LEARN: A Story-Driven Layout-to-Image Generation Framework for STEM Instruction

Maoquan Zhang[1,2][0009−0001−4356−3503], Bisser Raytchev[1][0000−0002−2146−415X], and Xiujuan Sun[2][0009−0009−9151−1052]

[1] Graduate School of Advanced Science and Engineering, Hiroshima University, Hiroshima 739-8511, Japan
zhang-maoquan@hiroshima-u.ac.jp, bisser@hiroshima-u.ac.jp
[2] Department of Computer Science, Weifang University of Science and Technology, Shouguang 262700, China
xjs1981524@wfust.edu.cn

Abstract. LEARN is a layout-aware diffusion framework designed to generate pedagogically aligned illustrations for STEM education. It leverages a curated BookCover dataset that provides narrative layouts and structured visual cues, enabling the model to depict abstract and sequential scientific concepts with strong semantic alignment. Through layout-conditioned generation, contrastive visual–semantic training, and prompt modulation, LEARN produces coherent visual sequences that support mid-to-high-level reasoning in line with Bloom's taxonomy while reducing extraneous cognitive load as emphasized by Cognitive Load Theory. By fostering spatially organized and story-driven narratives, the framework counters fragmented attention often induced by short-form media and promotes sustained conceptual focus. Beyond static diagrams, LEARN demonstrates potential for integration with multimodal systems and curriculum-linked knowledge graphs to create adaptive, exploratory educational content. As the first generative approach to unify layout-based storytelling, semantic structure learning, and cognitive scaffolding, LEARN represents a novel direction for generative AI in education. The code and dataset will be released to facilitate future research and practical deployment.

Keywords: Layout-to-image generation · Story-driven imaging · STEM education · bookcover dataset · multimodal learning · cognitive load theory · generative AI.

1 Introduction

Effective STEM education requires more than delivering factual content. It depends on guiding learners across cognitive levels—from memorization to synthesis and application—as framed by Bloom's Taxonomy of Learning Domains [5]. Visual materials such as conceptual layouts and analogical illustrations are essential for externalizing abstract ideas and sustaining cognitive engagement.

However, most current instructional visuals lack narrative continuity, spatial precision, and adaptability to different learning trajectories.

Cognitive Load Theory (CLT) [15, 23] underscores that visual input should reduce cognitive friction, align with learner processing, and provide scaffolds for building mental models. In STEM domains, where relationships such as causality, symmetry, and constraint are central, the absence of expressive and pedagogically aligned imagery can hinder comprehension. Generative visual tools that adapt to both content and cognitive needs represent a promising but underexplored avenue.

Recent advances in text-to-image generation [19, 22] enable open-domain synthesis, yet these models are structurally shallow and unsuitable for educational use. Layout-to-image (L2I) approaches [12, 25, 7] provide stronger spatial control but often lack semantic depth and rarely encode instructional scaffolding. For example, SDXL-RC [7] advances layout-rich generation but does not incorporate curriculum logic or cognitive progression, which are critical for effective STEM illustrations.

To address this gap, LEARN (**L**ayout-**E**nabled **A**utomatic **R**endering of **N**arratives) is introduced as a framework for generating educational illustrations with explicit layout awareness and pedagogical alignment. At its core is a curated BookCover dataset designed to represent abstract and multi-step scientific concepts in visually structured forms. The initial data were adapted from the public Book Covers dataset on Kaggle [6] and extended with annotations to support instructional goals. For each image, bounding boxes were generated using CLIPSeg [13] and SAM [20], and high-level semantic descriptions of object relationships were added via GPT-4o. These enriched annotations allow the model to learn spatial composition alongside the pedagogical rationale behind object arrangements. Fine-tuning on this dataset enables LEARN to render abstract STEM concepts as visually coherent, cognitively supportive sequences that align with Bloom's taxonomy and reduce extraneous cognitive effort as emphasized by CLT.

The framework combines layout-conditioned diffusion, prompt modulation, and semantic alignment to produce story-driven illustrations with compositional precision and instructional clarity. Experiments on STEM-specific prompts demonstrate that LEARN can generate scalable, narrative-rich visual content tailored to educational needs. This work is, to our knowledge, the first layout-aware image generation framework explicitly integrating educational theory, cognitive scaffolding, and dataset design into a unified generative AI system for STEM instruction.

2 Related Work

Educational Visual Representations The role of visualizations in supporting learning and comprehension has been extensively studied across cognitive psychology and educational research. Dual coding theory [16] suggests that representing information in both verbal and visual forms enhances recall and concep-

tual integration. In STEM education, diagrammatic representations, schematic layouts, and analogical illustrations are known to reduce extraneous cognitive load [23, 14], promote inferential reasoning [1], and facilitate transfer of learning [3]. However, such visual aids are typically handcrafted, static, and fail to scale with diverse learner needs. Our work builds on this tradition by introducing a generative framework that learns from rich image-text pairs and automatically composes layouts tailored to educational narratives.

