Automatic and Reliable Evaluation for Academic Caption-to-Figure Generation with LMMs

Guanghui Ye[1], **Huan Zhao**[1]*, **Qin Zhu**[1], **Fengnan Li**[2]
Jiaqi Li[3], **Yixian Shen**[4], **Zhonghao Ren**[1], **Zhihua Jiang**[5]*
[1]College of Computer Science and Electronic Engineering, Hunan University, China
[2]Department of Biostatistics Bioinformatics, Duke University, USA
[3]Department of Computer Science, University of Warwick, UK
[4]Informatics Institute, University of Amsterdam, Netherlands
[5]Department of Computer Science, Jinan University, China

Abstract

Existing datasets for evaluating text-to-image generation focus mostly on real-life images, which poses challenges for assessing academic figure generation given real scientific captions, which is a hot topic in AI for Science. To fill the gap, we propose **HE4AFG**, a novel dataset which first provides a **H**olistic **E**valuation for **A**cademic caption-to-**F**igure **G**eneration (AFG). Specifically, HE4AFG collects real figure captions from 8 scientific domains and finally generates 3,900 evaluation samples (particularly, including multi-panel figures) using 5 mainstream large multimodal models (LMMs). For each sample, we provide high-quality human ratings in terms of three aspects—*scientific aesthetic (SA)*, *topic relevance (TR)*, and *attribute correctness (AC)*. Moreover, we present two trainable models: (1) **HE4AFG-E**, an automated **E**valuation model for AFG, which generates aspect-aware training examples and then use them to train three aspect-specific evaluation modules via contrastive learning; (2) **HE4AFG-R**, an automated **R**efinement model, which generates and utilizes feedback on the quality of the figures (e.g., unfaithful elements) to continuously improve AFG. Extensive experiments on HE4AFG demonstrate the effectiveness and performance advantages of our models.

1 Introduction

The rapid acceleration of academic research requires innovative artificial intelligence (AI) tools to explore new concepts and tackle complex challenges (Zhao et al., 2025; Su et al., 2025; Xu et al., 2025). Figures presented in academic papers (we call them *academic figures* for short) play a pivotal role in scientific communication (Roberts et al., 2024; Li et al., 2024; Zhang et al., 2025a), serving as essential tools for researchers to convey complex ideas and data. Although recent large multimodal models (LMMs) have demonstrated impressive capabilities in scientific research (Li et al., 2025; D'Souza et al., 2025), there are unique challenges to evaluate the quality of text-to-image (T2I) generation in academic domains. In particular, when given a long prompt such as real figure captions, instead of simply generating the visual objects, the resulting figure should more emphasize the precision in binding entity attributes and addressing long-range dependencies. As exemplified in Fig.1, the prompt specifies "seven green dashed lines maintain equal spacing", but the image shows "unequal spacing between them". These discrepancies indicate a failure in attribute binding.

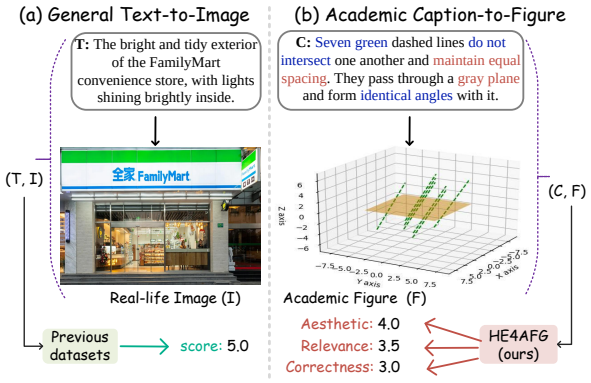


Figure 1: Different from general T2I, AFG focuses more on scientific details (e.g., correct/incorrect attribute binding). Instead of merging all evaluations into a single score, HE4AFG provides more **fine-grained scores** across multiple dimensions (scientific aesthetic, topic relevance, attribute correctness).

Scientific diagrams are a scarce resource due to the difficulty in extraction and annotation. We list some recent datasets in Table 1. The key insights are: (i) Large-scale datasets like Multi-modal ArXiv (Li et al., 2024) and SciCap (Hsu et al., 2021) lack human annotation, while smaller datasets like SciImage (Zhang et al., 2025a) and Science-T2I (Li et al., 2025) are human-annotated, ensuring higher quality and reliability for their specific tasks. (ii) Datasets are highly specialized and focus on different tasks within the academic domain. The largest datasets (SciCap, Multi-modal

*Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 44406–44423
July 2-7, 2026

Dataset	Size	Human-Annotated?	Tasks
SciCap (Hsu et al., 2021)	2M image-caption pairs	No	image captioning
Multi-modal ArXiv (Li et al., 2024)	6.4M images and 3.9M captions	No	image captioning; question-answering
SciFIBench (Roberts et al., 2024)	2,000 image-caption pairs	No	image captioning; question-answering
CharXiv (Wang et al., 2024)	2,323 images	No	question-answering
DaTikZ (Belouadi et al., 2024)	120,000 text-image pairs	No	text-to-image
VGBench (Zou et al., 2024)	4,279 / 5,845 text-image pairs	No	question-answering / text-to-image
ScImage (Zhang et al., 2025a)	3,000 images and 404 prompts	Yes	text-to-image (evaluation)
Science-T2I (Li et al., 2025)	20,000 images and 9,000 prompts	Yes	text-to-image (hallucination detection)
HE4AFG (ours)	3,900 images and 650 captions	Yes	caption-to-figure (evaluation)

Table 1: Comparisons between existing scientific image datasets and HE4AFG.

ArXiv) are primarily designed for image captioning or question answering. Newer datasets such as DaTikZ (Belouadi et al., 2024) and VGBench (Zou et al., 2024) are designed for the T2I task. ScImage is explicitly created to evaluate T2I models, while Science-T2I focuses on the detection of hallucinations in the generated images. Unlike ScImage, we use real figure captions that originate from academic papers. Compared to synthetic prompts, they are more natural and coherent, following the writing styles of researchers (see Appendix A).

Human evaluation of text-image alignment is costly and time-consuming. Therefore, it is crucial to establish automatic and reliable evaluation models that strongly correlate with human judgments (Zhang et al., 2025a; Li et al., 2025; D'Souza et al., 2025). However, existing methods face their key limitations. (i) Human experts can provide precise scores as "ground-truth", but the cost is highly expensive (Zhang et al., 2025a). (ii) Roughly most LMMs can take interleaved text and figures as input and therefore be used to evaluate the T2I task. However, they often have a low correlation with human scores (Hessel et al., 2021; Saxon et al., 2024; Kirstain et al., 2023). (iii) Existing metrics, such as TIFA (Hu et al., 2023), measure faithfulness by generating questions from the given prompt and then answering these questions based on the generated figure. However, they mostly focus on natural figures. (iv) The GPT-4 series (e.g., GPT-4o) perform best when serving as a multimodal evaluator due to their amazing abilities (Xiong et al., 2025), while still being hindered by expensive APIs.

