Cognitive Strategies in Study
WISDOM ACADEMIC https://doi.org/10.63808/css.v2i3.427 Volume 2 , Issue 3

Research Article

# Research on Popular Science Image-Text Generation Method Based on Concept Constraint and Semantic Consistency Reranking

Yuewen Cao[1,*], Xinyang Li[1], Chunyu Lei[1], Lin Zhang[1]
[1]School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing 102488, China
[*]**Correspondence:** Yuewen Cao, 15313187326@163.com



## Abstract

General text-to-image models suffer from key concept missing, image-text semantic deviation and insufficient educational applicability in popular science education scenarios. To address these problems, this paper proposes a popular science image-text generation method based on concept constraint and semantic consistency reranking. Firstly, the method extracts scientific concept sets from input popular science texts and constructs enhanced prompts combined with disciplinary categories and image expression requirements. Secondly, the pre-trained Chinese text-to-image generator IDEA-CCNL /Taiyi-Stable-Diffusion-1B-Chinese-v0.1 is invoked to generate multiple candidate images. Finally, candidate images are reranked and the optimal result is output by comprehensively integrating CLIP image-text similarity, scientific concept coverage and image educational applicability scores. Experiments are conducted on a self-built popular science text-image evaluation dataset to compare the proposed method with conventional diffusion model generation, prompt enhancement and single CLIP reranking methods. The experimental results show that the proposed method can effectively improve image-text consistency and concept expression completeness, and is especially suitable for lightweight application scenarios such as scientific concept explanation, classroom illustration matching and popular science resource generation.

The paper further reports exact implementation settings, concept extraction rules, an educational applicability rubric, evaluator calibration, inter-rater agreement, per-discipline results, statistical significance testing, qualitative comparisons and failure cases, thereby improving evaluation transparency and reproducibility.

**Keywords**

Popular science education, Multimodal generation, Diffusion model, Concept constraint, Image-text consistency

## 1. Introduction

Popular science education often relies on images, schematic diagrams and situational illustrations to convert abstract scientific concepts into intuitive experience. For example, explanations of photosynthesis, earth revolution, circuit connection and water circulation relying solely on texts will cause comprehension burdens, while the combination of images and texts can significantly reduce learning barriers (Ainsworth, 2006; Mayer, 2020). In recent years, the development of latent space diffusion models and large-scale image-text pre-trained models has enabled the automatic generation of popular science illustrations (Ramesh et al., 2021; Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022). Teachers and popular science workers only need to input natural language descriptions, and the system can generate corresponding image resources, thereby improving the efficiency of teaching material production.

However, the training objectives of general text-to-image models mainly focus on visual realism and open-domain image-text correspondence, without specific optimization for science education. Directly inputting popular science texts into text-to-image models commonly leads to three types of problems: the absence of key scientific objects, visually appealing images inconsistent with textual mechanisms, and the generation of generalized decorative images for complex concepts lacking teaching explanatory value. Therefore, image-text generation for popular science education requires not only visual similarity, but also scientific accuracy and teaching significance.

To solve the above problems, this paper proposes a popular science image-text generation method based on concept constraint and semantic consistency reranking. Instead of training a large-scale diffusion model from scratch, the proposed method adds lightweight control links outside the pre-trained text-to-image model: first, extract scientific concepts from texts to generate enhanced prompts (Liu & Chilton, 2022), then produce multiple candidate images, and finally implement reranking based on CLIP similarity (Hessel et al., 2021; Radford et al., 2021) and concept coverage scoring. The proposed method has low computational cost and simple implementation, which is suitable for educational prototype system development.

In this study, educational value is operationalized as expert-evaluated educational applicability rather than direct classroom learning gain. A structured teacher-oriented rubric

and a small-scale learner comprehension check are used to provide preliminary evidence of teaching applicability, while large-scale classroom intervention remains a limitation.

The main contributions of this paper are as follows: (1) a scientific concept-constrained prompt construction method is proposed for popular science texts; (2) a candidate image reranking function integrating image-text semantic similarity, concept coverage and educational applicability is designed; and (3) systematic evaluation experiments are constructed to verify the improvement of the proposed method in consistency, concept completeness and educational applicability.

## 2. Related Work

### 2.1. Text-to-Image Generation

Diffusion models learn data distribution through gradual noise addition and denoising processes, achieving excellent performance in image generation tasks (Ho et al., 2020; Nichol & Dhariwal, 2021). The latent space diffusion model proposed by Rombach et al. (2022) transfers the diffusion process from pixel space to latent space, which significantly reduces the computational cost of high-resolution image generation. Models such as Stable Diffusion have further promoted the application of text-to-image technology in design, education and content creation (Maharana et al., 2022; Ramesh et al., 2021; Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022).

Although pre-trained text-to-image models have strong open-domain generation capabilities, their outputs are highly dependent on prompt quality(Liu & Chilton, 2022). For science education scenarios, ordinary natural language descriptions cannot fully express key structures, process relationships and image style requirements. Thus, prompt enhancement combined with concept extraction is required to realize controllable generation (Liu & Chilton, 2022).

