, 84 from kindergarten (4-6 year olds) and 64 from lower elementary level (7-9 year olds) from a public Montessori school. These two groups align with two stages of drawing development, where ages 4-6 are considered *pre-schematic* and ages 7-9 *schematic* (Lowenfeld, 1957).

Data collection took place in the classroom in small groups. A trained research assistant explained the task. Then each child was given a thick pencil and a piece of paper with the stimulus printed on it. They were given 5 minutes to complete their drawing, after which the next stimulus was given. Each child completed three drawings, one for each of the stimuli shapes G, I and R.

Since the children drew freely on paper, they sometimes ignored instructions and flipped the paper by 180°, drawing on the inverted stimulus. Stimulus 'R' (resembling an inverted house) was flipped most often, with 60% drawings on 'R' being inverted. Across the whole dataset, about 30% drawings (nearly the same ratio in both pre-schematic and schematic) were drawn on inverted stimuli.

Adults Data comprises of 444 drawings from 148 participants, who each completed drawings for stimuli shapes G, I, and R on the MTCI hosted by Barbot's Crealyx platform, an open-access, online testing platform dedicated to the assessment of creativity.

AI We treated the MTCI task as an inpainting task. We prompted Open AI's Dall-e (v2), through their API in image editing mode using three prompts (see box below) for each of three stimuli shapes G, I, R. We collect a total of 50 images per prompt per stimuli, resulting in a final dataset of $(50 + 50 + 50) \times 3 = 450$ images.

The first prompt, based on Chen (2023) contained clear stylistic instructions (colour, pen type, thickness, art style etc.) with no explicit instruction for content (to match the task instructions given to humans), and was set as the base prompt (prompt 1). Prompts 2 and 3 extended the base prompt with explicit content instructions for creating objects/scenes (prompt 2) and living figures (prompt 3).

Prompts

Prompt 1: *creative minimalist black-on-white drawing, lineart-style on white background, drawn with digital fineliner, no color or shading*

Prompt 2: *creative minimalist black-on-white drawing of day-to-day object or scene, lineart-style on white background, drawn with digital fineliner, no color or shading*

Prompt 3: *creative minimalist black-on-white drawing of living figures (human, animal or creature), lineart-style on white background, drawn with digital fineliner, no color or shading*

Two-thirds of the generated drawings did not follow the instructions (either ignored the stimulus or added colors), so we produced 1350 drawings to obtain 450 valid ones.

Preprocessing

Since the children, adults and AI data came from different sources, it was crucial to preprocess the images for valid comparisons. We controlled for size, colours and line thickness using computer vision techniques. The children and Dall-e drawings are cropped and resized to 400×400 size to match the adults. The cropping was specified in a way to ensure the stimulus size and position was aligned across all drawings. The children's drawings were made with pencil and are therefore shades of gray. All drawings were binarized and cast to black-and-white. Children drawings also had tiny scattered pencil spots which were removed using image erosion. The lines for children and AI drawings were dilated to match the line thicknesses of adult drawings. Post processing, we confirmed a non-significant difference in line thickness across the three groups using a Kruskal-Wallis test ($p > 0.1$).

Measures

We developed computational measures of style and content to characterise drawing creativity.

Style We quantify (1) ink density, (2) fraction of ink inside the stimuli shape, (3) number of components, and (4) number of lines.

(1) Ink density: The percentage of the drawing covered in ink, measured by dividing the number of black pixels by the total number of pixels times 100 (Figure 1 Style panel 1.).

(2) Fraction of Ink inside the Stimulus Shape: Quantified by the amount of ink inside the stimulus's bounding box divided by the total amount of ink. Bounding boxes for the base stimuli were obtained by identifying the extreme ink points defining their boundaries. (Figure 1 Style panel 2.).

(3) Number of Components: Counts the number of visually distinct regions in the drawing, based on the graph theoretic property of reachability (based on (Nath, Brändle, et al., 2024), Figure 1 Style panel 3.).

