Solving Visual Analogies in Born-Digital Documents with an Attention-Based Framework

Amresh Kumar Singh[1], Sandeep Khanna[2], Chiranjoy Chattopadhyay[2], Romi Banerjee[1], and Rahul Kumar Ray[2]

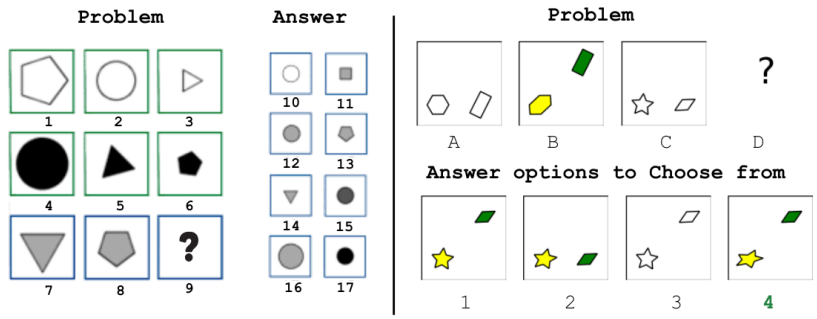
[1] Dept. of Computer Science and Engineering, IIT Jodhpur, Rajasthan, India
[2] School of Computing and Data Sciences, FLAME University, Maharashtra, India

Abstract. Visual analogical reasoning in born-digital documents requires solving proportional analogies (A:B::C:D), a task distinct from sequence completion. It demands the precise mapping of abstract relations from a source pair (A:B) to a target context (C:?). Although born-digital diagrams provide high resolution data, they lack explicit semantic labels for the underlying governing logic. In this paper, we present DigiAnalogy-10K, a structured large-scale benchmark of 10,000 problems for visual proportional analogy (A:B::C:D) in born-digital documents. Unlike benchmarks for progressive matrix completion, DigiAnalogy-10K requires models to identify a latent transformation in an exemplar pair (A to B) and apply it precisely to a query context (C to ?). We also propose VISAA-BD (Visual Instruction and Structure-aware Attention for Analogies in Born-Digital documents), an end-to-end framework that treats analogy solving as a structural-to-semantic mapping task. By utilizing a structure-aware attention architecture, the model encodes the latent transformation between A and B and generates an interpretable rule that can guide the corresponding transformation from C to D. Furthermore, VISAA-BD functions as a visual instruction teacher, generating consistent descriptions in natural language of the governing transformation rules (BLEU-4 = 0.96). To rigorously evaluate this logical mapping, we introduce the Transformation Inference Score (TIS), a task-specific metric that jointly measures operational correctness and hierarchical order preservation. Our work demonstrates that VISAA-BD effectively leverages the structural integrity of born-digital formats to bridge the gap between low-level graphics and high-level interpretable diagrammatic reasoning.

Keywords: Visual Analogical Reasoning · Transformation Rule Generation · Born-Digital Documents · Structure-aware Attention · Diagrammatic Reasoning

1 Introduction

Visual reasoning, the ability to interpret and manipulate graphical representations, is a cornerstone of human intelligence. Within this domain, analogical reasoning is the capacity to perceive and apply relational similarities across distinct contexts, and is a fundamental and distinct cognitive process, classically



(a) Progressive Matrix Completion (b) Visual Proportional Analogy

Fig. 1: Distinction between (a) Progressive Matrix Completion (e.g., RAVEN [24]) and (b) Visual Proportional Analogy. The analogy task applies relation T from the source pair (A, B) to guide the transformation from C to D.

expressed as a proportional analogy: $A : B :: C : D$ [12,8,16]. In digital environments, this translates to the precise mapping of abstract visual transformations (e.g., rotation, color change) from an exemplar pair (A,B) to a query context (C,?). Born-digital diagrams provide high-fidelity structural data ideal for this task, yet they critically lack explicit semantic labels for the underlying logic, creating a gap between low-level graphics and high-level interpretable reasoning. Early work on diagram understanding emphasized explicit object and spatial representations as a prerequisite for higher-level inference, which aligns with our focus on learning transformation rules rather than only recognizing primitives [7]. This work focuses on visual proportional analogies, a task structurally and cognitively distinct from the more commonly studied progressive matrix completion (e.g., [17]) , and closely related to relational reasoning problems studied in pattern recognition and document analysis communities [12]. As illustrated in Figure 1, the former involves inducing a rule from a sequence to complete a pattern (e.g., in a 3×3 grid), while the latter requires mapping an isolated relation between two pairs ($A \rightarrow B$, $C \rightarrow D$).

Prior research has largely focused on the progressive matrix paradigm. Benchmarks such as PGM [2] and RAVEN [24] advanced large-scale evaluation, but are constrained by fixed 3×3 layouts, lack explicit rule annotations, and do not target the proportional analogy structure. Concurrently, models evolved from relation networks [19,2] and program induction [6,10] to modern transformers [26][18], yet they primarily solve for the "next in sequence" and seldom generate human-interpretable rule descriptions. In [3], techniques for parsing visual primitives were proposed; however, they fail to bridge analogical reasoning. Related challenge and competition work on extracting structured information from diagrams and graphics further highlights the importance of synthetic or weakly-structured supervision for reasoning over visual primitives. Unlike natural image reasoning benchmarks, our work operates entirely on born-digital diagrammatic artifacts that exhibit document-like properties such as structured layout, symbolic primi-

Table 1: Summary of related work and key limitations. Our contributions address the highlighted gaps.