To alleviate these issues, we propose **HE4AFG**, a novel dataset which first provides a **H**olistic **E**valuation for **A**cademic caption-to-**F**igure **G**eneration. In the context of AFG, essential evaluations measure how closely the synthesized figure matches its input caption, emphasizing precision in incorporating target objects, reasoning about spatial relations, and binding entity attributes. Specifically, we introduce key aspects: (i) *scientific aesthetic (SA)—the figure features a clean, clear, vector-style aesthetic; (ii) topic relevance (TR)—the generated figure is relevant to the caption; and (iii) attribute correctness (AC)—the attribute binding in the figure is correct*. We first collect 650 original caption-figure pairs that span 8 academic domains. Particularly, these pairs cover composite figure types such as multi-panel figures (accounting for about 10%). We then utilize 5 mainstream generative LMMs (GPT-5 (OpenAI, 2025), Qwen3 (QwenTeam, 2025), DALL-E (Cho et al., 2023), Llama (Kassianik et al., 2025), Stable Diffusion (SD) 3.5 (Rombach et al., 2022)) to generate additional figures given collected captions as instructions. Subsequently, we employ an expert team to annotate each caption-figure pair in terms of three quality dimensions (SA, TR, AC). Finally, HE4AFG is positioned as a high-quality, multi-domain, type-rich dataset, consisting of a total of 3,900 academic caption-to-figure samples.

Moreover, we present **HE4AFG-E**, which is specifically developed as an automated **E**valuation model on HE4AFG. Specifically, we train three aspect-specific sub-evaluators to generate reliable scores. Considering that CLIP (Radford et al., 2021) remains a strong baseline, we propose a Scientifically adapted CLIP (SCLIP) framework, where we improve the perception of academic figures of the CLIP text and visual encoders. We start from existing scientific image datasets (e.g., ArXivCap (Li et al., 2024)) to generate positive and negative pairs with respect to (w.r.t.) each dimension. We then train the CLIP encoders via intra- and cross-modal contrastive learning (Wu et al., 2024; Higa et al., 2023). Although hard negative generation is standard in contrastive learning, our approach differs in how negatives are constructed considering the targeted dimensions and how they are integrated with the training objective. Furthermore, we introduce **HE4AFG-R**, which serves as an automated **R**efinement model on HE4AFG. HE4AFG-R follows a three-step process: (1) generating feedback on figure quality during data curation; (2) fine-tuning a lightweight LMM using the generated feedback; and (3) performing a feedback-based iterative figure generation using the fine-tuned model.

This work represents a fundamental yet emerging research direction aimed at developing novel datasets and automatic models for academic figure generation, which is highly significant for understanding the alignment between multimodal content and academic semantics in the fields of computer vision and natural language processing. To our knowledge, our work proposes **the first evaluation dataset & model specifically designed for this target task**. Our contributions are four-fold:

- We propose a novel HE4AFG dataset, which aims to evaluate academic figure generation driven by real captions from academic papers. HE4AFG covers multiple disciplinary fields, contains complex graph types, and adds expert-level scores for model performance measures. **It fills the gap of comprehensive evaluation** in academic figure benchmarks.
- We introduce an accompanying HE4AFG-E model, which can automatically provide reliable scores to assess AFG. Specifically, we train aspect-specific evaluation modules by constructing hard negatives and adapting modality encoders, which exhibits **a unique solution to academic figure challenges**.
- We also explore a refinement model, HE4AFG-R, which generates and leverages feedback to iteratively improve generated figures, **creating a virtuous cycle** where evaluation drives generation, and generation, in turn, adds value to the evaluation.
- Our experiments span multiple approach categories, including a large collection of frontier LMMs and task-specific evaluation metrics. HE4AFG-E achieves **the highest correlation with human judgments**, surpassing the absolute points of GPT-4o $\sim$3% on average.

2 Related Work

For the space limit, we present detailed descriptions of related work in Appendix B. Here, we highlight the associations between previous studies and ours.

Based on existing datasets, (a) we start with ArXivCap (Li et al., 2024), where we select correct caption-figure pairs as positive examples and then generate negative caption-figure pairs; (b) we select caption samples from SciCap (Hsu et al., 2021) to generate assessed figures. This shows that our training and evaluation data do not overlap to ensure fairness; (c) we also examine our model on two previous datasets: Q-Eval-100K (Zhang et al., 2025b) and ScImage (Zhang et al., 2025a), showing its generalization. Compared to existing models, our HE4AFG-E model offers several merits: (i) it serves as an automatic tool that avoids excessive human annotations; (ii) it is reference-free, capable of assessing responses without relying on gold answers; (iii) it is tailored to academic figure tasks, a setting largely neglected by prior models; and (iv) it remains low-cost, relying only on pretrained modality encoders and lightweight LMMs.

3 Dataset: HE4AFG

To provide a holistic evaluation of AFG, we propose HE4AFG, as exemplified in Fig.2. We detail the data curation and annotation process below.

Caption-Figure Pair Extraction. We collect a set of caption-figure pairs sourced from (i) existing datasets such as SciCap (Hsu et al., 2021), which focus on Computer Science papers, and (ii) ArXiv papers that span other academic domains. The caption-figure pairs are extracted from the original LaTeX files by matching the syntax pattern and then cleaned according to the rules designed manually. After these processes, 650 pairs are generated by performing an additional human inspection to ensure that all pairs contain clear figures and correct captions. Table 2 presents all domain names, figure types, and data statistics. Compared to existing datasets, HE4AFG already represents a relatively broad domain coverage. It also covers a variety of figure types, ranging from simple figures such as statistical analyzes charts to composite figures such as multi-panel figures (mpf) where distinct sub-figures share a common caption.

Domain Name	Full Name	Percentage
cs	Computer Science	18.8%
econ	Economics	14.8%
q-fin	Quantitative Finance	13.7%
q-bio	Quantitative Biology	10.3%
eess	Electrical Engineering and Systems Science	12.2%
stat	Statistics	9.9%
physics	Physics	10.7%
math	Mathematics	9.6%
Figure Category	Full Name	Percentage
spf	single-panel figure	90.2 %
mpf	multi-panel figure	9.8 %

Table 2: Key statistics of HE4AFG.

Evaluation Sample Generation. We then utilize latest LMMs representing a variety of model architectures and training strategies to generate the evaluated samples. These models include GPT-5, Qwen3, DALL-E, Llama[1], and SD 3.5. Using each

[1]Following ScImage, we adopt the "text-code-image" mode for Llama, i.e., the model first generates intermediate codes (e.g., python) and then compiles them into images.