(4) Number of Lines: Skeletonizes the drawing to extract its structural outline and then applies the Hough Transform to detect and count straight lines (Figure 1 Style panel 4.).

Content We use the `clip` image-embedding model (OpenAI, `clip-vit-large-patch14`) to obtain image embeddings per drawing. We compute the cosine distance between the embeddings of the drawing and the corresponding base stimulus shape (G, I or R). Based on the nature of `clip`'s training data, we know the distance measures induced by these encoders are largely influenced by semantic rather than stylistic attributes (Udandarao, Burg, Albanie, & Bethge, 2023; Rashtchian et al., 2023).

For captions, we use `GPT4o` to generate image descriptions for each drawing. We explicitly state in the prompt to give short captions (<15 words), describing the content (and not style) of the drawing, and to caption "hard to interpret" drawings as such. We then used `gtelarge` text-embedding model to obtain caption-embeddings per caption per drawing. Using these, we (1) cluster the embeddings hierarchically to arrive at core semantic themes expressed in the drawings. (2) We define a measure of semantic uniqueness for each drawing by computing the mean cosine distance of the caption-embedding to its ten nearest neighbours. The nearer a caption-embedding is to others, the more popular and less unique the drawing concept is in the dataset of drawings.

Annotation

Drawing content was also annotated by an expert. Each drawing was classified into a minimum of one and a maximum of three (from most salient to least salient) concept categories. If the drawing was hard to interpret, it was assigned to a "hard to interpret" category. Using these categories, we evaluate conceptual diversity for the different groups by diving the number of unique categories per group by the total number of unique categories.

Drawing captions generated by `GPT4o` were also validated by an expert with a correct/incorrect label per caption per drawing. We found that `GPT4o` captions scored nearly 83% correct across all drawings.

To measure process aspects, we obtained flexibility (conceptual diversity within the outputs of the same agent) scores by two expert raters. Flexibility was scored in a range of [0-2] for the three drawings by the same participant. Since there are no explicit participants for the Dall-e drawings, we randomly sample sets of three Dall-e drawings, one from each stimulus shape under the same prompt. We sample 50 random samples sets per prompt (without replacement so that each image is sampled once) and obtain flexibility scores for each set.

Creativity Scoring

Creativity was scored by four sources. Two experts rated each drawing on a scale of [0-4] following the MTCI scoring protocol. Two automated scoring tools AuDrA (Patterson et al., 2024) and Open Creativity Scoring - Figural (Acar et al., 2023) produced originality ratings in the range [0-1] per drawing.

One expert rater also scored utility in the range [0-2] per drawing, depending on how well the drawing incorporated the base shape.

Results

1. How do Children, Adults and AI Drawings Differ?

We compare the drawings of children, adults and AI based on style and content. We present the subgroups of children and AI separately to note differences across childhood development and prompt-guided content manipulation.

Style Figure 2 presents the boxplots comparing the style measures. We see significant differences in ink densities across the three groups (the within group differences were not significant, Figure 2a). We find that AI drawings have the highest ink density followed by children and then adults with the least. From Figure 2b we see that most of this density resides inside the stimulus shape for children and adults with no significant difference between them (there was a significant difference within the subgroups of children), but resides outside for AI drawings. Further, the children's drawings had a higher number of components compared to adults, and AI had the least (Figure 2c). This means the children draw more separable elements in their drawings whereas the AI produces a connected chunk of ink. Finally, from Figure 2d we see that AI drawings contain a large number of straight lines, significantly higher than children or adult drawings (with no significant difference between children and adults groups or subgroups). Within AI, prompt 3 drawings had significantly lesser lines than prompt 1 or 2, suggesting that the instruction of producing living figures resulted in more curved strokes. Lastly, we found that while children and adults nearly always seamlessly incorporated the stimulus shape in their drawings, AI only did so about 50% of the times.

Content To study the content of the drawings, we identify the common themes across drawings using two methods.