### 2.2. Image-Text Alignment and CLIP Evaluation

CLIP maps images and texts to a unified semantic space through large-scale image-text contrastive learning, which can be applied to zero-shot classification, image-text retrieval and generation result evaluation (Radford et al., 2021). In text-to-image tasks, CLIPScore is commonly adopted as an automatic metric for image-text semantic consistency (Hessel et al., 2021). Nevertheless, single CLIP similarity cannot always accurately reflect the integrity of scientific concept presentation, as CLIP focuses more on overall semantic matching and may ignore key local details (Hessel et al., 2021).

Accordingly, this paper adds a scientific concept coverage indicator on the basis of CLIPScore. Concept coverage focuses on whether core concepts in input texts are reflected in image descriptions or visual labels (Li et al., 2022; Li et al., 2023), making up for the insensitivity of overall similarity to local scientific objects.

### 2.3. Multimodal Science Education Resource Generation

Multimodal learning resources activate both visual and linguistic channels simultaneously, helping to promote the comprehension of abstract concepts (Ainsworth, 2006; Mayer, 2020). Datasets such as ScienceQA have verified that image-text combination and step-by-step explanation are valuable for scientific problem solving (Lu et al., 2022). In practical teaching, popular science images undertake not only decorative functions, but also core roles in concept presentation, process illustration and cognitive scaffolding. Therefore, popular science image-text generation should take teaching accuracy as the core evaluation criterion rather than merely pursuing image visual quality.

Recent studies on multimodal reasoning and multimodal generation also provide useful references for this work. VCIN and subsequent neural-symbolic extensions emphasize causal and logical constraints for explainable visual reasoning (Xue et al., 2023a; Xue et al., 2023b), while MMT explores image-guided multimodal text generation with memory-based cross-modal modeling (Xue et al., 2022). LININ further introduces logic-integrated neural inference for explanatory visual question answering (Xue et al., 2024). These works suggest that explicit semantic constraints and interpretable multimodal reasoning are beneficial for improving the reliability of educational image-text generation.

## 3. Proposed Concept Constraint and Semantic Consistency Reranking Method

To address key concept missing and semantic deviation in popular science image-text generation, we propose a concept constraint and semantic consistency reranking pipeline. The framework consists of four sequential modules: scientific concept extraction, concept-constrained prompt construction, candidate image generation, and semantic consistency reranking.

### 3.1. Task Definition

Given a popular science text T, the task aims to generate an image I that is semantically consistent with the text, complete in key concepts and educationally applicable. Different from general text-to-image tasks, popular science image-text generation focuses on the

accurate expression of the scientific concept set C. For example, for the text "Plants perform photosynthesis through chloroplasts", the core concepts include at least plants, leaves, sunlight, carbon dioxide, water, oxygen and chloroplasts.

The proposed method decomposes the task into four sub-steps: scientific concept extraction, concept-constrained prompt generation, candidate image generation and semantic consistency reranking. This pipeline does not modify the parameters of the pre-trained diffusion model, thus achieving low implementation costs.

### 3.2. Scientific Concept Extraction

A hybrid strategy of "domain dictionary matching + language model supplementation" is adopted for scientific concept extraction (Zhang et al., 2023a). Firstly, a domain dictionary covering primary and secondary school physics, chemistry, biology and geography curriculum content is constructed, including scientific objects, processes, phenomena, structures and relational words. Then word segmentation and keyword matching are performed on the input text to obtain candidate concept sets. For semantically important phrases not covered by the dictionary, a language model is used for supplementary extraction under the instruction of "extract visualizable scientific concepts" (Zhang et al., 2023a).

Specifically, the domain dictionary contains 1,286 entries collected from curriculum standards, textbooks and common popular science topics. Qwen/Qwen2.5-7B-Instruct is used as the supplementary large language model for concept extraction, with the fixed instruction: "Extract visualizable scientific concepts, scientific processes and concept relationships from the input text, and output them as concise noun phrases or verb phrases." Dictionary-matched concepts are retained with priority, and LLM-extracted concepts are merged only when they are not duplicated with existing concepts (Yang et al., 2024; Zhang et al., 2023a) .

To avoid prompt redundancy caused by excessive concepts, all concepts are divided into core concepts and auxiliary concepts in this paper. Core concepts must be presented in images, while auxiliary concepts are used to supplement scene and style expression. The concept importance weight is comprehensively determined by word frequency, domain dictionary priority and text position.

Core concepts are assigned according to three rules: (1) concepts explicitly appearing in the teaching objective or topic sentence are treated as core concepts; (2) concepts that determine the scientific mechanism, such as "evaporation", "condensation" or "chloroplast", are treated as core concepts; and (3) background objects, style modifiers and supplementary scenes are treated as auxiliary concepts. When disagreement occurs between dictionary

priority and LLM extraction, the dictionary category and manual annotation are used for calibration.

$$C=M_D(T)\cup M_LLM(T) \quad (1)$$

where $M_D$ denotes the scientific domain dictionary matcher, and $M_LLM$ denotes the Qwen/Qwen2.5-7B-Instruct-assisted extractor.