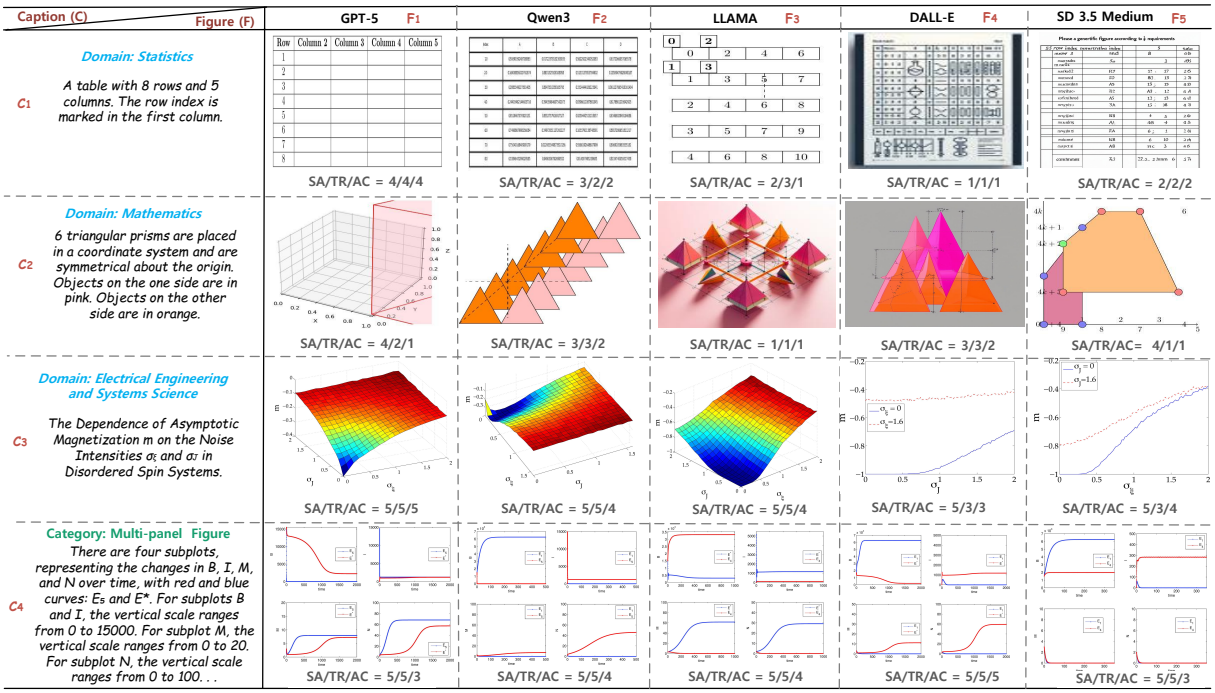


Figure 2: Examples from HE4AFG. For each line-wise example, we present an input caption and five output figures generated by distinct LMMs, each with the scores annotated by participants. Captions C1∼C4 represent different cases. C1: the captions involving abstract concepts are challenging for all models (all have low scores). C2: fine-grained captions involving spatial, quantitative, and attribute requirements show the disparity between models (some have high scores, while some have low scores). C3: simple, clear captions are helpful to generate high-quality figures across distinct models (all have high scores). C4: if a figure contains multiple subplots, its caption should clearly describe the purpose of each one (such as "For subplots B and I").

collected caption, we generate five synthetic figures through these models. This process results in a total of 3,900 evaluation samples (3,250 generated samples + 650 original samples) in the final dataset. Both domain diversity (e.g., 8 academic domains) and model diversity (e.g., 5 generative LMMs) inherently help dilute the specific biases of individual samples, making the dataset more representative.

Human Annotation. We represent a holistic evaluation of caption-to-figure generation using three criteria: scientific aesthetic (SA), topic relevance (TR), and attribute correctness (AC). Each criterion is rated on a scale of 1 to 5, with 5 being the highest score. We employ a panel of 3 Ph.D students from distinct disciplines, ensuring domain expertise in evaluating academic visualizations. We provide each annotator with detailed annotation guidelines (Appendix D.2). We assign each sample to each annotator and calculate a mean of 3 participants as the final score. We also calculate the averaged weighted kappa value for all sampled instances and get a high score of 0.81, demonstrating good agreement between data specialists. We tax the value of our evaluation at roughly 900 USD, with up to 3 annotators involved for up to 15 hours each, all working for 20 USD per hour, on average.

4 Evaluation Model: HE4AFG-E

HE4AFG-E achieves the estimation of the caption-figure alignment by maximizing consistency. These training objectives combine rich learning signals from matched and mismatched caption-figure pairs. Therefore, we develop a task-aligned training framework, comprising three phases based on contrastive learning, to incorporate the features of academic figures into existing generative models.

4.1 Training Data Curation

Positive Example Generation. Since our target is to train HE4AFG-E to assign a relatively high score for aligned caption-figure pairs while a relatively low score for those unaligned, we need to construct positive examples (positives) and negative examples (negatives). We use ArXivCap (Li et al., 2024) as the source dataset. Specifically, we randomly select 32K caption-figure pairs $\langle C, F \rangle$ that span 32 academic domains (we select 1K for each domain). Each caption-figure pair in ArXivCap comes from real ArXiv papers, thus regarded as a positive example $\langle C_G, F_G \rangle$ (i.e., gold pair).

Negative Example Generation. As Fig.3 shows, we perform the following steps:

(1) Aesthetic Negatives: Because aesthetic evaluation is relatively subjective (there are no proper

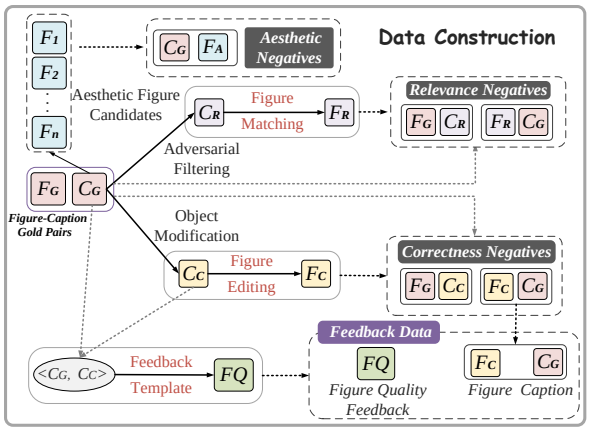


Figure 3: Training data generation for HE4AFG-E. **(1)** We use gold caption-figure pairs $\langle C_G, F_G \rangle$ as positives. **(2)** We construct aesthetic negatives $\langle C_G, F_A \rangle$ via human ratings. **(3)** We create relevance negatives $\langle C_R, F_G \rangle$ / $\langle C_G, F_R \rangle$ via adversarial filtering. **(4)** We build correctness negatives $\langle C_C, F_G \rangle$ / $\langle C_G, F_C \rangle$ via object modifications. **(5)** We generate textual feedback on figure quality FQ to train HE4AFG-R later.

automation tools yet), we manually annotate the aesthetic quality score of the figures in the training set and then select those with low scores as negatives. Specifically, we construct a small subset of figures ($\{F_1, F_2, ..., F_n\}$) and assign an aesthetic score ranging from 1 to 5 to each figure. To construct hard negatives, for a gold caption C_G, we first calculate its cosine similarity with each figure F_i using the vanilla CLIP. When similarity exceeds a certain threshold (e.g., 0.8), we select the figure F_A (which is highly relevant to C_G) with the lowest aesthetic score, and produce a negative pair $\langle C_G, F_A \rangle$ in terms of aesthetic criterion.