We use expert annotated categories to calculate concept diversity as the number of unique categories per subgroup divided by the total number of unique categories across subgroups (counted to be 253). Adult drawings displayed the maximum content diversity, with their drawings encompassing nearly 50% of all identified categories, followed by children ($\sim$ 30%) and then AI ($\sim$ 18%) (wherein prompt 2 had a significantly higher diversity than other prompts). In the AI drawings, prompt 2 produced more object themes and prompt 3 produced more living themes, compared to prompt 1, suggesting that the prompting was effective.

To understand the classes of themes more closely, we visualise them by hierarchical clustering of `GPT4o` caption-embeddings (Figure 3). We see that adults display the maximum number of themes, some of them also encompassing complex ideas e.g. gravestone, music or leisure activities. The children and AI drawings had a roughly similar number of categories but the themes differed significantly. Children themes included many imaginative concepts such as cartoon characters, wings, antennae, rockets, ice cream or ghosts. The large proportion of houses can be attributed to 60% of 'R' stimuli drawings being inverted. Many of the AI drawings

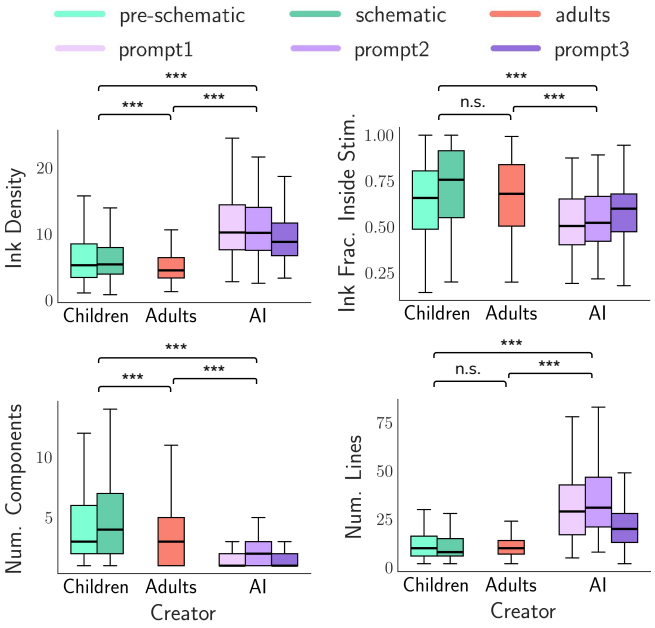

Figure 2: Comparing style measures of ink density, ink fraction inside stimulus, number of components and number of lines in drawings per subgroup. *** denotes $p < 0.01$.

contained abstract themes such as abstract shapes, figures or faces, while some others were complex themes such as furniture or captivity. Children and AI had a relatively similar fraction of drawings which were hard to interpret ($\sim$ 25%), higher than those in adults (16%).

We plot the mean flexibility scores for each group to study process level differences (Figure 4). We note are adults are generally highly flexible with mean flexibility scores largely above 1. Within children, the pre-schematic group has relatively low flexibility with many children scoring a mean under 1. This means these children redrew the same themes across their drawings, disregarding the stimulus. However, in the schematic group, children are strikingly more flexible ($p < 0.01$), rarely repeating themes across their drawings. A similar shift towards higher flexibility is visible as a result of prompting. The base prompt yielded inflexible drawings, whereas prompt 2 and 3 sees a rightward shift in the values, yielding a significant difference in case of prompt 2 (also in line with the greater conceptual diversity in prompt 2).