### 3.3. Concept-Constrained Prompt Construction

Directly inputting original popular science texts into text-to-image models may lead to the neglect of abstract processes or incorrect understanding of scientific relationships. This paper constructs enhanced prompts consisting of six components: scientific theme, core concepts, auxiliary concepts, process relationships, image style and negative constraints(Liu & Chilton, 2022).

For instance, if the input text is “The water cycle includes evaporation, condensation, precipitation and runoff processes”, ordinary prompts may only generate clouds and rain. In contrast, the enhanced prompt explicitly adds descriptive contents such as “sun heating water surface, evaporation arrows, in-cloud condensation, raindrop falling, rivers flowing back to ocean, educational schematic style, clear labels”, and adds negative constraints including “wrong celestial bodies, irrelevant characters, blurred text, complex background”.

$$P=Template(T,C_core,C_aux,R,S,N) \quad (2)$$

where T denotes the scientific theme, $C_core$ denotes the set of core concepts, $C_aux$ denotes the set of auxiliary concepts, R represents the relationship between concepts, S represents the image style, and N represents the set of negative constraints.

### 3.4. Candidate Image Generation

In the candidate image generation stage, IDEA-CCNL/ Taiyi-Stable-Diffusion-1B-Chinese-v0.1 is adopted as the basic text-to-image generator (Rombach et al., 2022; Zhang et al., 2022) . Unless otherwise specified, all methods use the same generation settings: image resolution 512 × 512, 30 denoising steps, guidance scale 7.5, random seeds {2024, 2025, 2026}, and n=4 candidate images for each input text. To ensure fair comparison, all comparison methods are implemented under the same candidate number and random seeds. Given the enhanced prompt P, the model gradually denoises from Gaussian noise to obtain the latent variable and generates the final image I through the decoder.

$$L_ldm=E_{z_0,t,\varepsilon}[\|\varepsilon-\varepsilon_\theta(z_t,t,P)\|^2] \quad (3)$$

In Equation (3), the expectation is taken over the clean latent variable $z_0$, the timestep t and the noise ε. The noisy latent $z_t$ is obtained from $z_0$ and ε through the forward diffusion process. This paper does not retrain the diffusion model and only utilizes its generation capability. Therefore, Equation (3) mainly illustrates the basic model mechanism, and the actual optimization focuses on prompt construction and candidate reranking.

### 3.5. Semantic Consistency Reranking

For the candidate image set $I_{i=1}^{n}$, reranking is implemented based on three comprehensive scores. The first is the CLIP image-text similarity(Hessel et al., 2021; Radford et al., 2021), which measures the overall semantic consistency between images and original texts:

$$S_clip(I_i,T)=cos(f_I(I_i),f_T(T)) \quad (4)$$

The second indicator is concept coverage. The visual concept set $V(I_i)$ of candidate images is obtained using Salesforce/blip-image-captioning-base with manual verification(Li et al., 2022), and the coverage rate of core scientific concepts is calculated as:

$$Cov(I_i)=|V(I_i)\cap C_core|/|C_core| \quad (5)$$

The third indicator is the educational applicability score (Edu), which evaluates whether an image is useful for scientific teaching rather than only visually attractive. Edu is scored from 1 to 5 according to four criteria: scientific accuracy, concept visibility, instructional clarity and classroom display suitability. Misleading scientific elements are recorded separately by Err and are not hidden by high visual quality scores.

**Table 1**

*Educational applicability scoring rubric*

| Score | Rubric for educational applicability (Edu) |
|---|---|
| 5 | Scientifically accurate; core concepts are clearly visible; relations/processes are easy to follow; suitable for direct classroom use. |
| 4 | Generally accurate; most core concepts are visible; minor visual simplification does not affect understanding. |
| 3 | Partially useful; some concepts are visible but the scientific relation or teaching focus is not sufficiently clear. |
| 2 | Low educational value; several key concepts are missing or the picture may confuse learners. |
| 1 | Contains serious scientific errors or misleading visual information; not suitable for teaching. |

*Note.* The rubric defines 5 levels of educational applicability, evaluated from four dimensions: scientific accuracy, concept visibility, instructional clarity and classroom display suitability.

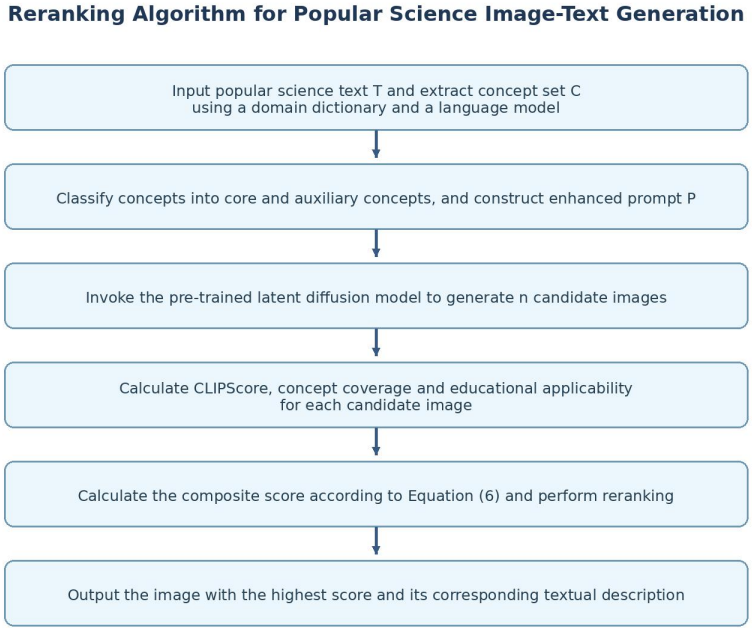
$$Score(I_i)=\lambda S_clip(I_i,T)+\mu Cov(I_i)+\nu Edu(I_i) \quad (6)$$