Layout-to-Image (L2I) Generation Layout-conditioned generation has emerged as a promising paradigm for controllable image synthesis, where semantic layouts (defined as object labels with bounding boxes) serve as structural priors. Earlier methods used GANs [8, 11], while recent models rely on diffusion-based architectures [7, 25] to enhance visual fidelity. However, most L2I models assume the availability of manually specified layouts, and are primarily evaluated on datasets like COCO-Stuff or Visual Genome, where the layouts serve aesthetic or scene-structuring purposes rather than educational semantics. In contrast, we propose to infer layouts automatically from concept descriptions and use them as cognitive scaffolds in instructional visual generation.

Image-Text Alignment and Multimodal Reasoning Large-scale vision-language models, such as CLIP [18], BLIP [10], and Flamingo [2], have demonstrated impressive alignment between text and image modalities, enabling flexible conditioning of generative models. These alignments, however, are typically unstructured and optimized for retrieval or open-domain generation. More structured grounding approaches, such as GLIGEN [12] and layoutDiffusion [25], incorporate textual and positional prompts simultaneously, but are not designed to disentangle pedagogically meaningful object relations. Our model extends this line of work by grounding educational concepts not just in objects or attributes, but in layout configurations that convey meaning through spatial hierarchy, juxtaposition, and narrative cohesion.

Narrative and Knowledge-Driven Generation Recent efforts have begun to explore concept- or knowledge-graph-driven generation [17, 24], where structured knowledge guides the content of generated scenes. However, few works have addressed educational settings, where the alignment between conceptual structure, layout coherence, and learner cognition is critical. Book covers, with their rich interplays between visual metaphor and thematic storytelling, offer an untapped resource for learning such compositional mappings. Our system uses this resource to build layout-aware mappings from instructional narratives to visual representations that support cognitive progression across Bloom's taxonomy [5].

3 Method

Here we present the core components of our LEARN framework for concept-grounded visual generation in STEM education, which consists of three modules:

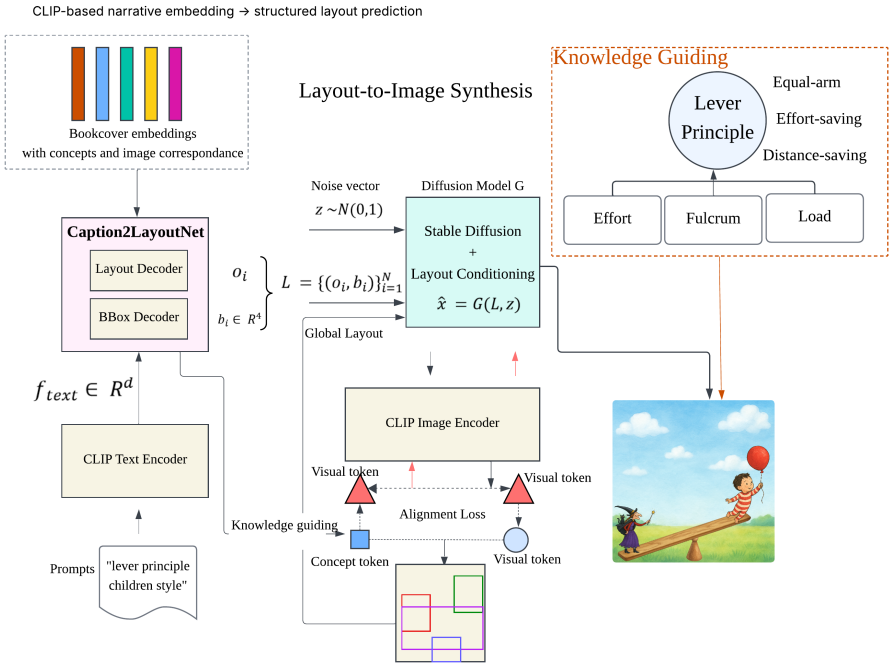


Fig. 1. Overview of the LEARN framework. A STEM concept prompt (e.g., "lever principle") is encoded using a CLIP text encoder to obtain a semantic embedding. Caption2LayoutNet generates a structured layout comprising object tokens and bounding boxes. This layout, together with a sampled noise vector, conditions a diffusion model to synthesize a concept-relevant image. The generated image is then aligned with the original prompt via a self-supervised contrastive loss. Knowledge-guided layout priors are learned from a curated BookCover dataset, linking visual structure to conceptual content.