2. Who is More Creative?

To test which group received the highest creativity scores, we first check the agreement within and between the expert and automated scores. Between the two expert raters, and between the two automated methods, the average fixed raters ICC score was respectively 0.82 and 0.90 ($p < 0.01$), indicating high inter-rater reliability and consistency, and therefore we use mean expert score (normalised using min-max scaling to the range 0-1) and mean automated score for analyses. Interestingly, the ICC between mean expert score and mean automated score is only 0.48 and Spearman correlation co-

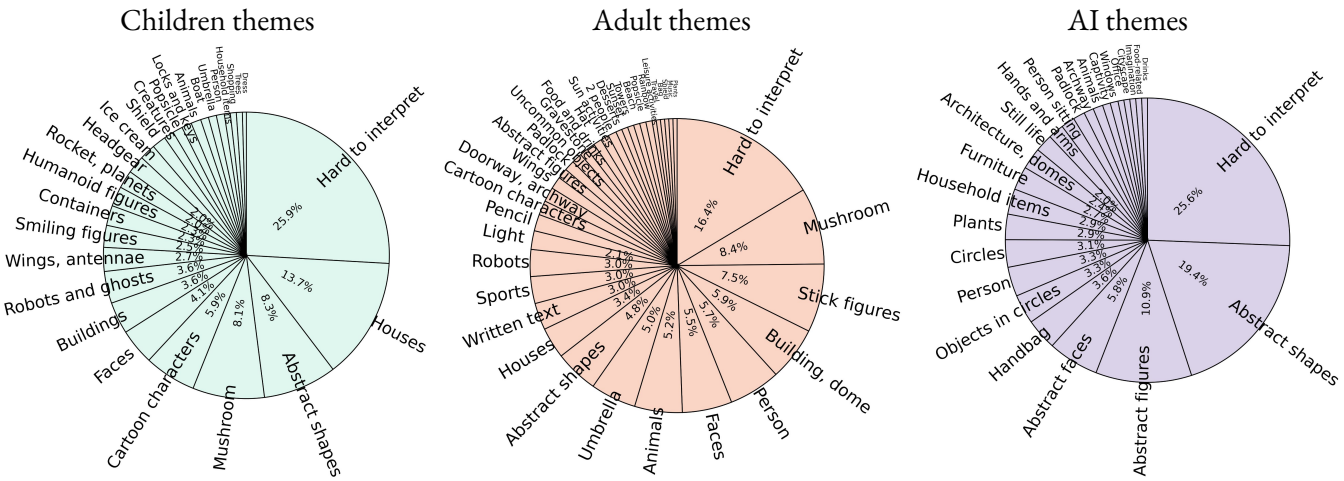

Figure 3: Conceptual themes present in children, adult and AI drawings.

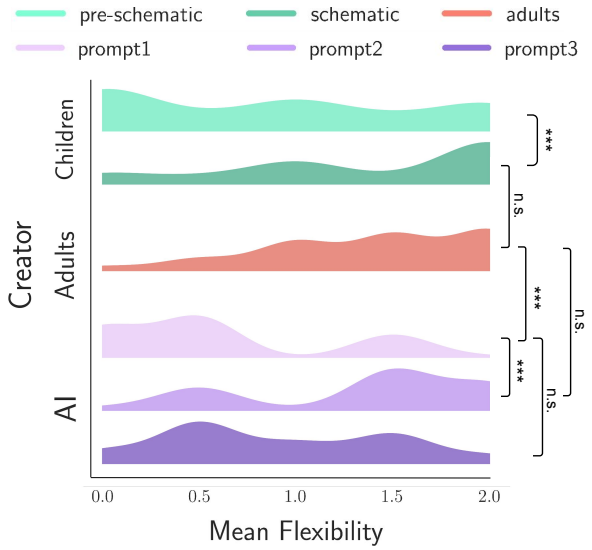

Figure 4: Comparing drawing flexibility across subgroups. *** denotes $p < 0.01$.

efficient is 0.36 ($p < 0.01$). This indicates that the experts and automated scores, though both reliable, do not agree with each other fully on how creative the drawings are.

Further, we test the agreement within the groups to find that the Spearman correlation between mean expert score and mean automated score is higher in adults (0.66, $p < 0.01$), compared to children (0.59, $p < 0.01$, with 0.56 in pre-schematic), and lowest in AI (~ 0.30, $p < 0.01$ for each prompt subgroup). This can be explained by the fact that the automated models were trained on adult drawings and can therefore not generalize easily to other subdomains such as drawings by a different group of humans like pre-schematic children, or to non-human, artificial drawings by AI.