(2) Relevance Negatives: Following SciFIBench (Roberts et al., 2024), we generate hard negatives in terms of topic relevance via adversarial filtering, to further enhance the discriminative ability of the model. Specifically, we select the caption-figure pair most similar to $\langle C_G, F_G \rangle$ from the ArXivCap dataset as follows: We first use CLIP to compute the embedding for each caption; We then create a vector database of caption embeddings using Faiss (Douze et al., 2024); For each embedding, we search for the nearest neighbor based on Euclidean distance; Finally, we traverse the caption C_R that is best matched to the embedding of C_G in the vector dataset and then find the figure F_R that is paired with C_R from ArXivCap. Thus, we generate two negative pairs $\langle C_R, F_G \rangle$ and $\langle C_G, F_R \rangle$.

(3) Correctness Negatives: We first generate C_C by making a minor revision to C_G. Specifically, we adopt an attribute dictionary D from SciImage (Zhang et al., 2025a), which defines key elements (e.g., objects, attributes, spatial relations, etc) relevant to scientific images (Appendix C). Using a dictionary-based approach enables controllable modifications to the original caption while achieving the goal of generating hard negatives. We select only one attribute from D at a time and randomly assign it a valid value. Because C_C is almost identical to C_G, it is infeasible to select a nearest neighbor solution like $\langle C_R, F_R \rangle$. Thus, we use a lightweight image editor (e.g., SEED-X (Ge et al., 2024)) to generate a new figure F_C given both F_G (as input figure) and C_C (as editing requirements). After this, we generate two negative pairs $\langle C_C, F_G \rangle$ and $\langle C_G, F_C \rangle$ in terms of correctness.

4.2 Training HE4AFG-E (SA, TR, AC)

We train three pairs of textual encoders $CLIP_t^{SA/TR/AC}$ and visual encoders $CLIP_v^{SA/TR/AC}$ using aspect-sensitive training data under the scientifically adapted CLIP (SCLIP). Specifically, we use positive pairs (such as $\langle C_G, F_G \rangle$) and negative pairs (such as $\langle C_G, F_A \rangle$/$\langle C_G, F_R \rangle$/$\langle C_G, F_C \rangle$) for contrastive learning. During inference, given a pair $\langle C, F \rangle$, an aspect-specific score will be derived by $s(CLIP_t^{SA/TR/AC}(C), CLIP_v^{SA/TR/AC}(F))$, where s is the cosine similarity function.

SCLIP is designed to be more sensitive to scientific details. The idea of contrastive learning is to bring positive pairs closer while pushing negative pairs apart. Due to multimodal input, we leverage the intra-modal ($\mathcal{L}_{IM}$) and cross-modal ($\mathcal{L}_{CM}$) contrastive loss. (1) The insight in $\mathcal{L}_{IM}$ is to keep the representations of these samples as far as possible in the feature space. Specifically, we define $\mathcal{L}_{IM}$ in two scenarios: $\mathcal{L}_{IM_t}$ (text) and $\mathcal{L}_{IM_v}$ (visual). Let $\mathbf{Z}_{C_T}$ and $\mathbf{Z}_{C_F}$ be feature vectors of C_T and C_F respectively, where C_T (T – True) represents a positive caption such as C_G while C_F (F – False) represents a negative caption such as C_R. Then, $\mathcal{L}_{IM_t}$ can be derived by $s(\mathbf{Z}_{C_T}, \mathbf{Z}_{C_F})$. Similarly, we calculate $\mathcal{L}_{IM_v}$ using visual feature vectors. (2) The insight in $\mathcal{L}_{CM}$ is to bring the projected representations of an aligned caption-figure pair as close as possible and also to keep an unaligned caption-figure pair as far as possible. Thus, we define $\mathcal{L}_{CM}$ in two scenarios: $\mathcal{L}_{CM}^P$ (positive) and $\mathcal{L}_{CM}^N$ (negative). First, we can define $\mathcal{L}_{CM}^P$ based on two aligned caption-figure pairs $\langle C_T, F_T \rangle$ and $\langle C_F, F_F \rangle$: $\mathcal{L}_{CM}^P = \exp(s(\mathbf{Z}_{C_T}, \mathbf{Z}_{F_T})/\tau) + \exp(s(\mathbf{Z}_{C_F}, \mathbf{Z}_{F_F})/\tau)$ ("exp" is the exponential function and τ is a temperature parameter). We can

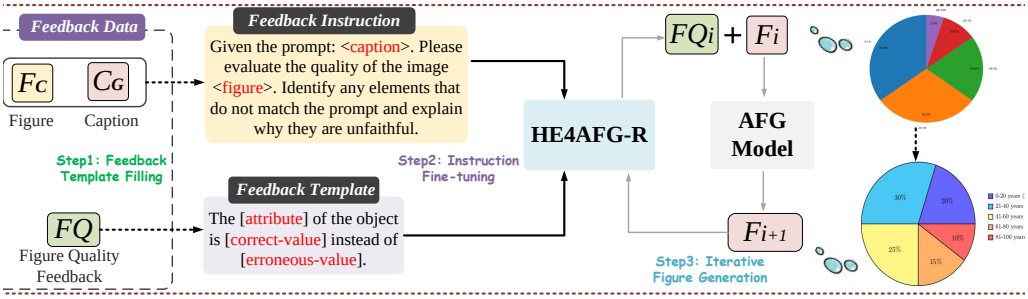


Figure 4: Overall, HE4AFG-R first fine-tunes a feedback generator (left) and then employs it to iteratively guide AFG (right).

also define $\mathcal{L}_{CM}^{N}$ based on two pairs of unaligned caption-figure $\langle C_T, F_F \rangle$ and $\langle C_F, F_T \rangle$: $\mathcal{L}_{CM}^{N} = \exp(s(\mathbf{Z}_{C_T}, \mathbf{Z}_{F_F})/\tau) + \exp(s(\mathbf{Z}_{C_F}, \mathbf{Z}_{F_T})/\tau)$. Referring to the InfoNCE loss (van den Oord et al., 2018), we generate the final $\mathcal{L}_{CM}$ by connecting $\mathcal{L}_{CM}^{P}$ (maximized) and $\mathcal{L}_{CM}^{N}$ (minimized) with a negative logarithmic function. Finally, we integrate $\mathcal{L}_{CM}$ and a weighted $\mathcal{L}_{IM}$: $\mathcal{L} = \mathcal{L}_{CM} + \lambda\mathcal{L}_{IM_t} + (1-\lambda)\mathcal{L}_{IM_v}$, where λ is a weight parameter.