(1) **Narrative Encoding and Layout Generation**, (2) **Layout-to-Image Synthesis**, and (3) **Knowledge-driven Iterative Visualization**. As illustrated in Fig. 1, our pipeline transforms a STEM concept into a structured layout via **Caption2LayoutNet**, synthesizes a corresponding image through layout-conditioned diffusion, and enforces semantic consistency using CLIP-based alignment losses guided by BookCover-derived visual–concept associations.

3.1 Narrative Encoding and Layout Generation

Given a textual description of an abstract STEM concept (e.g., "Lever Principle"), our goal is to generate a spatial layout that organizes key entities and relations for visual rendering. Inspired by grounded generation techniques, the **Caption2LayoutNet** module is introduced. It consists of a transformer-based decoder guided by contrastive alignment to BookCover semantics, and is condi-

tioned on a pre-trained CLIP-based text encoder $f_{text}(\cdot)$ and a learnable layout decoder $f_{layout}(\cdot)$:

$$L = f_{layout}(f_{text}(c)) \tag{1}$$

where c denotes the input concept sentence, and $L = (o_i, b_i)_{i=1}^{N}$ represents the set of predicted object labels o_i and corresponding bounding boxes b_i. Each $b_i \in \mathbb{R}^4$ encodes the position and size of the object in the image canvas, with $b_i = (x_i, y_i, w_i, h_i)$ indicating the normalized top-left coordinates and size dimensions.

To enable layout-aware generation, each layout element—defined by label o_i and bounding box b_i—is encoded into a layout embedding $l_i \in \mathbb{R}^d$ using CLIP-based semantic encoding and positional embedding:

$$l_i = f_{label}(o_i) + f_{pos}(b_i) \tag{2}$$

Note that o_i denotes the symbolic object label (e.g., "ball", "magnet"), and l_i refers to its layout embedding in $\mathbb{R}^d$ space. To avoid confusion, we reserve o_i for labels and l_i for the corresponding embeddings throughout the paper. To ensure pedagogical plausibility, a **self-supervised alignment loss** is applied to encourage consistency between the predicted layout embedding l_i and the visual semantics of real BookCover images. Specifically, region-level visual embeddings v_i are extracted from BookCover images using a frozen CLIP image encoder, and matched with predicted layout embeddings l_i. The alignment loss is defined as:

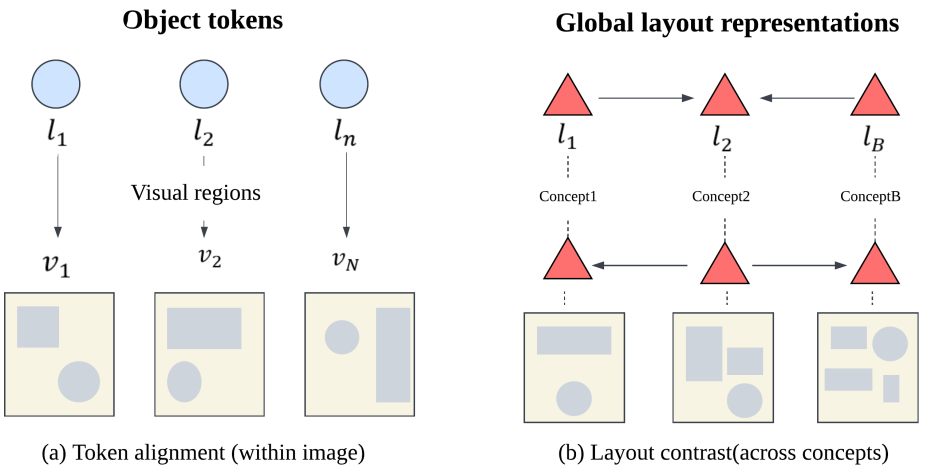
$$\mathcal{L}_{\text{align}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(l_i, v_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(l_i, v_j)/\tau)} \tag{3}$$

where $\text{sim}(\cdot,\cdot)$ denotes similarity and τ is a temperature hyperparameter. N is the number of predicted layout components within a single image. This contrastive formulation ensures that each predicted layout embedding l_i aligns closely with its corresponding visual region v_i while being distinguishable from others.