Table 1 reports the mean and standard deviation of mean creativity score by experts and automated tools for each sub-

		Mean score expert	Mean score automated
Children	Pre-schematic	0.38 ($\pm$ 0.26)	0.50 ($\pm$ 0.12)
	Schematic	**0.50** ($\pm$ 0.24)	0.53 ($\pm$ 0.11)
Adults		0.43 ($\pm$ 0.25)	0.49 ($\pm$ 0.12)
AI	Prompt1	0.21 ($\pm$ 0.16)	0.57 ($\pm$ 0.08)
	Prompt2	0.36 ($\pm$ 0.17)	**0.62** ($\pm$ 0.08)
	Prompt3	0.32 ($\pm$ 0.18)	0.61 ($\pm$ 0.07)

Table 1: Mean and standard deviation of drawing creativity scores by expert raters and automated tools per subgroup. Bold incates the subgroup with the highest score.

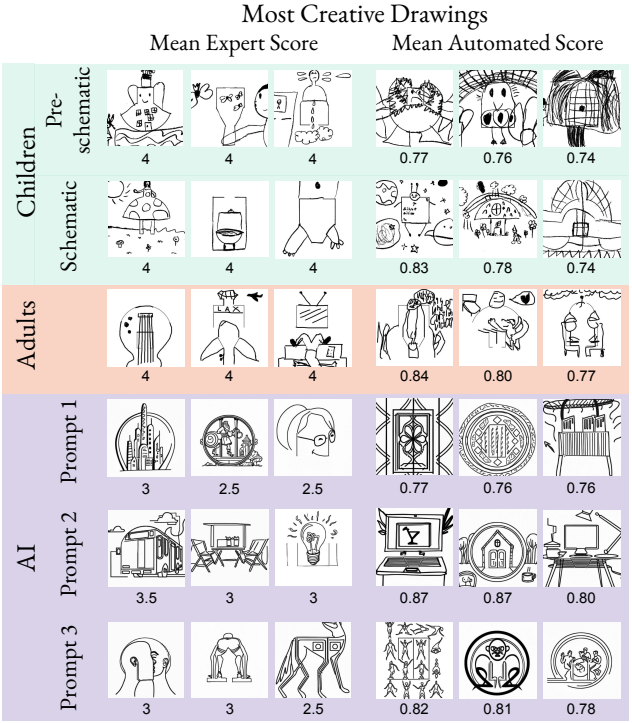

Figure 5: Top three drawings per subgroup receiving the highest mean creativity scores by expert raters, and automated tools. Value indicates mean creativity score.

Model	R^2_{test}	Cor_{test}
expert $\sim$ (used_stim + hard_to_interpret) * (10NN_text + dist_from_stim) + (1 | subgroup)	0.60	0.79
automated $\sim$ ink_density + dist_from_stim + 10NN_image + (1 | subgroup)	0.59	0.78

(a) Best models predicting expert and automated scores.

	used_stim	hard_to_interpret	10NN_text	dist_from_stim	ink_density	10NN_image
expert	**0.16***	**-0.77***	**0.14***	**0.35***	0.15*	-0.04
automated	0.00	0.03	0.07*	**0.34***	**0.44***	**0.10***

(b) Feature contributions for the two best models. * indicates $p < 0.01$. Bold marks the higher weight across the two models.

Table 2: Comparison of model performance and feature contributions

group. We see that the experts rated the schematic group as most creative, followed by adults and rated AI drawings the least creative. On the other hand, the automated tools rated adults the least creative, followed by children and scored the AI drawings as most creative. Across expert and automated scoring, within children's drawings, the schematic group scored higher than the pre-schematic group, and within AI drawings, prompt 2 scored the highest, followed by prompt 3 and the base prompt was scored the lowest.

We visualize the drawings that received the highest scores in Figure 5. Interestingly, there are no overlaps in the top three highest scored drawings by expert and automated scores in any subgroup. This suggests that experts and the automated tools use very different strategies to rate creativity. We test this more formally in the next section.