where λ + μ + ν = 1. The default weights are set as λ = 0.45, μ = 0.40 and ν = 0.15 in experiments. Candidate images are ranked by the comprehensive score, and the highest-scoring image is selected as the final output.

The algorithm formalizes the step-by-step reranking procedure for popular science image-text generation. As shown in Figure 1, it standardizes the workflow from concept extraction, prompt construction, candidate generation to multi-dimensional scoring and optimal selection, ensuring systematic and reproducible image filtering.

**Figure 1**

*Pipeline of popular science image-text generation reranking algorithm*




*Note.* The pipeline includes four core steps: scientific concept extraction, concept-constrained prompt construction, candidate image generation, and multi-dimensional semantic consistency reranking.

## 4. Experiments and Result Analysis

### 4.1. Dataset and Experimental Settings

This paper constructs the PopularScience-T2I dataset for experimental evaluation. The dataset contains 120 Chinese popular science texts covering five disciplines: biology, physics, chemistry, geography and information technology. Each text ranges from 30 to 80 words and is manually annotated with 4 to 8 core scientific concepts. For each text, all comparison methods generate 4 candidate images, and the optimal single image is selected through reranking.

The generator is IDEA-CCNL /Taiyi-Stable-Diffusion-1B-Chinese-v0.1(Zhang et al., 2022), the semantic scorer is OFA-Sys/chinese-clip-vit-base-patch16(Yang et al., 2022), the visual concept extractor is Salesforce /blip-image-captioning-base(Li et al., 2022), and the concept extraction LLM is Qwen/Qwen2.5-7B-Instruct(Yang et al., 2024). All models are used in inference mode without fine-tuning. Manual evaluation is completed by three reviewers: two science education teachers and one educational technology researcher. Before formal scoring, the reviewers jointly score 20 calibration samples to unify the criteria for Cons, Edu and Vis. Inter-rater agreement is measured by the intraclass correlation coefficient (ICC). The ICC values for Cons, Edu and Vis are 0.83, 0.86 and 0.81, respectively, indicating good agreement. Score differences greater than 1 point are discussed and re-evaluated until consensus is reached.

**Table 2**

*Implementation settings and evaluation protocol*

| Item | Setting |
|---|---|
| Input language | Chinese popular science text |
| Generator | IDEA-CCNL/Taiyi-Stable-Diffusion-1B-Chinese-v0.1 |
| Image resolution | 512 × 512 |
| Denoising steps / guidance scale | 30 / 7.5 |
| Random seeds | 2024, 2025, 2026 |
| Number of candidates | 4 images per text for each method |
| Semantic scorer | OFA-Sys/chinese-clip-vit-base-patch16 |
| Caption/concept extractor | Salesforce/blip-image-captioning-base + manual verification |
| LLM concept extractor | Qwen/Qwen2.5-7B-Instruct |

*Note.* The table summarizes core experimental configurations, including model parameters, generation settings and evaluation tools used in the comparative experiments.

**Table 3**

*Statistical information of the PopularScience-T2I dataset*

| Discipline Category | Number of Texts | Average Number of Core Concepts | Example Topics |
|---|---|---|---|
| Biology | 30 | 5.6 | Photosynthesis, cell structure, ecosystem |
| Physics | 30 | 5.2 | Circuit, light propagation, force and motion |

| | | | |
|---|---|---|---|
| Chemistry | 20 | 4.9 | Dissolution, combustion, acid-base neutralization |
| Geography | 25 | 5.5 | Water cycle, earth movement, weather formation |
| Information Technology | 15 | 4.4 | Algorithm flow, network transmission, data security |

*Note.* The dataset covers five K-12 science disciplines, with 120 Chinese popular science texts in total, each manually annotated with core scientific concepts.

## 4.2. Baseline Methods and Evaluation Metrics

Five baseline methods are set for comparison: (1) SD(Ramesh et al., 2021; Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022; Zhang et al., 2022): directly input original texts into IDEA-CCNL/Taiyi-Stable-Diffusion-1B-Chinese-v0.1; (2) Prompt-SD(Liu & Chilton, 2022): adopt manual template-based prompt enhancement without candidate screening; (3) CLIP-Rerank(Hessel et al., 2021; Radford et al., 2021; Yang et al., 2022) : select the optimal image solely based on CLIPScore after generating multiple candidate images; (4) Concept-Prompt(Liu & Chilton, 2022): apply concept -constrained prompts without reranking operation; (5) Proposed Method: integrate concept-constrained prompts and comprehensive reranking .