Category	**Representative Work**	**Core Contribution**	**Limitations / Gap**
Progressive Matrix Benchmarks	PGM [2], RAVEN [24]	Large-scale, procedural RPM-style datasets.	Fixed 3×3 layout; task is sequence completion, not analogy.
Abstract Reasoning Models	WReN [2], CoPINet [25], PAtNet [21], Transformers [26]	Neural models for RPM and visual rule reasoning.	Mainly target answer selection, with limited explicit rule generation for proportional analogies.
Rule Inference & Explanation	[20]	Recover generative operations or programs.	Outputs are latent codes or symbolic programs, not fluent language.
Document & Diagram Understanding	[3], [5]	Parse symbols and structure in born-digital graphics.	Focus on detection/parsing, not high-level relational reasoning.
Ours (VISAA-BD)	DigiAnalogy-10K, VISAA-BD, TIS metric	Proportional analogy benchmark, interpretable rule generation, structure-aware evaluation.	Directly addresses the task, interpretability, and evaluation gaps.

tives, and rule-governed transformations. This positions visual proportional analogy as a document analysis problem, where understanding structural relations is essential for semantic interpretation, rather than merely recognizing objects. We identify three core gaps, summarized in Table 1:

1. **Task Misalignment:** Existing benchmarks and models target progressive rule induction, not explicit analogical mapping.
2. **Lack of Interpretability:** Systems are "black-box"; they do not produce natural language explanations of the inferred rule.
3. **Evaluation Gap:** Standard metrics (e.g., accuracy) do not evaluate the relational mapping or the correctness of the inferred rule sequence.

To bridge these gaps, we present three key contributions:

1. **DigiAnalogy-10K:** A large-scale benchmark of **10,000 visual proportional analogy problems** (A:B::C:D) in born-digital format. It features stochastic transformations (randomized rotation, shear, and color fill) and a clear $1\times4+4$ layout tailored for relational mapping, not sequence completion.
2. **VISAA-BD (Visual Instruction and Structure-aware Attention for Analogies in Born-Digital documents):** An end-to-end framework that

treats analogy solving as structural-to-semantic mapping. The structure-aware attention module encodes the latent relation T from (A, B) and generates an interpretable transformation rule that can guide the mapping from C to D. Simultaneously, it acts as a *visual instruction teacher*, generating fluent natural language descriptions of the governing rules.

3. **Transformation Inference Score (TIS):** A novel evaluation metric that rigorously assesses analogical mapping by jointly measuring the correctness of inferred operations and the preservation of their hierarchical order, addressing the limitations of standard text similarity metrics.

VISAA-BD achieves high accuracy in rule generation (BLEU-4 = 0.96) and robust analogical mapping. Our work demonstrates how leveraging the structural integrity of born-digital data can bridge low-level graphics and high-level interpretable reasoning for proportional analogies.

The remainder of the paper is structured as follows: Section 2 introduces the dataset. Section 3 describes the proposed framework. Section 4 presents the evaluation metric. Section 5 reports experimental results, including quantitative, qualitative, and robustness analyses, followed by the user study. Finally, Section 6 concludes the paper.

2 Proposed Visual Analogy Dataset: DigiAnalogy-10K

DigiAnalogy-10K is a large-scale benchmark dataset developed for the assessment of the proportional analogy task. It consists of $10,000$ problems. Each problem instance comprises a structured set of panels: three context images (A, B, C) and four candidate answer images. These panels collectively form a complete visual analogy problem in a $1 \times 4 + 4$ configuration. The images are rendered with precise geometric consistency, ensuring that the shapes, spatial relations, and transformation effects are clearly conveyed. It supports eleven geometric primitives: *arrow, star, triangle, rectangle, ellipse, pentagon, hexagon, plus, diamond, heart,* and *circle*. To increase visual diversity, the shapes in panels A and C are subjected to randomized initial rotations sampled uniformly from $[0°, 360°]$. Two additional transformations are introduced, horizontal and vertical shear. Each object also undergoes a color-fill transformation chosen from four discrete colors (red, green, blue, yellow). Other transformations include clockwise

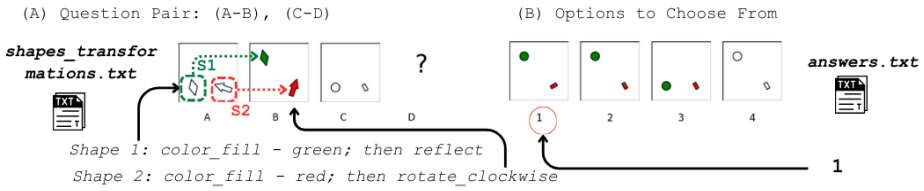


Fig. 2: A sample of DigiAnalogy-10K and its corresponding annotations.