3. What do human experts and automated tools value when rating visual creativity?

We use linear mixed effects regression to test the role of our style and content measures in explaining variance in mean creativity scores by experts and automated tools. We perform stratified, 3-fold cross validation and report R^2_{test} and Speaman correlation Cor_{test} metrics for the models with the lowest BICs (Table 2a). We find that the expert score depended on originality, in both image and text space—*i.e.* on how far the drawing is from the stimulus in image space ($dist\_from\_stim$), and on how similar the caption is to its ten nearest neighbours in text space ($10NN\_text$). Importantly, their scores also depended on how seamlessly the drawing incorporated the stimulus ($used\_stim$), and whether the drawing was easy to interpret ($hard\_to\_interpret$), which can be considered as evaluations of drawing utility. On the other hand, the automated score is largely influenced by the amount of ink ($ink\_density$) and originality measures operating purely largely in image space ($10NN\_image$), ignoring the multimodal nature of creativity.

We see this dissociation clearly in Table 2b, which reports the regression coefficients for a combined linear model with all the predictors from the individual best models (VIFs are under 5). We see that the predictors useful for explaining experts scores, namely $used\_stim$, $hard\_to\_interpret$ and $10NN\_text$ are not effective predictors of mean automated score, while $ink\_density$ and $10NN\_image$ are not effective predictors of mean expert score. $Dist\_from\_stim$ was the only predictor contributing significantly to both models.

Discussion

Our work develops a novel dataset and a computational framework based on *style* and *content* to study creative drawings of children (across two age groups), adults and AI (across three prompts). This framework was useful in studying differences in drawings across the groups—children's drawings had more visual components, and depicted imaginative themes (Latham & Ewing, 2018). Adult drawings displayed more conceptual diversity, and AI drawings had the highest ink density. We saw a striking increase in flexibility from pre-schematic to schematic children (Spensley & Taylor, 1999). Also, content-driven prompting improved creativity in AI (Hao, Chi, Dong, & Wei, 2024). Unlike most textual creative tasks (Bellemare-Pepin et al., 2024; Marco, Rello, & Gonzalo, 2024), Dall-e drawings were hard to produce and differed greatly from human outputs. Methods beyond prompting, for example agentic interactions (Vinker et al., 2024) or fine-tuning could help align AI drawings more with humans' (Liang et al., 2024).

Our framework also helps understand creativity and its measurement. Interestingly, which group was most creative depended on the evaluator. We confirm a self-bias where human experts preferred human drawings, and automated models preferred AI drawings (Magni, Park, & Chao, 2024). As with textual creativity, both the automated methods and expert ratings value originality (distance from stimulus) (Kenett, 2019) and novelty (distance from other responses) (Runco & Acar, 2012). But, the automated tools preferred ink density while expert scores valued utility. This highlights an important shortcoming of automated tools which lack understanding of underlying concepts and are therefore unable to incorporate the utility (effectiveness) dimension in creativity evaluation. However, this reflects positively on our search for domain-agnostic determinants of creativity—in line with past work, originality and utility interacted to predict creativity (Diedrich, Benedek, Jauk, & Neubauer, 2015).

Despite challenges in scale unification across intelligent agents and manual curation of features, our framework, decomposing creativity into *what* and *how*, can extend beyond the visual domain to other forms of creative expression.

References

Acar, S. (2023). Creativity assessment, research, and practice in the age of artificial intelligence. *Creativity Research Journal*, 1–7.

Acar, S., Organisciak, P., & Dumas, D. (2023). Automated scoring of figural tests of creativity with computer vision. *The Journal of Creative Behavior*.

Ardila, A. (2004). There is not any specific brain area for writing: From cave-paintings to computers. *International Journal of Psychology*, *39*(1), 61–67.

Augustin, M. D., Defranceschi, B., Fuchs, H. K., Carbon, C.-C., & Hutzler, F. (2011). The neural time course of art perception: An erp study on the processing of style versus content in art. *Neuropsychologia*, *49*(7), 2071–2081.