4.3 Extension to Multi-Panel Figures

We design an extension scheme that employs a three-stage progressive approach to evaluate multi-panel figures (*mpf*). Specifically, it takes an *mpf* and its shared caption as input and outputs a final aggregate result. Stage 1, Panel Segmentation. We utilize detection models (e.g., Detectron2 (Wu et al., 2019)) to automatically identify sub-figure regions within *mpf* and extract figure labels (e.g., flow chart, statistical chart), while handling shared elements (such as legends and axes) to prevent over-segmentation. Stage 2, Panel-Caption Alignment. We first utilize LLMs (e.g., SciBERT (Beltagy et al., 2019)) to parse the shared caption and decompose it into sub-figure description fragments. We then use object detection tools (e.g., YOLOv8 (Ultralytics, 2023)) to identify key visual objects within each panel. We last select the most appropriate caption fragment for each panel according to multimodal similarity mechanisms (e.g., CLIP-Score (Hessel et al., 2021)). Stage 3, Single-Figure Evaluation. After Stages 1∼2, each evaluation sample is in single-figure scenarios. Thus, we are able to use HE4AFG-E to assess each sub-pair and generate a final score (e.g., averaging over the scores achieved for sub-pairs). This extension scheme elevates HE4AFG-E from a single-figure evaluation to an effective evaluation of composite figures.

5 Refinement Model: HE4AFG-R

We also present HE4AFG-R (Fig.4) which generates and then uses feedback to continuously improve the precision of the figures. This establishes a beneficial cycle: evaluation facilitates generation, and generation renders evaluation more valuable.

Step1: Feedback Template Filling. An object part that violates the required attribute binding is called "unfaithful element", which enables evaluators to pinpoint specific inaccuracies in model-generated content. We use these elements as an explicit signal of figure quality feedback (FQ). Specifically, we compare the input caption C_G and the output figure F_C and then generate FQ samples using a pre-defined template such as "The [attribute] of the object is [correct-value] instead of [erroneous-value]"), where the slots can be filled with the items identified from C_G and F_C. E.g., the caption C_G specifies "seven green dashed lines", but the figure F_C shows "seven red dashed lines", thus generating a specific sample such as "The [color] of the object is [green] instead of [red]".

Step2: Instruction Fine-tuning. Supervised Fine-Tuning (SFT) is a technique that conducts targeted training on a pre-trained large model using labeled data, enabling the model to better adapt to specific applications. Therefore, we fine-tune mPLUG-owl3 (8B) (Ye et al., 2025) as a feedback generator using the feedback samples generated.

Step3: Iterative Figure Generation. We design a feedback-based iterative process to enhance figure generation. Concretely, we first use the fine-tuned mPLUG-owl3 to generate a current feedback FQ_i conditioned on the gold caption C_G and the current figure F_i. Then, we use FQ_i as a refinement instruction for the AFG model to correct the errors in F_i while assuming that the generation is ongoing. Next, we integrate the new figure F_{i+1} and C_G to start the next iteration. This enables us to leverage feedback that represents a relevant and meaningful quality assessment to guide generation.

Model[1]	cs	econ	q-fin	q-bio	eess	stat	physics	math
Closed-source LMMs								
GPT-4V	(68.3, 63.2, 68.9)	(52.8, 50.1, 53.1)	(58.5, 55.4, 58.3)	(60.4, 57.8, 62.6)	(64.3, 60.9, 64.9)	(56.3, 52.8, 57.1)	(68.4, 67.6, 68.9)	(62.1, 59.2, 63.0*)
GPT-4 Turbo	(68.8, 62.7, 69.0)	(51.1, 48.8, 53.4)	(57.3, 54.7, 58.1)	(59.4, 56.7, 59.9)	(65.5, 62.3, 66.4)	(54.9, 51.0, 55.0)	(63.2, 60.4, 63.9)	(67.2, 64.5, 67.5)
GPT-4o	(71.7, 66.8, 72.3)	(58.8, 55.6, 59.4)	(64.4, 61.7, 64.4)	(64.7, 61.9, 65.6)	(70.5, 67.8, 70.4)	(63.2, 60.0, 64.5)	(68.8, 65.9, 68.7)	(63.9, 60.1, 64.8)
Gemini Pro Vision	(57.8, 53.0, 58.0)	(42.0, 39.4, 43.0)	(48.3, 45.5, 49.1)	(50.1, 46.8, 50.8)	(55.0, 51.7, 55.7)	(46.1, 42.6, 46.8)	(53.2, 52.6, 54.3)	(56.0, 53.4, 56.9)
Gemini 1.5 Pro	(71.2, 66.7, 70.4)	(58.3, 55.4, 58.5)	(64.5, 62.1, 65.0*)	(64.4, 61.2, 65.0)	(70.0, 66.8, 70.1)	(60.3, 57.8, 60.8)	(67.8, 66.1, 67.4)	(64.1, 61.7, 64.8)
Claude 3 Haiku	(57.4, 51.3, 59.0)	(43.8, 40.9, 46.0)	(49.2, 46.8, 50.0)	(49.7, 45.9, 49.8)	(55.9, 53.1, 56.4)	(41.7, 36.6, 42.4)	(50.8, 48.3*, 51.4)	(58.2, 55.8, 58.6)
Claude 3 Opus	(59.0, 53.7, 59.6)	(44.3, 42.0, 47.0)	(50.5, 48.4, 50.9)	(49.8, 47.5, 50.5)	(50.9, 49.1, 51.5)	(44.5, 44.1, 45.6)	(54.6, 52.5, 55.0)	(58.2, 56.4, 59.1)
Open-source LMMs								
IDEFICS-9b-Instruct	(23.9, 21.7, 26.4)	(13.5, 13.2, 17.4)	(19.1, 20.8*, 22.0)	(21.3, 21.0, 25.7)	(25.9, 25.4, 26.7)	(17.3, 15.9, 19.0)	(23.1, 24.5*, 25.8)	(28.2, 28.0, 29.6)
IDEFICS-80b-Instruct	(26.6, 25.0, 28.6)	(14.5, 13.0, 18.3)	(28.4, 26.8, 29.6)	(24.9, 23.0, 28.8)	(29.8, 26.9, 30.2)	(24.6, 22.4, 26.3)	(28.6, 26.8, 30.2)	(34.0, 31.6, 33.7)
Emu2	(24.3, 22.6, 25.8)	(18.5, 15.9*, 24.0)	(17.7, 17.0, 18.7)	(24.6, 22.5, 25.6)	(26.9, 24.9, 27.5)	(16.3, 16.0, 18.4)	(22.5, 21.6, 23.6)	(30.3, 28.7, 32.6)
InternLM-XComposer-7b	(38.7, 36.5*, 39.0)	(29.4, 28.2, 31.3)	(35.3, 33.9, 36.0)	(35.5, 35.0, 36.8)	(34.2, 33.7, 36.0)	(33.1, 31.8, 36.7)	(39.4, 38.2, 39.9)	(40.9, 39.4, 41.7)
InstructBLIP-FlanT5-xl	(39.3, 37.7, 40.4)	(29.5, 29.0, 30.6)	(31.0, 30.4, 32.2)	(35.5, 33.5, 35.5)	(41.9, 38.9, 42.6)	(33.7, 30.6, 36.0)	(39.7, 38.9, 41.6)	(37.9, 37.8, 40.2)
InstructBLIP-Vicuna-7b	(41.3, 39.2, 43.6)	(25.7, 25.4, 26.6)	(37.4, 36.8, 38.3)	(37.6, 38.2, 37.9)	(40.5, 37.9, 40.9)	(35.5, 34.9, 37.2)	(41.1, 40.8, 41.3)	(44.7, 43.5, 45.6)
InstructBLIP-Vicuna-13b	(41.4, 39.3, 42.0)	(31.7, 30.5, 32.6)	(37.0, 35.5, 37.8)	(37.5, 36.4, 38.4)	(42.0, 40.9, 43.0)	(27.8, 26.6, 30.9)	(41.8, 38.9, 40.9)	(45.4, 44.7, 45.6)
Qwen-VL-Chat	(27.6, 25.8, 28.9)	(17.6, 15.2, 25.6)	(23.1, 21.8, 27.6)	(26.8, 24.9, 28.9)	(32.7, 31.9, 33.3)	(21.7, 20.0, 22.9)	(30.8, 27.9, 30.6)	(34.9, 32.6, 35.3)
Text-Image Alignment Metrics								
CLIPScore	(42.8, 40.5, 43.6)	(31.5, 30.6, 32.0)	(39.6, 38.5, 39.9)	(39.7, 38.9, 40.2)	(45.6, 43.7, 47.0)	(37.8, 37.0, 39.6)	(40.0, 39.1, 41.4)	(41.0, 39.6, 41.5)
VQAScore	(40.5, 38.8, 42.0)	(32.2, 30.4, 34.7)	(39.3, 35.8, 40.6)	(39.2, 37.4, 40.7)	(43.1, 39.1, 42.6)	(36.5, 33.5, 38.2)	(37.7, 34.8, 39.1)	(45.0, 41.7, 47.4)
BLIP2Score	(42.3, 41.1, 43.8)	(32.5, 30.9, 36.1)	(37.0, 36.6, 37.9*)	(39.5, 39.2, 40.4)	(45.4, 45.2, 48.3)	(37.5, 36.6, 38.9)	(39.1, 39.0, 40.7)	(45.5, 45.4, 47.5)
TIFA	(45.0, 42.6, 46.6)	(33.4, 33.0, 34.9)	(39.7, 39.2*, 40.7)	(39.5, 39.6, 40.6)	(45.6, 45.6, 45.8)	(37.5, 36.9, 38.8)	(41.4, 41.3, 42.5)	(45.7, 46.2, 46.3)
Our Evaluation Model								
HE4AFG-E (ours)	(73.8, 69.3, 73.9)	(62.9, 59.1, 64.3)	(68.4, 65.3, 68.7)	(63.7, 59.0, 64.5)	(72.7, 68.9, 72.3)	(66.6, 62.9, 67.0)	(71.1, 67.4, 72.4)	(76.1, 72.4, 76.6)