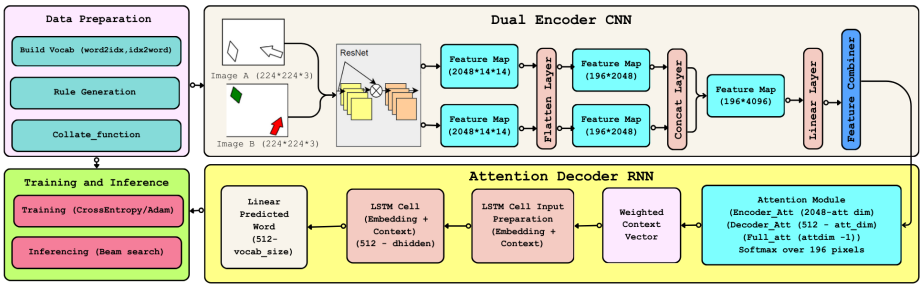


Fig. 3: An end-to-end dual-image encoder with attention-guided LSTM decoder for automatic transformation rule generation from visual analogy pairs.

or counterclockwise rotations ($45°, 90°, 135°$, or $180°$), scale-up ($\times 1.5$) or scale-down ($\times 0.75$), and vertical reflection. An example image of the DigiAnalogy-10K dataset is shown in Fig. 2, along with the annotation texts.

Each instance includes a detailed annotation of the exact visual operations used to generate panel D from C and panel B from A. These transformation logs specify symbolic operation sequences that serve as ground-truth explanations of the visual changes. This symbolic supervision supports training rule-generation models to map pixel-level differences to structured linguistic descriptions. The annotations are precisely aligned with the geometry, providing reliable targets for sequence prediction. To support text modeling, a unified vocabulary is built from all transformation annotations in the dataset. It includes operation names, numerical parameters, separators, symbolic tokens, and color identifiers, ensuring each transformation component is represented by a consistent lexical unit. During training, this vocabulary controls tokenization and decoding, enabling robust handling of variable-length rule sequences.

Together, the structured images, precise transformation annotations, and curated vocabulary make DigiAnalogy-10K a strong benchmark for visual–symbolic alignment and for evaluating models that infer and verbalize abstract visual transformations. DigiAnalogy-10K is designed to mimic the rule-based diagrams common in educational and assessment materials rather than natural images. DigiAnalogy-10K, along with its annotations and generation scripts, has been prepared for research use and will be made available upon request.

3 Methodology

The VISAA-BD framework (Fig. 3) is a dual-visual attention-based encoder–decoder model for generating transformation rules from image pairs. It first encodes the source and transformed states through a shared visual encoder, fuses their features, and then uses an attention-guided LSTM decoder to generate the rule sequence.

Given two input images, $I_A \in \mathbb{R}^{3\times H\times W}$ and $I_B \in \mathbb{R}^{3\times H\times W}$, representing the source (A) and transformed (B) shapes, respectively, the task is to generate a

textual sequence $S = \{w_1, w_2, ..., w_n\}$ that describes the transformation applied to obtain I_B from I_A. Formally, we model this as follows.

$$P(S|I_A, I_B; \theta) = \prod_{t=1}^{T} P(w_t|w_{1:t-1}, I_A, I_B; \theta) \tag{1}$$

where θ denotes all trainable parameters in both the encoder and decoder.

3.1 Data Preprocessing and Vocabulary Construction

All transformation rules are tokenized using a regex pattern,

$$\text{tokens} = \text{regex}(\texttt{r'\textbackslash w+|[;:]'}, \text{text.lower}()) \tag{2}$$

which extracts alphanumeric words and punctuation symbols that mark operation boundaries (e.g. "rotate 90: clockwise"). This ensures that operation names, parameters, and separators are preserved as meaningful units instead of being merged or discarded. The vocabulary maintains bijective mappings word2id : $w_i \rightarrow i$ and idx2word : $i \rightarrow w_i$, and includes the standard tokens $<$ pad $>$, $<$ sos $>$, $<$ eos $>$, and $<$ unk $>$ for padding, sequence control, and unseen words. This design keeps the model stable for variable-length rule descriptions and handles numerical parameters and transformation keywords without requiring handcrafted parsing.

3.2 Visual Encoder

Visual encoding transforms raw pixels into feature representations that capture semantic and spatial information. In VISAA-BD, dual encoders independently process source and target images, producing hierarchical visual embeddings that are later fused to emphasize transformational cues. Each image I_A, I_B is passed through a shared convolutional encoder $f_{enc}(.)$ based on ResNet-50 pretrained on ImageNet. Mathematically, $V_A = f_{enc}(I_A)$, $V_B = f_{enc}(I_B)$. Here $f_{enc}(I) = \text{ResNet}_{50}(I) \in \mathbb{R}^{C \times H' \times W'}$, $C = 2048$, $H' = W' = 14$. Flattening yields feature grids $E_A = \text{reshape}(V_A) = [v_A^1, v_A^2, \ldots, v_A^{196}], \quad v_A^i \in \mathbb{R}^{2048}$ with each $v \in \mathbb{R}^{2048}$ representing a spatial patch. The same equation as above applies for E_B.