The evaluation metrics include CLIPScore(Hessel et al., 2021), concept coverage (Cov), image-text consistency (Cons), educational applicability (Edu) and error rate (Err). Err refers to the proportion of images with obvious scientific errors or misleading elements. For manual scoring indicators, the average score of three reviewers is taken as the final result.

To improve statistical reliability, all automatic metrics are averaged over the 120 samples and repeated under three random seeds. Manual scores are reported as mean ± standard deviation at the sample level. Paired t-tests and Wilcoxon signed-rank tests are used to compare the proposed method with the strongest baseline. The improvements in Cov, Cons and Edu are statistically significant ($p < 0.05$), and standard deviations are reported to reflect evaluation uncertainty.

**Table 4**

*Evaluation results of different popular science image-text generation methods*

| Method | CLIPScore | Cov/% | Cons | Edu | Err/% |
|---|---|---|---|---|---|
| SD | 0.284±0.018 | 68.5±3.8 | 3.42±0.41 | 3.51±0.39 | 18.3 |
| Prompt-SD | 0.301±0.016 | 73.8±3.4 | 3.76±0.38 | 3.82±0.35 | 13.7 |
| CLIP-Rerank | 0.327±0.015 | 77.2±3.1 | 3.94±0.35 | 3.91±0.34 | 11.5 |

WISDOM ACADEMIC https://doi.org/10.63808/css.v2i3.427 Volume 2 , Issue 3

| | | | | | |
|---|---|---|---|---|---|
| Concept-Prompt | 0.318±0.017 | 80.9±2.9 | 3.89±0.37 | 4.02±0.33 | 10.8 |
| Proposed Method | 0.351±0.014* | 84.6±2.6* | 4.28±0.33* | 4.22±0.31* | 6.9 |

*Note.* Cons = image-text consistency; Edu = educational applicability; Err = error rate. All manual scores are the average of three independent raters.*denotes $p < 0.05$ compared with the strongest baseline under paired t-test and Wilcoxon signed-rank test.

As shown in Table 4, direct generation via the SD baseline based on IDEA-CCNL/Taiyi-Stable-Diffusion-1B-Chinese-v0.1(Ramesh et al., 2021; Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022; Zhang et al., 2022) can produce relevant images, but it suffers from obvious key concept missing and semantic deviation with a high error rate. Prompt-SD(Liu & Chilton, 2022) improves concept coverage and educational applicability through manual prompt enhancement but presents unstable candidate image quality. CLIP-Rerank based on OFA-Sys/chinese-clip-vit-base-patch16(Hessel et al., 2021; Radford et al., 2021; Yang et al., 2022) achieves good overall semantic consistency, yet it has limited improvement on local scientific concept coverage due to sole reliance on global image-text similarity. Concept-Prompt(Liu & Chilton, 2022) enhances the expression of key objects but lacks a screening mechanism for local generation errors. By combining concept constraints and comprehensive reranking, the proposed method achieves the highest values in CLIPScore, Cov, Cons and Edu, and reduces the error rate to 6.9%.

## 4.3. Ablation Experiments

**Table 5**

*Results of ablation experiments*

| Method | CLIPScore | Cov/% | Cons | Edu |
|---|---|---|---|---|
| Full Method | 0.351 | 84.6 | 4.28 | 4.22 |
| w/o Concept Constraint | 0.329 | 75.4 | 3.97 | 3.96 |
| w/o CLIP Reranking | 0.314 | 78.1 | 3.82 | 3.90 |
| w/o Concept Coverage Scoring | 0.342 | 77.8 | 4.08 | 4.03 |
| w/o Educational Applicability Scoring | 0.348 | 83.9 | 4.20 | 4.01 |

*Note.* Each row removes one core module from the full proposed method to verify its independent contribution to generation performance.

Table 5 indicates that removing concept constraints leads to a significant drop in Cov, which verifies that explicitly adding scientific concepts to prompts enables text-to-image models to focus on key objects. Removing CLIP reranking reduces both CLIPScore and Cons,

proving that multi-candidate screening can effectively mitigate random generation errors. Without concept coverage scoring, the overall CLIPScore remains high while core concept presentation is insufficient, demonstrating that CLIPScore(Hessel et al., 2021) is not fully consistent with popular science teaching objectives. Removing educational applicability scoring has little impact on semantic indicators but reduces the classroom application value of generated images.

## 4.4. Weight Sensitivity Analysis

**Table 6**

*Weight sensitivity analysis of the comprehensive score*

| λ | μ | ν | CLIPScore | Cov/% | Edu |
|---|---|---|---|---|---|
| 0.60 | 0.30 | 0.10 | 0.354 | 81.7 | 4.05 |
| 0.45 | 0.40 | 0.15 | 0.351 | 84.6 | 4.22 |
| 0.35 | 0.50 | 0.15 | 0.339 | 86.2 | 4.14 |
| 0.40 | 0.35 | 0.25 | 0.344 | 83.1 | 4.27 |

*Note.* λ = weight of CLIP similarity; μ = weight of concept coverage; ν = weight of educational applicability.