[1] We calculate SRCC using normalized annotation scores and model scores. All values are statistically significant due to *p-value* < 0.05 unless marked by *.

Table 3: Performance comparisons of the competitors on HE4AFG. **We report the SRCC values (%) across three dimensions (TR, AC, SA) for each model and each domain**. The higher, the better. The column-wise highest value is highlighted in green.

6 Experiments

6.1 Experimental Setup

Training & Evaluation Datasets: we train HE4AFG-E/R using our curated data sourced from ArxivCap. We evaluate HE4AFG-E/R and all compared methods on HE4AFG. All experiments of our model on HE4AFG were completed in 3 hours (2.5 hours for training and 0.5 hours for inference) on 4 NVIDIA A100 GPUs. **Competitors**: we select a rich set of 19 competitors that can be categorized into three groups (Appendix F): (1) Closed-Source LMMs: GPT-4V (OpenAI, 2023a), GPT-4 Turbo (OpenAI, 2023b), GPT-4o (OpenAI, 2024), Gemini Pro Vision (Anil et al., 2023), Gemini 1.5 Pro (Reid et al., 2024), Claude 3 Opus (Anthropic, 2024), and Claude 3 Haiku (Anthropic, 2024). (2) Open-Source LMMs: IDEFICS (Laurençon et al., 2023), Emu2 (Sun et al., 2024), InternLM-Composer (Team, 2023), InstructBLIP (Dai et al., 2023), and Qwen-VL (Bai et al., 2023). (3) Existing Text-Image Alignment Metrics: CLIPScore (Hessel et al., 2021), VQAScore (Lin et al., 2024), BLIP2Score (Li et al., 2023) and TIFA (Hu et al., 2023). The mainstream LMMs (e.g., GPT-4o), while not tailored to our task, represent state-of-the-art (SOTA) approaches in the field. Their performance indicates how well general-purpose large models transfer to our task. The details of experimental settings (including computational cost, model versions, hyperparameter, and training & inference details) can be found in Appendix E.

6.2 Main Findings

Each evaluator is asked to output three scores (TR, AC, SA). For all output scores, we calculate the Spearman Ranking Correlation Coefficient (**SRCC**) (Appendix E.1). We report the results of HE4AFG-E and 19 competitors across 8 academic domains in Table 3. The confidence intervals (CI) are 95%. This reveals a significant positive association (e.g., SRCC = 0.741, 95% CI [0.724, 0.757], $p < 0.05$). The key findings are as follows:

HE4AFG-E performs best. It significantly outperforms all other competitors, including GPT-4o, in 7 out of 8 domains. This shows the performance advantages and generalization of HE4AFG-E across various domains. Among the weaker models, Gemini 1.5 Pro performs on a par with GPT-4o. The weakest models are IDEFICS and Emu2, showing great room for improvement.

Closed-source models are noticeably better than open-source models. Considering the CS subset, there is a difference of 30.3%/27.5%/30.3% in terms of TR/AC/SA between the result of the best closed-source model (GPT-4o) and that of the best open-source model (InstructBLIP-Vicuna-13b). Moreover, InstructBLIP-13b under-performs the worst closed-source model (Claude 3 Haiku).

Performance does not necessarily scale with model size. Considering the models for which we evaluate multiple checkpoint sizes (e.g., IDEFICS, InstructBLIP, etc.), we find that the larger model does not significantly outperform its smaller checkpoint in our task. This shows that a larger model (e.g., greater than 10b) is not necessary, especially when more practical factors (e.g., reference time and storage overload) are concerned.

HE4AFG-R is effective. Table 4 illustrates the performance gain of an AFG model (e.g., SD) in different iterations, evaluated by GPT-4o and

HE4AFG-E. The core findings include: (i) Overall, the AFG model achieved noticeable performance gains across all three quality dimensions after three iterations. The performance boost is concentrated primarily in Iteration 1. Subsequent iterations (2 and 3) show a trend of diminishing marginal returns. (ii) Both GPT-4o and HE4AFG-E show highly consistent trends. (iii) TR scores rose from 2.94 (Initial) to 3.11 (Iteration 1), finally reaching 3.21 at Iteration 3. It indicates that iterative optimization significantly improved the model's core ability to understand feedback and generate better content. (iv) AC scores improved significantly from 2.49 (Initial) to 2.90 (Iteration 1), stabilizing around 2.98. This dimension saw the most dramatic progress. The initial score was the lowest. After iterations, the AC score approached the 3.0 mark, which indicates a qualitative increase in visual quality. (v) SA scores steadily rose from 3.31 (Initial) to 3.52 (Iteration 3). SA is the strongest dimension for this model (highest scores) and its steady growth indicates that the model maintains strong stylistic semantic during generation.