Augustin, M. D., Leder, H., Hutzler, F., & Carbon, C.-C. (2008). Style follows content: On the microgenesis of art perception. *Acta psychologica*, *128*(1), 127–138.

Barbot, B. (2018). The dynamics of creative ideation: Introducing a new assessment paradigm. *Frontiers in psychology*, *9*, 2529.

Barbot, B., & Tinio, P. P. (2015). *Where is the "g" in creativity? a specialization–differentiation hypothesis* (Vol. 8). Frontiers Media SA.

Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with semdis: An open platform for computing semantic distance. *Behavior research methods*, *53*(2), 757–780.

Bellemare-Pepin, A., Lespinasse, F., Thölke, P., Harel, Y., Mathewson, K., Olson, J. A., ... Jerbi, K. (2024). Divergent creativity in humans and large language models. *arXiv preprint arXiv:2405.13012*.

Chan, D. W., & Zhao, Y. (2010). The relationship between drawing skill and artistic creativity: do age and artistic involvement make a difference? *Creativity Research Journal*, *22*(1), 27–36.

Cropley, D. H., & Marrone, R. L. (2022). Automated scoring of figural creativity using a convolutional neural network. *Psychology of Aesthetics, Creativity, and the Arts*.

Diedrich, J., Benedek, M., Jauk, E., & Neubauer, A. C. (2015). Are creative ideas novel and useful? *Psychology of aesthetics, creativity, and the arts*, *9*(1), 35.

Fan, J. E., Bainbridge, W. A., Chamberlain, R., & Wammes, J. D. (2023). Drawing as a versatile cognitive tool. *Nature Reviews Psychology*, *2*(9), 556–568.

Hao, Y., Chi, Z., Dong, L., & Wei, F. (2024). Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, *36*.

Hart, Y., Kosoy, E., Liquin, E. G., Leonard, J. A., Mackey, A. P., & Gopnik, A. (2022). The development of creative search strategies. *Cognition*, *225*, 105102.

Heard, D. (1988). Children's drawing styles. *Studies in Art Education*, *29*(4), 222–231.

Ismayilzada, M., Stevenson, C., & van der Plas, L. (2024). Evaluating creative short story generation in humans and large language models. *arXiv preprint arXiv:2411.02316*.

Kaufman, J. C., & Baer, J. (2012). Beyond new and appropriate: Who decides what is creative? *Creativity Research Journal*, *24*(1), 83–91.

Kenett, Y. N. (2019). What can quantitative measures of semantic distance tell us about creativity? *Current Opinion in Behavioral Sciences*, *27*, 11–16.

Kotovenko, D., Sanakoyeu, A., Lang, S., & Ommer, B. (2019). Content and style disentanglement for artistic style transfer. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 4422–4431).

Latham, G., & Ewing, R. (2018). Children's images of imagination: The language of drawings. *The Australian Journal of Language and Literacy*, *41*(2), 71–81.

Laurito, W., Davis, B., Grietzer, P., Gavenčiak, T., Böhm, A., & Kulveit, J. (2024). Ai ai bias: Large language models favor their own generated content. *arXiv preprint arXiv:2407.12856*.

Levin, I., & Bus, A. G. (2003). How is emergent writing based on drawing? analyses of children's products and their sorting by children and mothers. *Developmental psychology*, *39*(5), 891.

Liang, Y., He, J., Li, G., Li, P., Klimovskiy, A., Carolan, N., ... others (2024). Rich human feedback for text-to-image generation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 19401–19411).

Lowenfeld, V. (1957). *Creative and mental growth*. Macmillan.

Magni, F., Park, J., & Chao, M. M. (2024). Humans as creativity gatekeepers: Are we biased against ai creativity? *Journal of Business and Psychology*, *39*(3), 643–656.

Marco, G., Rello, L., & Gonzalo, J. (2024). Small language models can outperform humans in short creative writing: A study comparing slms with humans and llms. *arXiv preprint arXiv:2409.11547*.