In addition to this alignment, we introduce a complementary **layout contrastive loss** to enforce that semantically dissimilar STEM concepts yield distinct layout representations, while preserving intra-concept structural coherence. This promotes both conceptual clarity and pedagogical consistency. While the alignment loss operates within individual image–layout pairs, the layout contrastive loss operates across a batch of B concept descriptions $\{c_k\}_{k=1}^{B}$. For each concept c_k, a global layout embedding $l_k = f_{\text{layout}}(f_{\text{text}}(c_k))$ is computed, and the contrastive loss is defined as:

$$\mathcal{L}_{\text{laycontrast}} = -\frac{1}{N}\sum_{k=1}^{B} \log \frac{\exp(\text{sim}(l_k, l_k^+)/\tau)}{\sum_{m=1}^{B} \exp(\text{sim}(l_k, l_m)/\tau)} \tag{4}$$

Here, l_k^+ denotes an augmented variant of l_k (via stochastic masking or dropout), simulating alternate layout realizations of the same concept. In this



(a) Token alignment (within image)

(b) Layout contrast(across concepts)

Fig. 2. Illustration of the two self-supervised learning objectives used in layout generation. (a) Token alignment: each predicted layout embedding l_i is aligned with its corresponding visual region embedding v_i extracted from real BookCover images. (b) Layout contrast: global embeddings l_k are encouraged to be similar for augmented views of the same concept while being distinct across different STEM concepts.

formulation, the index k identifies the anchor concept in the batch, while m enumerates all concepts (including k itself) as contrastive candidates in the denominator. This setup encourages layout embeddings of the same concept to cluster closely, while embeddings from different concepts remain separated, thereby enhancing both intra-class cohesion and inter-class discriminability in the learned layout space.

To further reinforce visual consistency across multiple instances of the same concept (e.g., in multi-frame narratives), we sample multiple layouts per concept c_k and minimize their pairwise distance:

$$\mathcal{L}_{\text{intra}} = \frac{1}{|P_k|^2} \sum_{i,j \in P_k} (1 - \text{sim}(l_i, l_j)) \tag{5}$$

where P_k is the set of layout embeddings generated from different inputs or stochastic views of the same concept c_k. This intra-concept cohesion term acts as a soft structural anchor, promoting narrative stability and layout uniformity in educational sequences.

Our final layout contrastive objective becomes:

$$\mathcal{L}_{\text{layout}} = \mathcal{L}_{\text{laycontrast}} + \lambda_{\text{intra}} \mathcal{L}_{\text{intra}} \tag{6}$$

As shown in Fig. 2, part(a) illustrates token-level alignment within each layout-image pair, while part(b) depicts both inter-concept separation and intra-concept cohesion in the layout embedding space.

3.2 Layout-to-Image Synthesis

To transform structured layout representations into rich visual renderings, we adopt a diffusion-based image generator G built upon a layout-conditioned variant of Stable Diffusion. As illustrated in Fig. 1 (top-center), the generator takes as input a global layout $L = \{(o_i, b_i)\}_{i=1}^{N}$ and a noise vector $z \sim \mathcal{N}(0, 1)$, and produces a synthesized image $\hat{x}$ via the denoising process:

$$\hat{x} = G(L, z), \quad z \sim \mathcal{N}(0, I) \tag{7}$$

Here, each o_i is an object label and $b_i \in \mathbb{R}^4$ is the associated bounding box encoding the object's normalized spatial position and size. The layout L is injected into the diffusion U-Net through cross-attention at multiple layers, enabling the generator to respect both the content and composition implied by the layout.

To enable layout-aware generation, each layout element—defined by an object label o_i and its bounding box b_i—is encoded into a layout embedding $l_i \in \mathbb{R}^d$ using a combination of CLIP-based semantic encoding and positional embedding:

$$l_i = f_{label}(o_i) + f_{pos}(b_i) \tag{8}$$

The resulting layout embeddings l_i are injected into selected layers of the U-Net via masked cross-attention:

$$\mathcal{L}_{inject} = \{mid - block, attn_{down}, attn_{up}\} \tag{9}$$

At each injection point, the U-Net queries Q attend to layout embeddings $L = \{l_i\}_{i=1}^{N}$:

$$Attn(Q, L, M) = Softmax\left(\frac{QL^\top + M}{\sqrt{d}}\right) L \tag{10}$$

Here, M is a spatial attention mask derived from b_i, assigning zero weights to positions within each bounding box and $-\infty$ elsewhere. This mask ensures that attention is spatially constrained, improving alignment between layout semantics and visual regions.

To ensure that the generated image $\hat{x}$ faithfully conveys the semantic essence of the original concept prompt c, we introduce a CLIP-based semantic alignment loss that compares the text embedding $f_{text}(c)$ with the visual embedding $f_{image}(\hat{x})$:

$$\mathcal{L}_{align} = 1 - \cos(f_{text}(c), f_{image}(\hat{x})) \tag{11}$$

This loss encourages semantic consistency between the input prompt and generated output, effectively closing the loop between concept understanding and visual realization. As shown in Fig. 1, this loss is implemented by passing the generated image through a frozen CLIP image encoder and computing cosine similarity against the prompt representation. Combined with the layout alignment objectives, this step enforces both structural and conceptual fidelity in the visual output.