The AFG Model (SD) (5-Point scale)↑	**TR**		**AC**		**SA**	
	GPT-4o	**HE4AFG-E**	**GPT-4o**	**HE4AFG-E**	**GPT-4o**	**HE4AFG-E**
SD (Initial)	2.56	2.94	2.70	2.49	3.27	3.31
SD (after 1 iteration)	2.77	3.11	3.01	2.90	3.35	3.49
SD (after 2 iterations)	2.83	3.20	3.08	2.97	3.44	3.51
SD (after 3 iterations)	2.83	3.21	3.07	2.98	3.42	3.52

Table 4: Results that show how HE4AFG-R improves AFG.

6.3 Ablation Study

HE4AFG-E is not a pipeline approach; therefore, we did not conduct ablation study on individual modules. The three components HE4AFG-E (TR, AC, SA) are trained independently, using different training data. As Table 5 shows, we perform an ablation study on the temperature parameter τ over the range (0.05, 0.07, 0.1, 0.2, 0.5, 1.0). Finally, we set τ as 0.07 for the best performance. Similarly, we conduct an ablation test on the weight parameter λ over the range (0.1, 0.3, 0.5, 0.7, 0.9) and finally determine $\lambda = 0.5$ for the best performance.

τ (temperature)	SRCC (TR, AC)	λ (weight)	SRCC (TR, AC)
0.05	(72.8, 68.1)	0.1	(69.0, 64.8)
0.07	**(73.8, 69.3)**	0.3	(72.8, 68.1)
0.1	(73.5, 68.8)	**0.5**	**(73.8, 69.3)**
0.2	(71.0, 66.5)	0.7	(72.5, 67.9)
0.5	(68.7, 62.3)	0.9	(68.3, 63.1)
1.0	(63.4, 58.3)	-	-

Table 5: Ablation test of super-parameters on the CS subset.

6.4 In-Depth Analysis

(I) Effectiveness on existing datasets. To validate the generalization of HE4AFG-E, we experiment with previous expert-annotated datasets: (1) Q-Eval-100K (Zhang et al., 2025b), which assesses the alignment level for text-to-vision. We use 60k natural figures to perform evaluations at both instance-level and model-level, following the original paper. (2) ScImage (Zhang et al., 2025a), which provides expert-level annotations to assess scientific text-to-image. We use 3,000 images from the entire dataset. Notably, the input text of both two datasets is synthetic prompts instead of real captions. Observed in Table 6, our model consistently outperforms other competitors in these scenarios.

Model *SRCC (%)*↑	**Q-Eval-100K** (Natural figures)		**ScImage** (Scientific figures)	
	Instance-level	**Model-level**	**Correctness**	**Relevance**
TIFA	59.8	81.3	30.3	39.2
CLIPScore	77.0	95.8	27.4	36.7
InstructBLIP-13b	76.6	93.3	27.0	37.4
GPT-4o	81.5	95.4	40.3	49.3
HE4AFG-M (ours)	**82.3**	**96.5**	**45.2**	**53.2**

Table 6: Results of top competitors on existing datasets.

(II) The AFG ability of LMMs. We further analyze the AFG ability of 5 LMMs used in our dataset. The results are presented in Table 7. Scores of {Human, GPT-4o, HE4AFG-E} averaged across instances are shown, along with the diff between human and model scores (the smaller the better). We find that: (1) GPT-5 performs best, while SD 3.5 performs worst. This shows the disparity between LMMs; (2) Based on diff scores, our model generates results closer to humans than GPT-4o.

(III) Results on 650 original pairs. Reporting only aggregated scores over 3,900 samples might obscure the model's performance on high-quality faithful data. For this, we evaluate competitive methods strictly on the 650 original caption-figure pairs. These results in Table 8 show that: (1) Validation of robustness. Even when restricted to the high-quality "original" subset where figures are known to be faithful to captions, HE4AFG-E still significantly outperforms strong baselines like GPT-4o (e.g., +3.4/+2.3/+2.7 points in TR/AC/SA). This confirms that our model's superiority is not an artifact of the expanded dataset (3,900 total) but holds true for rigorous, real-world academic samples (650 original). (2) Addressing the aggregation concern. Comparing the results in Tables 3 and 8, we observe that the relative ranking of the models remains consistent. HE4AFG-E maintains its lead in both settings. This suggests that while current LMMs do struggle with the harder generated samples (Table 7), HE4AFG-E is capable of distinguishing quality across both easy (original) and hard (generated) distributions. (3) Clarification of the benchmark design. Including the 3,250 generated samples serves as a necessary stress test to

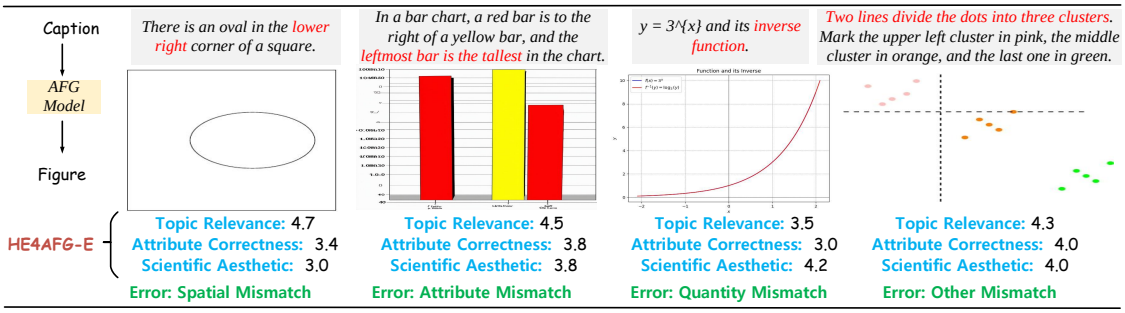


Figure 5: Case study (I). For each sample, HE4AFG-E generates three aspect-specific scores. Typical error types can be observed (unfaithful parts of the output figure to the input caption are highlighted in red).

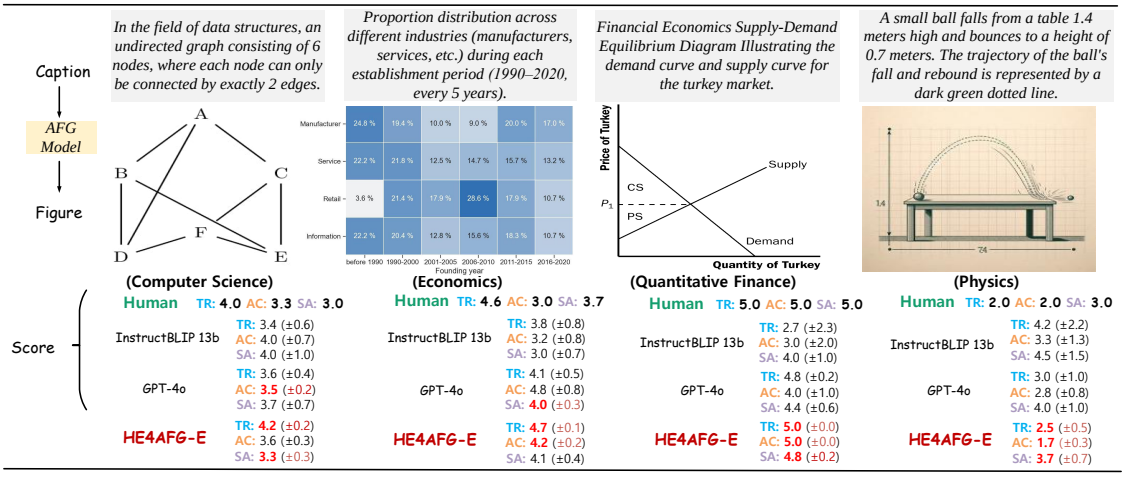


Figure 6: Case study (II). For each sample, we present the model scores and compare with human scores (i.e., diff).