Feature Combination Layer The encoder outputs are concatenated along the depth dimension $E_{AB}^{(i)} = [v_A^{(i)}; v_B^{(i)}] \in \mathbb{R}^{4096}$. A projection layer reduces dimensionality while learning a joint transformation.

$$F^{(i)} = W_c E_{AB}^{(i)} + b_c, W_c \in \mathbb{R}^{2048 \times 4096} \tag{3}$$

Hence, the combined encoder output is $F = [F^{(1)}, .., F^{(196)}] \in \mathbb{R}^{196 \times 2048}$, serving as the encoder output, encoding contextual transformation information between the paired images. This fused representation captures transformation cues between the two visual states.

3.3 Decoder: Attention-Based Recurrent Generator

The decoder maps fused visual features into a structured rule sequence using recurrent decoding with attention over transformation-relevant regions.

Sequence Modeling with LSTM Each word is generated sequentially conditioned on prior words and attended visual features:

$$h_t, c_t = \text{LSTMCell}\big([e_t; z_t], (h_{t-1}, c_{t-1})\big), \tag{4}$$

where $e_t \in \mathbb{R}^{d_e}$ is the token embedding at time step t, $z_t \in \mathbb{R}^{d_v}$ is the attention-weighted context vector derived from encoder features, and $[;]$ denotes concatenation.

The initial LSTM states are derived from the global average encoder feature $h_0 = W_h\overline{F} + b_h$, $c_0 = W_c'\overline{F} + b_c$, where $\overline{F} = \frac{1}{196}\sum_i F^{(i)}$. The output projection produces vocabulary logits, $\hat{y}_t = W_0 h_t + b_0, \hat{y}_t \in \mathbb{R}^{|\mathcal{V}|}$.

Additive Attention Mechanism Following Bahdanau-style additive attention, relevance scores between the hidden state of the decoder and the encoder features are computed as $e_{t,i} = w_a^T \tanh(W_e F^{(i)} + W_h h_t)$, with learned projection matrices $W_e \in \mathbb{R}^{d_a \times 2048}$, $W_h \in \mathbb{R}^{d_a \times 512}$, $w_a \in \mathbb{R}^{d_a}$, where d_a is the dimension of attention (i.e., 256). Normalized attention weights are calculated using

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_j \exp(e_{t,j})} \tag{5}$$

The context vector is generated using $z_t = \sum_i \alpha_{t,i} F^{(i)}$. The sigmoid-gated context modulates the relevance of attention $\tilde{z} = \sigma(W_\beta h_t \odot z_t)$, ensuring adaptive focus depending on the state of the decoder. This allows the decoder to dynamically choose how much visual information is needed at each generation step.

Training Strategy At decoding step t, the LSTM hidden state $h_t \in \mathbb{R}^{d_h}$ is projected to vocabulary logits $o_t = W_0 h_t + b_0$, where $W_0 \in \mathbb{R}^{|\mathcal{V}| \times d_h}$. The next-token probability is computed using softmax:

$$P(w_t = i | w_{1:t-1}, I_A, I_B) = \frac{\exp\left(o_{t,i}\right)}{\sum_{j=1}^{|\mathcal{V}|} \exp\left(o_{t,j}\right)} \tag{6}$$

The model generates words sequentially until it produces the end-of-sequence token $< eos >$, which marks the completion of the output.

Training the model involves minimizing the negative log-likelihood of the correct token sequence given the image pair. Joint optimization of both encoder and decoder ensures that the learned representation aligns visual transformations with textual semantics. The training objective is defined as

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T} \log P(w_t^* | w_{1:t-1}, I_A, I_B) \tag{7}$$

The model is trained using backpropagation through time with Adam at learning rate 10^{-3}. Variable-length sequences are handled using `pack_padded_sequence`, with optional gradient clipping for stability.

3.4 Probabilistic Inference

During inference, beam search maintains multiple partial hypotheses and selects the completed sequence with the highest accumulated log likelihood. For beam width k,

$$S_t^{(k)} = \arg\max_{S_{1:t-1}^{(k)}} \sum_{\tau=1}^{t} \log P(w_\tau \mid S_{1:\tau-1}, I_A, I_B) \tag{8}$$

The sequence with the highest accumulated log likelihood that ends with the token $< eos >$ is selected.

4 Proposed Metric: Transformation Inference Score

Textual metrics like BLEU or ROUGE emphasize local word order and n-gram overlaps. Jaccard and token-level F1 focus on operation accuracy; however, they ignore sequence. Exact-Match is precise but inflexible, offering no credit for nearly correct predictions. In visual analogy problems, the rules are inherently structured and sequential. Transformation Inference Score (TIS) evaluates rule-generation by jointly assessing operation correctness and execution order. It addresses cases where models output correct transformations in the wrong sequence or with extra/missing steps that conventional text metrics overlook. A high TIS indicates accurate, correctly ordered transformations; a medium TIS reflects partially correct sets with ordering or minor errors; and a low TIS signals substantial mismatches, linking symbolic accuracy with text similarity to quantify relational understanding depth.