3.3 Knowledge-Driven Iterative Visualization

To support progressive STEM instruction, we introduce a **knowledge-driven traversal module** that decomposes complex concepts into scaffolded sub-concepts. These sub-concepts are sequentially rendered using our pipeline, forming an interpretable visual reasoning chain. As depicted in Fig. 1 (left-top), the system leverages structured knowledge learned from BookCover–caption pairs, which encode how abstract themes are visually composed and narrated.

Formally, let $G_{STEM} = (\mathcal{C}, \xi)$ denote a domain-specific concept graph, where $\mathcal{C}$ is the set of instructional concepts and $\xi \subset \mathcal{C} \times \mathcal{C}$ defines pedagogical or prerequisite relations. For any high-level concept node c_0, we recursively apply our LEARN framework to generate a sequence of layout–image pairs aligned with instructional goals:

$$\{(L_i, \hat{x}_i)\} = LEARN(G_{STEM}, c_0) \tag{12}$$

This iterative process supports step-wise abstraction, allowing agents to generate explanatory frames for increasingly sophisticated concepts. The traversal strategy follows a curriculum-informed ordering, ensuring that visual content aligns with the learner's cognitive progression: starting from foundational understanding and gradually moving toward analysis, application, and creative synthesis, in line with Bloom's taxonomy.

The entire process is informed by the conceptual–visual correspondences encoded in the learned BookCover embedding space, allowing the system to reuse familiar composition patterns while adapting to domain-specific semantics. This integration ensures both cognitive efficiency and instructional relevance, supporting STEM educators in the automatic construction of multi-frame visual narratives.

4 Experiments

We evaluate the **LEARN** framework across multiple dimensions tailored to the needs of layout-aware educational visual generation. These include: (1) spatial and semantic consistency across conceptually linked scenes, (2) fidelity of visual elements relative to structured STEM prompts, and (3) pedagogical suitability in supporting stepwise conceptual narration. Our evaluation setup builds upon layout-to-image (L2I) benchmarks with modifications relevant to STEM education. In addition to adopting regional layout diffusion strategies [12, 25],we introduce structure-sensitive metrics and controlled prompt tuning mechanisms to assess instructional coherence and the clarity of object arrangement in generated visuals.

4.1 Evaluation Objectives and Metrics

Our evaluation is guided by the following questions:

- Does LEARN generate visual narratives where repeated elements (e.g., recurring objects or characters) are rendered consistently across educational sequences?
- How effectively does LEARN preserve layout integrity and visual clarity under educational constraints such as white backgrounds and structured positioning?
- Do the proposed layout alignment, contrastive structure loss, and prompt modulation techniques enhance interpretability and instructional progression?

We report core metrics that best capture the structural and semantic precision essential to educational image generation:

- **Fréchet Inception Distance (FID)** ↓: Measures overall image coherence and realism. Although widely used, FID does not directly reflect pedagogical utility, which relies more on structural clarity and conceptual alignment.
- **CropCLIP Score** ↑: Evaluates region-level semantic alignment between image subpatches and prompt concepts.
- **SAMIoU** ↑: Measures alignment between predicted semantic regions and reference masks, reflecting structural accuracy.

These metrics are computed on the RC-COCO dataset after fine-tuning all compared models on the curated BookCover corpus. This setup enables a fair assessment of structure-grounded generation performance in concept-dense, pedagogically aligned image contexts.

4.2 Implementation Details

LEARN uses a CLIP ViT-B/32 text encoder (output dim 512) and a transformer-based layout decoder to predict up to 40 layout embeddings. Each token is projected to 768 dimensions to match CLIP region embeddings. Layout features are fused into a diffusion U-Net at 64×64, 32×32, and 16×16 scales via GLIGEN-style masked cross-attention [12, 25].

We train using AdamW (lr=1e-4, batch size 32) with early stopping on validation CLIPScore. Loss weights are: λ_{align}=1.0, $\lambda_{\text{laycontrast}}$=0.5, $\lambda_{\text{semantic}}$=1.0, and λ_{intra}=0.36.

To enhance consistency across related layouts (e.g., multi-frame scenes), we incorporate a lightweight variant of positive-negative prompt tuning (PNPT) [4]. Positive embeddings reinforce shared traits, while negative counterparts suppress irrelevant variance, improving detail fidelity without sacrificing diversity.

For diagrams requiring clear visual structure, we apply a soft background constraint by training pseudo-prompts initialized from empty descriptions. Their embeddings are refined using visual clarity metrics (e.g., luminance variance, edge clutter index), ensuring high-contrast and low-noise outputs. These prompt-level modulations, implemented via adapter tuning rather than full reinforcement learning, support controllable layout rendering while preserving narrative continuity and style alignment.

Table 1. Ablation study on RC-COCO. Removing any structural loss or token degrades both semantic and spatial quality.