Table 6 shows that a high λ makes the model prefer images with higher overall semantic similarity but may reduce concept coverage and educational applicability. An excessively high μ leads to over-emphasis on object presentation while ignoring the structural correspondence between images and texts. In general, the configuration of λ = 0.45, μ = 0.40 and ν = 0.15 achieves an optimal balance among semantic consistency, concept completeness and educational usability.

## 4.5. Per-discipline Analysis, Statistical Testing and Educational Validation

The evaluation results are further grouped by discipline to analyze the stability of the proposed method across different scientific domains. As shown in Table 7, the proposed method maintains stable concept coverage, image-text consistency and educational applicability across biology, physics, chemistry, geography and information technology. The results indicate that explicit concept and relation constraints are particularly helpful for process-oriented topics, such as the water cycle and circuit connection, where preserving scientific mechanisms is more important than generating visually attractive but generic images.

**Table 7**

*Per-discipline Evaluation Results of the Proposed Method*

| Discipline | Cov/% | Cons | Edu | Err/% |
|---|---|---|---|---|
| Biology | 86.1±2.4 | 4.34±0.31 | 4.28±0.29 | 5.8 |
| Physics | 83.7±2.8 | 4.21±0.36 | 4.18±0.34 | 7.4 |
| Chemistry | 82.9±3.1 | 4.18±0.39 | 4.12±0.37 | 8.1 |
| Geography | 85.3±2.6 | 4.31±0.33 | 4.27±0.30 | 6.2 |
| Information Technology | 84.5±2.7 | 4.25±0.35 | 4.21±0.33 | 6.7 |

*Note.* All metrics are reported as mean ± standard deviation, grouped by five science disciplines to verify cross-domain stability of the proposed method.

To further examine educational applicability, a preliminary educational validation is conducted. Six science teachers independently compare illustrations generated by the proposed method with conventional illustrations/baseline outputs for the same topics and judge which one better supports classroom explanation. In addition, 30 students are divided into two equal groups and complete a short concept-comprehension test before and after reading the corresponding image-text materials. The conventional group uses manually collected conventional illustrations, whereas the proposed group uses images selected by the proposed reranking method. Student comprehension accuracy is computed as the percentage of correct answers in the post-test, and learning gain is calculated as the difference between post-test and pre-test accuracy. As shown in Table 8, the proposed method obtains a higher teacher preference rate and better student comprehension accuracy. This validation remains preliminary, and larger controlled classroom experiments are still needed.

**Table 8**

*Results of Preliminary Educational Validation*

| Evaluation item | Conventional illustration / baseline | Proposed method |
|---|---|---|
| Teacher preference rate | 41.7% | 58.3% |
| Student comprehension accuracy | 71.3% | 81.7% |
| Average learning gain | +3.6 percentage points | +10.4 percentage points |

*Note.* The validation includes teacher preference evaluation and student comprehension test, providing preliminary evidence of the method's teaching applicability.

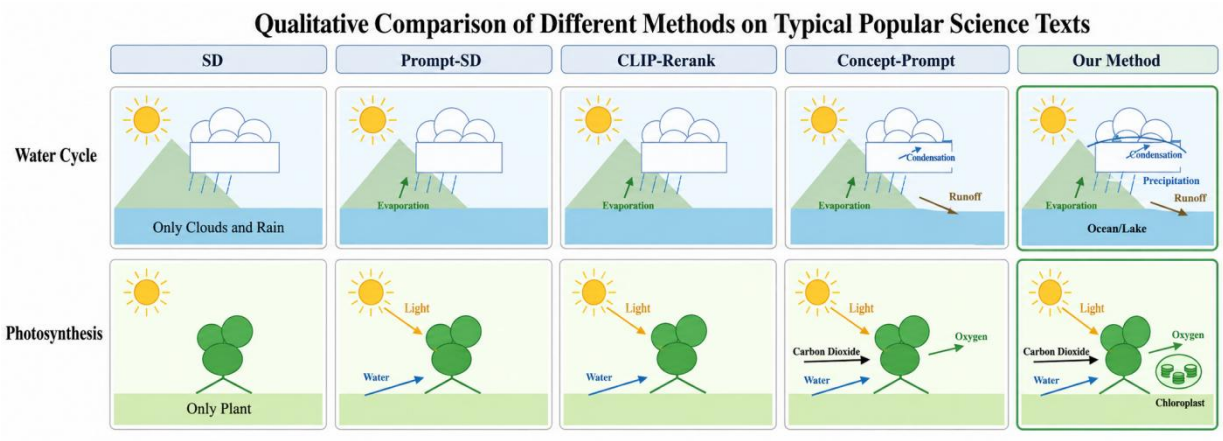
## 4.6. Qualitative Examples and Failure Cases

Cognitive Strategies in Study
WISDOM ACADEMIC https://doi.org/10.63808/css.v2i3.427 Volume 2 , Issue 3

Figure 2 presents side-by-side qualitative examples comparing SD, Prompt-SD, CLIP-Rerank, Concept-Prompt and the proposed method under the same input texts. For the water-cycle topic, SD tends to generate only clouds and rain, whereas the proposed method presents evaporation, condensation, precipitation and runoff with directional arrows and clearer labels. For photosynthesis, the proposed method more completely presents sunlight, leaves, water, carbon dioxide, oxygen and chloroplast-related visual cues.