Let $y = (y_1, \ldots, y_m)$ denote the ground-truth, and $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)$ be the predicted sequence. We compute two complementary components:

1. Token-level correctness: This metric measures whether the predicted operations match the reference set, regardless of order, and is defined as:

$$F_{\text{tok}} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}},$$

2. Order Preservation Ratio (OPR): This metric is defined as

$$\text{OPR} = \frac{\text{LCS}(\hat{y}, y)}{|y|},$$

where $\mathrm{LCS}(\hat{y}, y)$ is the length of the longest common subsequence between the predicted and reference operations, and $|y|$ is the number of true operations. The final Transformation Inference Score is then defined as a convex combination:

$$\mathrm{TIS} = \alpha F_{\mathrm{tok}} + (1 - \alpha)\mathrm{OPR},$$

where $\alpha \in [0, 1]$ balances the accuracy of operation and the preservation of order. In our experiments, we use $\alpha = 0.7$ as a task-motivated setting, giving slightly higher importance to correct operation prediction while still rewarding correct sequencing. We also perform a sensitivity analysis with $\alpha \in \{0.5, 0.6, 0.7, 0.8\}$, where TIS remains stable at 0.9843, 0.9839, 0.9835, and 0.9831, respectively, indicating that the conclusion is not sensitive to this choice.

5 Experiments and Results

5.1 Evaluation Setup

We evaluate VISAA-BD on $10,000$ rule-generation instances from DigiAnalogy-10K. DigiAnalogy-10K is split at the problem level into training, validation, and test subsets. Each problem is independently generated using randomized shape choices, transformation parameters, colors, and answer layouts; therefore, exact problem instances and visual configurations do not overlap across splits, although operation types may appear in multiple splits. The VISAA-BD configuration and evaluation protocol is defined in Table 2. The images of panels A and B are resized to 224×224 and normalized with ImageNet statistics. The generation vocabulary covers operation names, colors, and numeric parameters. We evaluate rule generation in a controlled synthetic setting. By controlled synthetic setting, we mean that visual panels and transformation descriptions are produced by the same rule generator using a fixed vocabulary of geometric operations, ensuring unambiguous image-rule alignment, avoiding annotation noise, and isolating reasoning performance from dataset bias or linguistic ambiguity.

Table 2: VISAA-BD configuration and evaluation protocol.

Component	**Setting**
Input	Resize 224×224; ImageNet normalization
Vocabulary	Derived from logs/generator
Encoder	ResNet-50 (pretrained), final conv features
Decoder	Pixel-wise attention LSTM; learned gating
Objective	Cross-entropy over token sequence
Optimizer	Adam; learning rate 1×10^{-3}; batch 32; 100 epochs
Inference	Beam search (width 3), deterministic
Metrics	BLEU, ROUGE, Jaccard, token P/R/F1, TIS

Table 3: Comparison of rule-generation performance across large language models and VISAA-BD on the DigiAnalogy-10K benchmark.

Metrics	GPT-4o	ChatGPT-4	Gemini	Claude	Ours
BLEU-1 [15]	0.48	0.44	0.50	0.44	0.99
BLEU-2 [15]	0.29	0.29	0.34	0.27	0.98
BLEU-3 [15]	0.20	0.19	0.20	0.21	0.97
BLEU-4 [15]	0.12	0.12	0.11	0.11	0.96
Jaccard Similarity [9]	0.33	0.31	0.33	0.32	0.97
Token Precision [23]	0.47	0.44	0.46	0.45	0.98
Token Recall [23]	0.64	0.53	0.60	0.53	0.98
Token F1 Score [23]	0.55	0.48	0.58	0.53	0.98
ROUGE-1 (P) [11]	0.85	0.84	0.84	0.83	0.99
ROUGE-2 (P) [11]	0.68	0.65	0.67	0.64	0.98
ROUGE-L (P) [11]	0.83	0.81	0.83	0.81	0.99
Transformation Inference Score (TIS)	0.78	0.76	0.77	0.77	0.98

5.2 Quantitative Results

Table 3 compares the predictive quality of GPT-4o[14], ChatGPT-4[13], Gemini[22], Claude[1], and the proposed VISAA-BD on the DigiAnalogy-10K benchmark using a range of established text-similarity and token-level agreement metrics. The foundation models are evaluated as zero-/few-shot reference systems without task-specific fine-tuning; therefore, the comparison is intended to provide contextual evidence rather than a like-for-like comparison with fine-tuned visual reasoning baselines. BLEU scores [15] measure the overlap of n-gram between predicted and reference rules, making them sensitive to local word order. The ROUGE metrics [11] capture sequence-level overlap, with ROUGE-1, ROUGE-2, and ROUGE-L reflecting unigram, bigram, and longest-common-subsequence matches, respectively. The Jaccard similarity [9] evaluates the set-level agreement between the predicted and the true transformation tokens. Token Precision, Recall, and F1 provide a more fine-grained view of semantic correctness by quantifying how many transformation operations were correctly included or missed, independent of order. Across all metrics, VISAA-BD shows a substantial advantage over large language models, demonstrating its stronger ability to infer, describe, and sequence visual transformation rules from visual analogy panels.