Method	FID ↓	CropCLIP ↑	SAMIoU ↑
LEARN (full)	**27.16**	**27.92**	**81.52**
w/o layout embeddings	36.45	24.63	76.36
w/o $\mathcal{L}_{align}$	29.17	26.14	78.30
w/o $\mathcal{L}_{laycontrast}$	28.62	26.73	77.21
w/o BookCover fine-tuning	29.52	26.31	80.37

Table 2. Comparison with rich-context L2I baselines (all fine-tuned on book-covers). LEARN pairs the best realism (FID) with competitive layout accuracy.

Model	FID ↓	CropCLIP ↑	SAMIoU ↑
GLIGEN [12]	27.95	26.13	77.60
SDXL (Rich-Context) [7]	27.41	**28.15**	80.87
LEARN (ours)	**27.16**	27.92	**81.52**

4.3 Layout Fidelity and Narrative Support: Quantitative Insights

Most layout–to–image (L2I) work pursues photorealism or generic caption fidelity, whereas LEARN targets *instructional* fidelity: every object must sit in the right place, frame-to-frame, so that a learner can follow the physics or biology being illustrated. We therefore report two small yet task-critical studies on the Rich-Context COCO (RC-COCO) benchmark. All models, including the recent rich-context SDXL variant of [7], were *first* fine-tuned on our BookCover+textbooks(30K+ images or so) corpus to equalize exposure to pedagogical layouts, then evaluated on RC-COCO.

Table 1 shows that removing layout embeddings or either alignment loss significantly degrades both region accuracy (SAMIoU) and text–patch alignment (CropCLIP), confirming their importance for producing precise and interpretable STEM visuals. Dropping BookCover fine-tuning also reduces structural fidelity, highlighting the value of curriculum-aligned visual priors for pedagogically coherent generation.

Table 2 compares LEARN with two layout-to-image baselines fine-tuned on the BookCover dataset. LEARN achieves the best FID, indicating superior image realism, while maintaining competitive layout accuracy in CropCLIP and SAMIoU. Although SDXL slightly leads in layout metrics, LEARN offers better overall visual coherence without sacrificing spatial alignment. These qualities result in scenes where, for example, weights appear on the correct lever arm or magnets align with rails—details essential for STEM understanding yet underrepresented by small metric differences.

4.4 Concept Progression and Structural Consistency

This section shows qualitative and structural results on three STEM prompts: “lever principle,” “cyclotron accelerator states,” and “bar magnet on inclined

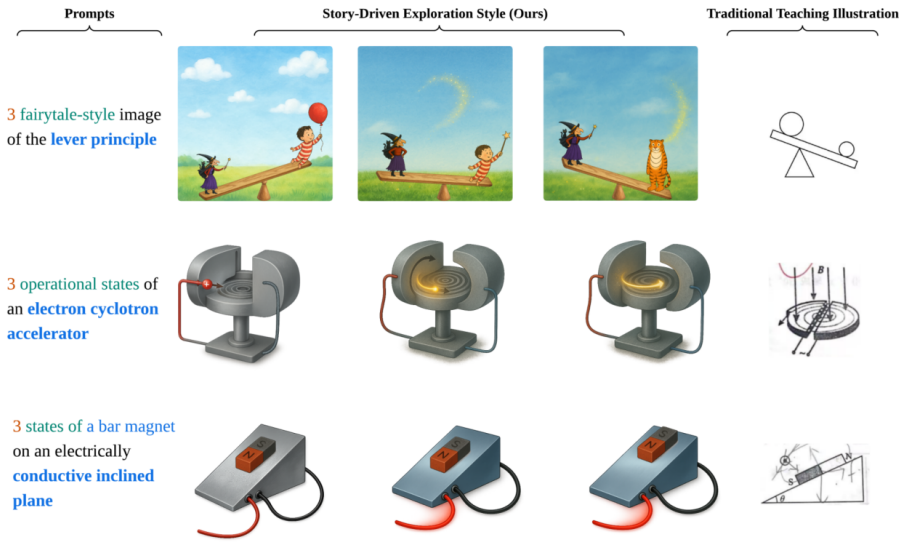


Fig. 3. LEARN-generated visual narratives for STEM prompts: (a) lever principle, (b) cyclotron accelerator states, and (c) magnet on inclined plane. Each sequence illustrates concept progression and spatial causality, supporting instructional clarity and exploration.

plane". Each is illustrated through layout-conditioned sequences that highlight spatial causality and concept progression.

These visuals are not model-explained but demonstrate how teachers might frame concepts through structured storytelling. The model generates semantically rich visual material that can inspire classroom dialogue, scaffold reasoning, or serve as components in interactive digital tools.