Marr, B. (2023). The intersection of ai and human creativity: can machines really be creative? *Forbes. Verkkosivu https://www. forbes. com/sites/bernardmarr/2023/03/27/the-intersection-of-ai-and-human-creativity-can-machinesreally-be-creative/(viitattu 15.3. 2024)*.

Narvaez, R., Polsley, S., & Hammond, T. (2024). Children's drawings speak: Comparing sketch features amongst development groups. In *Proceedings of the 23rd annual acm interaction design and children conference* (pp. 676–679).

Nath, S. S., Brändle, F., Schulz, E., Dayan, P., & Brielmann, A. (2024). Relating objective complexity, subjective complexity, and beauty in binary pixel patterns. *Psychology of Aesthetics, Creativity, and the Arts*.

Nath, S. S., Dayan, P., & Stevenson, C. (2024). Characterising the creative process in humans and large language models. In *Proceedings of the 15th international conference on computational creativity*. Retrieved from `https://arxiv.org/abs/2405.00899`

Oppenlaender, J., Linder, R., & Silvennoinen, J. (2024).

Prompting ai art: An investigation into the creative skill of prompt engineering. *International Journal of Human–Computer Interaction*, 1–23.

O'Toole, K., & Horvát, E.-Á. (2024). Extending human creativity with ai. *Journal of Creativity*, *34*(2), 100080.

Patterson, J. D., Barbot, B., Lloyd-Cox, J., & Beaty, R. E. (2024). Audra: An automated drawing assessment platform for evaluating creativity. *Behavior Research Methods*, *56*(4), 3619–3636.

Philippsen, A., Tsuji, S., & Nagai, Y. (2022). Quantifying developmental and individual differences in spontaneous drawing completion among children. *Frontiers in Psychology*, *13*, 783446.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527.

Rashtchian, C., Herrmann, C., Ferng, C.-S., Chakrabarti, A., Krishnan, D., Sun, D., ... Tomkins, A. (2023). Substance or style: What does your image embedding know? *arXiv preprint arXiv:2307.05610*.

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity research journal*, *24*(1), 66–75.

Spensley, F., & Taylor, J. (1999). The development of cognitive flexibility: Evidence from children's drawings. *Human Development*, *42*(6), 300–324.

Stevenson, C., & Chen, S. (2023, Dec). *Visual creativity in ai.* OSF. Retrieved from `osf.io/4j5nk`

Thomas, X., Powell, L., Polsley, S., Ray, S., & Hammond, T. (2022). Identifying features that characterize children's free-hand sketches using machine learning. In *Proceedings of the 21st annual acm interaction design and children conference* (pp. 529–535).

Torrance, E. P. (1966). Torrance tests of creative thinking. *Educational and psychological measurement*.

Udandarao, V., Burg, M. F., Albanie, S., & Bethge, M. (2023). Visual data-type understanding does not emerge from scaling vision-language models. In *The twelfth international conference on learning representations.*

Urban, K. K. (2005). Assessing creativity: The test for creative thinking-drawing production (tct-dp). *International Education Journal*, *6*(2), 272–280.

Van Geert, E., & Wagemans, J. (2020). Order, complexity, and aesthetic appreciation. *Psychology of aesthetics, creativity, and the arts*, *14*(2), 135.

Vinker, Y., Shaham, T. R., Zheng, K., Zhao, A., Fan, J. E., & Torralba, A. (2024). Sketchagent: Language-driven sequential sketch generation. *arXiv preprint arXiv:2411.17673*.

Weaver, M. B., Caldwell, B. W., & Sheafer, V. (2019). Interpreting measures of rarity and novelty: investigating correlations between relative infrequency and perceived ratings. In *International design engineering technical conferences and computers and information in engineering conference* (Vol. 59278, p. V007T06A008).

Zhang, Y., Zhang, Y., & Cai, W. (2018). Separating style and content for generalized style transfer. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 8447–8455).