Table 4 presents the ablation study of VISAA-BD. Replacing the ResNet-50 encoder with DenseNet-121 leads to a clear drop across BLEU, ROUGE, Jaccard, token-level scores, and TIS, suggesting that the original encoder provides more suitable visual features for transformation-rule generation. Removing the attention mechanism also reduces performance, indicating that attention helps the decoder focus on transformation-relevant visual regions during rule generation. Overall, the full VISAA-BD model achieves the best performance across all reported metrics.

We evaluated TIS across the rule descriptions generated in DigiAnalogy-10K. VISAA-BD achieved an average $TIS = 0.98$ (± 0.03), demonstrating a strong

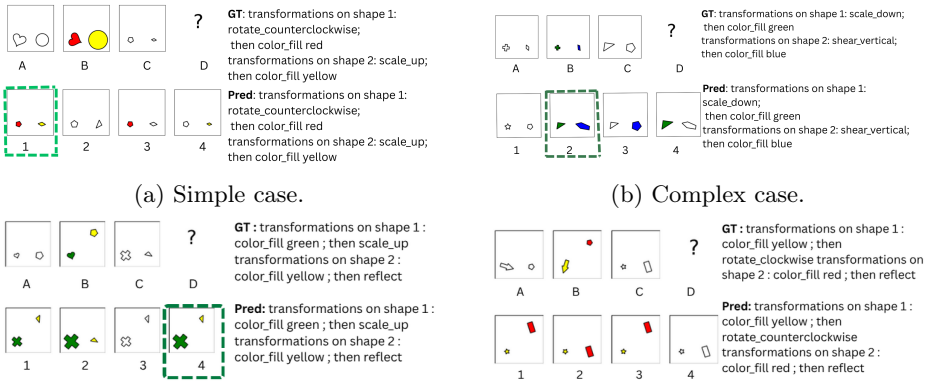
Table 4: Ablation study of VISAA-BD architectural components on DigiAnalogy-10K.

Metrics	**DenseNet-121**	**w/o Attention**	**VISAA-BD**
BLEU-1 [15]	0.91	0.96	0.99
BLEU-2 [15]	0.83	0.95	0.98
BLEU-3 [15]	0.73	0.95	0.97
BLEU-4 [15]	0.62	0.95	0.96
Jaccard Similarity [9]	0.80	0.93	0.97
Token Precision [23]	0.89	0.96	0.98
Token Recall [23]	0.88	0.96	0.98
Token F1 Score [23]	0.88	0.96	0.98
ROUGE-1 (P) [11]	0.91	0.97	0.99
ROUGE-2 (P) [11]	0.83	0.95	0.98
ROUGE-L (P) [11]	0.90	0.97	0.99
Transformation Inference Score (TIS)	0.89	0.96	0.98

alignment between predicted and reference transformations. This high score indicates that VISAA-BD not only identifies the correct visual operations, but also preserves their sequential structure accurately. To contextualize the effectiveness of TIS, we further computed it for several leading large language models. As reported in Table 3, all foundation-model baselines achieved substantially lower TIS values than VISAA-BD, indicating partial understanding of transformation semantics but weaker preservation of operation ordering. These systems show partial understanding of transformation semantics but often fail to preserve operation order. In contrast, VISAA-BD achieves a much higher TIS, demonstrating superior inference of both the correct operations and their sequence.

5.3 Qualitative Examples

Qualitative analysis provides crucial insights into the model's capabilities and limitations, complementing quantitative metrics. As illustrated in a–c of Fig. 4, VISAA-BD successfully generates fluent and accurate descriptions for a diverse range of transformations, including multi-step sequences involving novel shapes such as hearts and circles. It demonstrates a robust understanding of spatial operations such as shearing and rotation, accurately parsing the visual logic of the image pairs. However, a deeper examination of failure cases, categorized in d, reveals specific challenges. A common error mode involves the misordering of logically sequential transformations, where the model lists the correct operations but in an invalid execution sequence. Other failures include the omission of subtle transformations and the occasional lexical imprecision in attribute description. These findings are significant because they expose limitations that are largely overlooked by conventional text-generation metrics such as BLEU, which prioritize n-gram overlap over logical coherence. Consequently, this analysis directly validates the necessity of our proposed Transformation Inference Score



(a) Simple case.

(b) Complex case.

(c) Generalizations.

(d) Incorrect operation.

Fig. 4: Qualitative results. (a)–(c) demonstrate successful rule generation on challenging problems from DigiAnalogy-10K. (d) shows a typical failure mode where transformation operations are correct but their sequence is wrong.

(TIS), which is explicitly designed to penalize such logical and structural errors, providing a more aligned measure of a model's true relational understanding.