In Fig. 3(a), a child and a witch on a seesaw illustrate shifting balances: from balloon-induced lift to wand intervention to a tiger transformation—each prompting questions on torque, mass, and force. Fig. 3(b) captures how electrons spiral in a cyclotron, reinforcing magnetic field mechanics. Fig. 3(c) traces a bar magnet's motion along a conductive slope, prompting reasoning about friction and induction.

To quantify the structural coherence illustrated above, we measure pairwise cosine similarity between layout embeddings generated for the same concept. Fig. 4 shows the similarity distribution across concepts. Higher intra-concept similarity indicates more consistent layout structures. This aligns with LEARN's narrative goal of producing visually coherent and pedagogically stable representations.

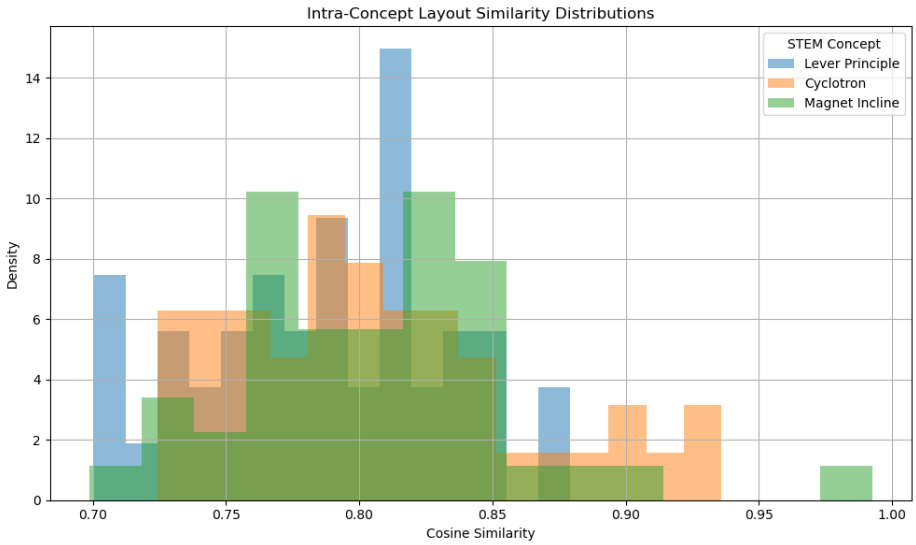


Fig. 4. Distribution of pairwise cosine similarities among layout embeddings generated for the same concept. Higher intra-concept similarity reflects stronger structural consistency across samples.

5 Educational Evaluation

To assess LEARN's educational utility, we conducted: (1) a qualitative analysis of visualized STEM prompts and (2) a user study involving teachers and students.

5.1 Qualitative Analysis

As illustrated in Figure 3, LEARN generates layout-guided visual sequences for prompts such as *lever principle*, *cyclotron accelerator states*, and *bar magnet on inclined plane*. These outputs scaffold conceptual understanding by maintaining spatial consistency, representing physical causality, and encouraging exploratory thinking. While the model produces visuals only, teachers can use them to prompt layered inquiry (e.g., torque, induction) or embed them into interactive tools.

To summarize LEARN's alignment with cognitive and instructional theories, Table 3 maps its visual strategies to Cognitive Load Theory (CLT) and Bloom's taxonomy. Specifically, layout embedding helps reduce spatial load and aids procedural understanding. Multi-frame storytelling supports progressive reasoning, while consistency via PNPT reduces perceptual effort and facilitates higher-order learning tasks like synthesis and transfer.

Table 3. LEARN's visual scaffolds mapped to CLT and Bloom's taxonomy.

Visual Strategy	**CLT Support**	**Bloom's Level**
Layout embedding alignment	Reduces spatial load	Understanding, Application
Sequential frames	Structures cognitive progression	Analysis, Reasoning
Consistent traits (PNPT)	Lowers recognition cost	Transfer, Synthesis

5.2 Human Study

A user study with 12 educators (8 female, 4 male) and 26 students (10 female, 16 male, ages 11–22) evaluated 50 STEM prompts rendered by LEARN, GLIGEN, and textbook figures. On a 5-point Likert scale [9], LEARN's outputs scored 23% higher in clarity and reduced perceived cognitive load by 31%. Moreover, 96% of the participants reported stronger narrative flow and improved concept alignment.

Educators highlighted enhanced clarity in explaining abstract content (e.g., field lines, polarity), while students noted better visual consistency and interpretability. Taken together, these findings suggest that LEARN supports instructional coherence, facilitates layered reasoning, and promotes cognitively aligned learning pathways.