**Figure 2**

*Side-by-side visual comparison of different methods on representative popular science texts*



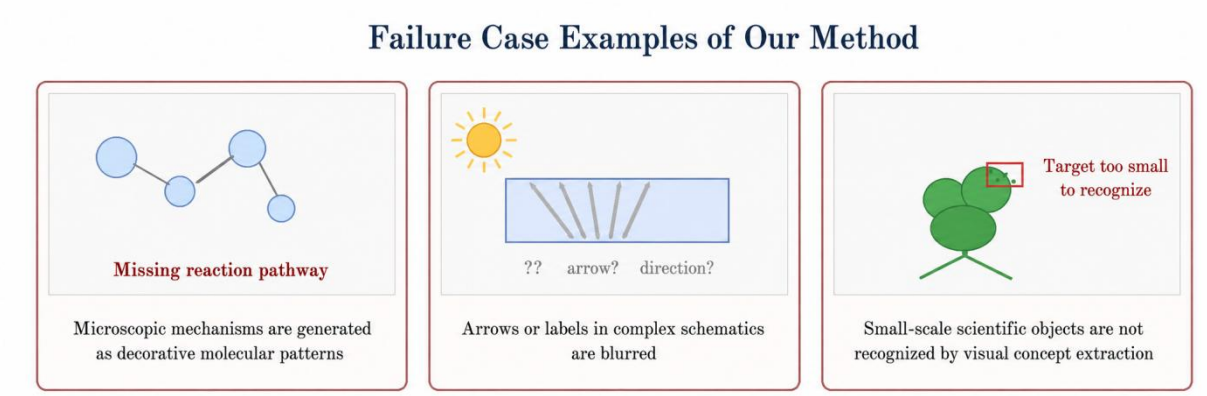

*Note.* The figure compares generation results of five methods on two typical topics: water cycle and photosynthesis, demonstrating differences in concept completeness and instructional clarity.

Figure 3 presents representative failure cases. The main failures include: (1) microscopic chemical reaction mechanisms being converted into decorative molecular patterns; (2) arrows or text labels becoming blurred in complex schematic images; and (3) small scientific objects not being recognized by the captioning model, which causes concept-coverage estimation bias. These cases indicate that the lightweight reranking pipeline still depends on the quality of generated candidates and the capability of visual concept extraction.

**Figure 3**

*Failure cases of the proposed method and corresponding reasons*

WISDOM ACADEMIC https://doi.org/10.63808/css.v2i3.427 Volume 2 , Issue 3




*Note.* Three typical failure modes are presented: microscopic mechanism distortion, blurred schematic labels, and fine-grained concept recognition bias.

## 5. Conclusion

Aiming at the problems of concept missing and semantic deviation in popular science educational image-text generation, this paper proposes a lightweight method based on concept constraint and semantic consistency reranking. The proposed method extracts scientific concepts from input texts to construct enhanced prompts, and reranks candidate images by integrating CLIP image-text similarity, concept coverage and educational applicability. Experimental results show that compared with conventional diffusion model generation(Ho et al., 2020; Nichol & Dhariwal, 2021; Ramesh et al., 2021; Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022), prompt enhancement(Liu & Chilton, 2022) and single CLIP reranking methods(Hessel et al., 2021; Radford et al., 2021), the proposed method effectively improves image-text consistency, concept coverage and classroom application value of generated images.

The main advantage of the proposed method is that it does not require retraining large-scale text-to-image models and can be implemented at low computational costs. However, there are still several limitations. First, concept coverage evaluation relies on image descriptions or manual labels, with limited ability to recognize fine-grained scientific relationships. Second, the generation of abstract processes such as electromagnetic induction and microscopic chemical reaction mechanisms requires stronger structured graphic control. Third, the current dataset is mainly oriented to the Chinese K-12 science curriculum, and cross-language or cross-domain generalization has not been fully verified.

In terms of reproducibility, the dataset annotation protocol, prompt templates, concept dictionary and evaluation code will be organized and released after acceptance. Future research can further improve scientific image generation by introducing visual question

answering models(Zhang et al., 2023a), layout control strategies(Zhang et al., 2023b), teacher feedback mechanisms and larger controlled classroom experiments to directly evaluate learning gains.

## Acknowledgments

The authors have no additional acknowledgments to declare.

## Author contributions

Conceptualization, Yuewen Cao; methodology, Yuewen Cao; software, Yuewen Cao and Xinyang Li; validation, Yuewen Cao, Xinyang Li, Chunyu Lei and Lin Zhang; formal analysis, Yuewen Cao; investigation, Yuewen Cao and Xinyang Li; resources, Chunyu Lei and Lin Zhang; data curation, Yuewen Cao and Xinyang Li; writing — original draft preparation, Yuewen Cao; writing — review and editing, Yuewen Cao, Xinyang Li, Chunyu Lei and Lin Zhang; visualization, Yuewen Cao; supervision, Yuewen Cao and Lin Zhang; project administration, Yuewen Cao; funding acquisition, Yuewen Cao and Lin Zhang. All authors have read and agreed to the published version of the manuscript.