5.4 Robustness Study

To examine robustness, we added Gaussian and Salt–Pepper noise to panels A and B. As shown in Fig. 5, VISAA-BD remains stable under Gaussian noise, achieving BLEU-4 of 0.78, token F1 of 0.90, and TIS of 0.90. Salt–Pepper noise is more disruptive because it corrupts shape boundaries, reducing BLEU-4 to 0.54 and TIS to 0.83. Overall, the results show that VISAA-BD can retain

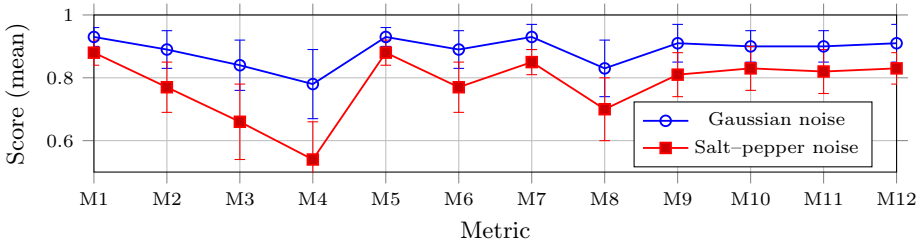


Fig. 5: Effect of Gaussian and salt–pepper noise on rule prediction performance across metrics M1–M12 (mean ± std). Metrics are abbreviated as: M1 = BLEU-1 [15], M2 = BLEU-2, M3 = BLEU-3, M4 = BLEU-4, M5 = ROUGE-1 (P) [11], M6 = ROUGE-2 (P), M7 = ROUGE-L (P), M8 = Jaccard Similarity [9], M9 = Token Precision, M10 = Token Recall, M11 = Token F1 Score [4], M12 = Transformation Inference Score (TIS).

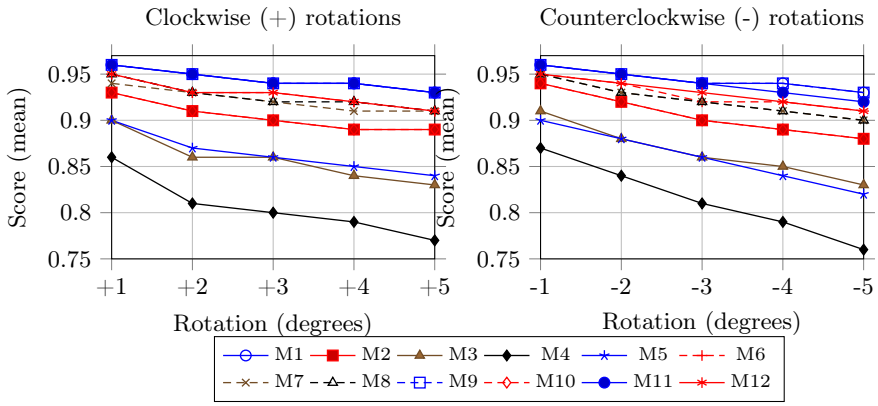
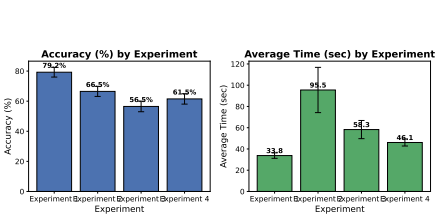


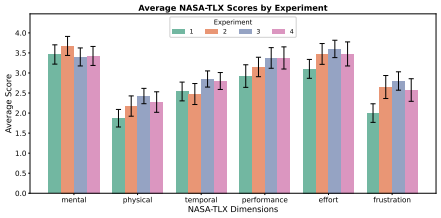
Fig. 6: Robustness of VISAA-BD under small rotations. Left: clockwise rotations. Right: counterclockwise rotations. Curves show mean performance per metric.

transformation-rule inference under moderate visual degradation, although high-frequency noise remains challenging.

To assess rotational robustness, we evaluated VISAA-BD on ten rotated variants of DigiAnalogy-10K, obtained by rotating both panels A and B by $\pm 1^\circ, \ldots, \pm 5^\circ$ while keeping the rule text unchanged. Fig. 6 reports the results. The metrics include BLEU-based n-gram overlap (M1–M4), ROUGE-based sequence overlap (M5–M7), Jaccard similarity (M8), token precision/recall/F1 (M9–M11), and TIS (M12). In all settings, BLEU remains above approximately 0.75, most ROUGE scores remain high, and TIS stays between 0.91 and 0.95. This indicates that VISAA-BD continues to infer correct operations and their order under perturbed panel alignment. Performance decreases gradually with rotation magnitude, but the degradation is symmetric for clockwise and counterclockwise directions, suggesting robustness.



(a) Accuracy and completion time.



(b) Average NASA-TLX subscale scores.

Fig. 7: Performance and workload measures across four experiments.