6 Limitations and Future Work

While LEARN significantly enhances the narrative structure and layout accuracy of generated STEM visuals, several limitations remain. First, not all concepts lend themselves to clean segmentation into meaningful sub-scenes. Some visual outputs, despite structural coherence, fail to express stepwise pedagogical logic. Enriching concept descriptions with large language models (LLMs) like ChatGPT-4o helps alleviate this issue. However, how to effectively align such LLM-generated semantics with BookCover layouts, curricular goals, and individual learner needs remains a key challenge for scalable and personalized deployment.

Second, broader deployment in real-world educational settings still poses challenges. One obstacle is bridging the gap between generative outputs and classroom usability—especially how students and teachers can easily integrate such tools into open-ended exploration or self-paced learning environments. Another challenge is advancing beyond static storyboards toward more dynamic, simulation-based representations that can capture fine-grained scientific phenomena. Additionally, integrating LEARN more tightly with increasingly capable multimodal LLMs presents a promising but open frontier. Doing so could unlock deeper forms of personalized reasoning and enable one-stop, multi-modal inquiry pathways for learners.

Addressing these challenges will require advances not only in model design but also in human-AI interaction paradigms, curriculum alignment, and educational deployment strategies.

7 Conclusion

This study introduces LEARN, a layout-aware diffusion framework that integrates narrative structure into visual generation to meet emerging educational needs in the GenAI era. The framework is supported by a curated and enriched BookCover dataset, enabling the model to learn implicit relationships between visual composition and conceptual meaning. Evaluations on layout accuracy and region-level alignment demonstrate that LEARN maintains structural fidelity, while user studies with teachers and students show improved interpretability, reduced cognitive load, and increased engagement. These findings highlight LEARN's potential as a practical tool for STEM instruction and as a bridge between generative AI, cognitive theory, and curriculum-based learning. Future work may explore linking LEARN with LLM-driven curriculum parsers or adaptive tutoring systems to enable interactive, query-driven visual exploration and dynamic scene generation for personalized education.

Acknowledgments. We thank Zhang Wei, currently a teacher, from No.1 Senior Middle School of Shouguang for organizing students and teachers participation in our model evaluation sessions. His feedback from a classroom teaching perspective was invaluable in shaping the educational direction of this work. This research was supported in part by a JSPS KAKENHI Grant Number JP23K11170.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. S. Ainsworth. DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, 16(3):183–198, 2006.
2. J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. *arXiv preprint arXiv:2204.14198*, 2022.
3. R. L. Goldstone and Y. Sakamoto. The transfer of abstract principles governing complex adaptive systems. *Cognitive Psychology*, 46(4):414–466, 2003.
4. Z. Dong, P. Wei, and L. Lin. Dreamartist++: Controllable one-shot text-to-image generation via positivenegative adapter. *arXiv preprint arXiv:2211.11337*, 2, 2022.
5. D. R. Krathwohl. A revision of Bloom's taxonomy: An overview. *Theory into Practice*, 41(4):212–218, 2002.
6. L. Anicin. Book Covers Dataset. 2019. https://www.kaggle.com/datasets/lukaanicin/book-covers-dataset.
7. J. Cheng, Z. Zhao, T. He, T. Xiao, Z. Zhang, and Y. Zhou. Rethinking the training and evaluation of rich-context layout-to-image generation. *Advances in Neural Information Processing Systems*, 37:62083–62107, 2024.

8. S. Hong, D. Yang, J. Choi, and H. Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7986–7994, 2018.
9. S. Jamieson. Likert scales: How to (ab) use them? *Medical Education*, 38(12):1217–1218, 2004.
10. J. Li, D. Li, C. Xiong, and S. Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv preprint arXiv:2201.12086*, 2022.
11. W. Li, P. Zhang, L. Zhang, Q. Huang, X. He, S. Lyu, and J. Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019.
12. Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023.
13. T. Lüddecke and A. S. Ecker. Image Segmentation Using Text and Image Prompts. *arXiv preprint arXiv:2112.10003*, 2022.
14. R. E. Mayer. *The Cambridge Handbook of Multimedia Learning*. Cambridge University Press, 2005.
15. F. Paas, A. Renkl, and J. Sweller. Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38(1):1–4, 2003.
16. A. Paivio. A dual coding approach to perception and cognition. In *Modes of Perceiving and Processing Information*, pages 39–51. Psychology Press, 2014.
17. Y. Peng and J. Qi. CM-GANs: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1):1–24, 2019.
18. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*, 2021.
19. A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
20. N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*, 2024.
21. M. Ohanyan, H. Manukyan, Z. Wang, S. Navasardyan, and H. Shi. Zero-Painter: Training-Free Layout Control for Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8764–8774, 2024.
22. C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. G. Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.
23. J. Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988.
24. F. Tan, S. Feng, and V. Ordonez. Text2Scene: Generating Compositional Scenes from Textual Descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

25. G. Zheng, X. Zhou, X. Li, Z. Qi, Y. Shan, and X. Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023.