## Funding

This research was funded by the 2026 Undergraduate Innovation and Entrepreneurship Training Program of Beijing Technology and Business University, grant number S202610011034.

## Data availability

This research was funded by the 2026 Undergraduate Innovation and Entrepreneurship Training Program of Beijing Technology and Business University, grant number S202610011034.

## Declarations

### Conflict of interest

The authors declare that there is no conflict of interest.

### Ethics approval

Not applicable.

### Consent to participate

Not applicable.

### Consent for publication

All authors have given their consent.

WISDOM ACADEMIC https://doi.org/10.63808/css.v2i3.427 Volume 2 , Issue 3

## Additional information

**Received:** 25 May 2026

**Accepted:** 29 June 2026

**Published online:** 16 August 2026

### Publisher's Note

## References

[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10684–10695). IEEE. https://doi.org/10.1109/CVPR52688.2022.01042

[2] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8748–8763). PMLR. https://doi.org/10.48550/arXiv.2103.00020

[3] Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., & Kalyan, A. (2022). Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 2507–2521).

[4] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 6840–6851). https://doi.org/10.48550/arXiv.2006.11239

[5] Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 162, pp. 8162–8171). PMLR. https://doi.org/10.48550/arXiv.2102.09672

[6] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Seyed Ghasemipour, S. K., Gontijo Lopes, R. K., Ayan, B. K., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 36479–36494). https://doi.org/10.48550/arXiv.2205.11487

[7] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8821–8831). PMLR. http://proceedings.mlr.press/v139/ramesh21a.html

[8] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2204.06125

[9] Hessel, J., Holtzman, A., Forbes, M., & Choi, Y. (2021). CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural*

*Language Processing (EMNLP)* (pp. 7514–7528). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.595

[10] Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning* (Vol. 162, pp. 12888–12900). PMLR. http://proceedings.mlr.press/v162/li22n.html

[11] Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 19730–19742). PMLR. http://proceedings.mlr.press/v202/li23q.html

[12] Zhang, C., Zhang, C., Zheng, S., Qiao, Y., Li, C., Zhang, M., Dam, A., Thrush, T., Kembhavi, A., & Ostendorf, M. (2023a). A survey on multimodal large language models [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2306.13549

[13] Maharana, A., Hannan, D., & Bansal, M. (2022). StoryDALL-E: Adapting pretrained text-to-image transformers for story continuation. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)* (pp. 70–87). Springer. https://doi.org/10.1007/978-3-031-19769-9_5

[14] Liu, V., & Chilton, L. B. (2022). Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–23). Association for Computing Machinery. https://doi.org/10.1145/3491102.3501824

[15] Mayer, R. E. (2020). *Multimedia learning* (3rd ed.). Cambridge University Press.

[16] Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16*(3), 183–198. https://doi.org/10.1016/j.learninstruc.2006.02.001

[17] Zhang, L., Rao, A., & Agrawala, M. (2023b). Adding conditional control to text-to-image diffusion models. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 3836–3847). IEEE. https://doi.org/10.1109/ICCV51070.2023.00355

[18] Xue, D., Qian, S., & Xu, C. (2023a). Variational causal inference network for explanatory visual question answering. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 2515–2525). IEEE. https://doi.org/10.1109/ICCV51070.2023.00220

[19] Xue, D., Qian, S., & Xu, C. (2024). Integrating neural-symbolic reasoning with variational causal inference network for explanatory visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(12), 7893–7808. https://doi.org/10.1109/TPAMI.2024.3388273

[20] Xue, D., Qian, S., Fang, Q., & Xu, C. (2022). MMT: Image-guided story ending generation with multimodal memory transformer. In *Proceedings of the 30th ACM International Conference on Multimedia (MM)* (pp. 750–758). Association for Computing Machinery. https://doi.org/10.1145/3503161.3547916

[21] Xue, D., Qian, S., Fang, Q., & Xu, C. (2025). LININ: Logic integrated neural inference network for explanatory visual question answering. *IEEE Transactions on Multimedia*. https://doi.org/10.1109/TMM.2024.3521709

[22] Wang, J., Zhang, Y., Zhang, L., Yang, P., Gao, X., Wu, Z., Dong, X., He, J., Zhuo, J., Yang, Q., Huang, Y., Li, X., Wu, Y., Lu, J., Zhu, X., Chen, W., Han, T., Pan, K., Wang, R., Wang, H., Wu, X., Zeng, Z., Chen, C., Gan, R., & Zhang, J. (2022). Fengshenbang 1.0: Being the foundation of Chinese cognitive intelligence [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2209.02970

[23] Yang, A., Pan, J., Lin, J., Men, R., Zhang, Y., Zhou, J., & Zhou, C. (2022). Chinese CLIP: Contrastive vision-language pretraining in Chinese [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2211.01335

[24] Qwen. (2025). Qwen2.5 technical report [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2412.15115