5.5 User Study

A controlled mixed-methods study examined the effect of AI-generated hints on diagrammatic reasoning using problems from DigiAnalogy-10K. Sixty university participants (mean age 24) completed four online experiments with increasing assistance levels (none to three hints), while accuracy, time-on-task, and workload were recorded. After onboarding with demographics, consent, and AI-exposure information, participants completed a NASA-TLX workload questionnaire and a short debrief. Optional interviews with 20 participants provided qualitative insights into hint usefulness, cognitive load, reasoning strategies, and speed–accuracy trade-offs under IRB-approved protocol 20250901EXP. Figure 7a shows that hints affect both accuracy and completion time, and participants complete later assisted tasks faster after adaptation. However, this improvement in speed may introduce a speed–accuracy trade-off. Figure 7b shows that physical demand remains low, while mental demand and effort reflect active engagement with tasks. Overall, the results suggest that calibrated hints can support efficient reasoning, but should be designed carefully to avoid unnecessary cognitive load.

6 Conclusion and Future Work

This paper introduces a framework for interpretable visual analogical reasoning in proportional analogy problems with the DigiAnalogy-10K dataset, expanding shape diversity and complex transformations. It includes structured annotations for rule-based and data-driven research. An attention-based encoder-decoder model, using ResNet-50 and pixel-wise attention LSTM, links visual cues to rule generation, showing strong performance in various metrics. The new Transformation Inference Score (TIS) assesses the correctness and order of transformation. The user study highlights the effectiveness of AI-generated hints. Future work will extend DigiAnalogy-10K to scanned and mixed-quality documents under realistic acquisition noise and explore large language and multimodal foundation models for improved rule generation, explanation generation, and human-interpretable visual reasoning.

References

1. Anthropic: Claude 3 model card. `https://www.anthropic.com` (2024), accessed: 2025-11-11
2. Barrett, D.G.T., Hill, F., Santoro, A., Morcos, A.S., Lillicrap, T.: Measuring abstract reasoning in neural networks. In: ICML (2018)
3. Burie, J., Fornés, A., Santosh, K.C., Luqman, M.M.: Deep learning for graphics recognition: document understanding and beyond. IJDAR **24**, 1–2 (2021)
4. Chinchor, N.: Muc-4 evaluation metrics. In: Message Understanding Conference. Morgan Kaufmann Publishers (1992)
5. Davila, K., Kota, B.U., Setlur, S., Govindaraju, V., Tensmeyer, C., Shekhar, S., Chaudhry, R.: Icdar 2019 competition on harvesting raw tables from infographics (chart-infographics). In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1594–1599. IEEE (2019)

6. Evans, R., Grefenstette, E.: Learning explanatory rules from noisy data. In: NeurIPS (2018)
7. Futrelle, R.P.: Strategies for diagram understanding: Object/spatial data structures, animate vision, and generalized equivalence. In: ICPR (1990)
8. Gentner, D.: Structure-mapping: A theoretical framework for analogy. Cognitive Science **7**(2), 155–170 (1983)
9. Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R., Vanhoutte, A.: Similarity measures in scientometric research: The jaccard index versus salton's cosine formula. Information Processing & Management **25**(3), 315–318 (1989)
10. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015)
11. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Proceedings of the ACL Workshop on Text Summarization Branches Out. pp. 74–81 (2004)
12. Lovett, A., Forbus, K., Usher, J.: Analogy with qualitative spatial representations can simulate solving raven's progressive matrices. In: Proceedings of the Annual Meeting of the Cognitive Science Society (2007)
13. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
14. OpenAI: Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/` (2024), accessed: 2025-11-11
15. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
16. Pedone, R., Hummel, J.E., Holyoak, K.J.: The use of diagrams in analogical problem solving. Memory & Cognition **29**(2), 214–221 (2001)
17. Raven, J.C.: Standardization of progressive matrices, 1938. British Journal of Medical Psychology **19**(1), 137–150 (1938)
18. Sahu, P., Basioti, K., Pavlovic, V.: Savir-t: Spatially attentive visual reasoning with transformers. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 460–476. Springer (2022)
19. Santoro, A., Raposo, D., Barrett, D.G.T., Malinowski, M., Pascanu, R., Battaglia, P.W., Lillicrap, T.: A simple neural network module for relational reasoning. In: NeurIPS (2017)
20. Sekh, A.A., Dogra, D.P., Kar, S., Roy, P.P., Prasad, D.K.: Can we automate diagrammatic reasoning? Pattern Recognition **106**, 107412 (2020)
21. Singh, A.K., Khanna, S., Chattopadhyay, C.: Advancing abstract reasoning for rpms with a path aggregation network and deep predictive reasoning. Machine Vision and Applications **37**(1), 3 (2026)
22. Team, G.D.: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2024)
23. Van Rijsbergen, C.J.: Information Retrieval. Butterworths, London (1979)
24. Zhang, C., Gao, F., Jia, B., Zhu, Y., Zhu, S.C.: Raven: A dataset for relational and analogical visual reasoning. In: CVPR (2019)
25. Zhang, C., Jia, B., Gao, F., Zhu, Y., Lu, H., Zhu, S.C.: Learning perceptual inference by contrasting. NeurIPS **32** (2019)
26. Zhao, S., You, H., Zhang, R.Y., Si, B., Zhen, Z., Wan, X., Wang, D.H.: An interpretable neuro-symbolic model for raven's progressive matrices reasoning. Cognitive Computation **15**(5), 1703–1724 (2023)