LMMs (5-Point scale)↑	TR			AC		
	Human	GPT-4o	HE4AFG-E	Human	GPT-4o	HE4AFG-E
GPT-5	3.42	3.61 (+0.19)	3.47 (+0.05)	3.10	3.40 (+0.30)	2.94 (-0.16)
Qwen3	3.36	3.62 (+0.26)	3.52 (+0.16)	3.01	3.35 (+0.34)	2.87 (-0.14)
LLAMA	2.94	3.48 (+0.54)	3.12 (+0.18)	2.51	3.01 (+0.50)	2.66 (+0.15)
SD 3.5-Medium	2.83	2.56 (-0.27)	2.94 (+0.11)	2.43	2.70 (+0.27)	2.49 (+0.06)
DALL-E	3.12	3.52 (+0.40)	3.36 (+0.24)	2.96	3.31 (+0.35)	2.88 (-0.08)

Table 7: Overall performance of LMMs. Scores averaged across instances are shown, along with the score differences.

differentiate evaluators. If we only used the 650 original pairs, the gap between the weaker evaluators and the stronger ones would be less discriminative. The expanded set reveals the ceiling of current evaluation capabilities, while the original set confirms baseline reliability.

Model	TR	AC	SA
TIFA	41.2	40.3	42.6
CLIPScore	39.6	38.3	40.7
InstructBLIP-13b	36.2	33.4	38.4
GPT-4o	62.5	60.7	65.6
HE4AFG-E (ours)	65.9	63.0	68.3

Table 8: Results on the 650 original pairs.

6.5 Case Study

We present some qualitative results. In Fig.5, we show the score samples generated only by HE4AFG-E. It is observed that the model demonstrates robust performance in distinguishing image quality, irrespective of caption complexity. Meanwhile, the score results also shed light on certain error patterns, such as inaccuracy in spatial relationships, attribute binding, or numerical values. In Fig.6, we compare the scores generated by 3 best-performing models from distinct groups. In particular, we selected representative figures from different domains to validate the domain adaptation. The highlighted in red show that HE4AFG-E correlates more closely with human judgment than other competitors (including GPT-4o), further confirming its reliability as an automatic model.

7 Concluding Remarks

The proposed dataset (HE4AFG) and its accompanied models (HE4AFG-E/R) address the critical need for a comprehensive, reliable evaluation platform (see broader impact in Appendix G). **First**, its fine-grained assessment provides quantitative scores, enabling researchers to pinpoint specific inaccuracies in the generated content. **Second**, leveraging lightweight scientifically adapted encoders, HE4AFG-E achieves high performance with significantly lower computational cost than large-scale valuators such as GPT-4o. **Third**, the work supports adaptation to new academic domains and integration with various backbone models, positioning the proposed dataset & models as a practical solution for evolving multimodal scientific tasks.

Limitations

Despite its effectiveness, our work has certain limitations that point to promising directions.

First, reliance on synthetic negative samples. In the data construction phase, we manufacture negative samples by modifying correct captions (e.g., altering attributes or quantities) or using image editing models to generate mismatched images, instead of directly collecting a large number of human-scored, real erroneous figures. Although this can produce hard negatives, the errors generated by modifying the correct captions via dictionary replacement or specific instruction editing are often "artificial" or "idealized" compared to the typical error distribution of real generative models in actual inference. Real model hallucinations are often complex and unstructured (e.g., logical errors, physically impossible structures).

Second, dependency on multi-panel figure segmentation. For complex figures containing multiple sub-figures, we use detection models to slice the figure into individual sub-figures before evaluating them separately. This construction method couples the accuracy of the image segmentation with the evaluation dataset. If the segmentation model fails (e.g., cuts the boundaries incorrectly), the subsequent "caption-to-subfigure" evaluation samples are inherently flawed. This results in noise within the ground truth of the evaluation dataset.

References

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Anthropic. 2024. Introducing the next generation of claude.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966.

Jonas Belouadi, Anne Lauscher, and Steffen Eger. 2024. Automatikz: Text-guided synthesis of scientific vector graphics with tikz. In *Proceedings of the International Conference on Learning Representations*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Association for Computational Linguistics.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. DALL-EVAL: probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3020–3031.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *CoRR*, abs/2401.08281.

Jennifer D'Souza, Hamed Babaei Giglou, and Quentin Münch. 2025. Yescieval: Robust llm-as-a-judge for scientific question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 13749–13783.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2024. SEED-X: multimodal models with unified multi-granularity comprehension and generation. *CoRR*, abs/2404.14396.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.

Kyota Higa, Masahiro Yamaguchi, and Toshinori Hosoi. 2023. ICCL: self-supervised intra- and cross-modal contrastive learning with 2d-3d pairs for 3d scene understanding. In *Proceedings of International Conference on Image Processing*, pages 1085–1089.

Ting-Yao Hsu, C. Lee Giles, and Ting-Hao Kenneth Huang. 2021. Scicap: Generating captions for scientific figures. In *Proceedings of Findings of the Conference on Empirical Methods in Natural Language Processing*, pages 3258–3264.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. 2023. TIFA: accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20349–20360.

Anil K. Jain, Karthik Nandakumar, and Arun Ross. 2005. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285.

Paul Kassianik, Baturay Saglam, Alexander Chen, Blaine Nelson, Anu Vellore, Massimo Aufiero, Fraser Burch, Dhruv Kedia, Avi Zohary, Sajana Weerawardhena, Aman Priyanshu, Adam Swanda, Amy Chang, Hyrum S. Anderson, Kojin Oshiba, Omar Santos, Yaron Singer, and Amin Karbasi. 2025. Llama-3.1-foundationai-securityllm-base-8b technical report. *CoRR*, abs/2504.21039.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Proceedings of Advances in Neural Information Processing Systems*.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. OBELICS: an open web-scale filtered dataset of interleaved image-text documents. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jialuo Li, Wenhao Chai, Xingyu Fu, Haiyang Xu, and Saining Xie. 2025. Science-t2i: Addressing scientific illusions in image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2734–2744.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 14369–14387.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part IX*, volume 15067 of *Lecture Notes in Computer Science*, pages 366–384. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

OpenAI. 2023a. Gpt-4v(ision) system card.

OpenAI. 2023b. New models and developer products announced at devday.

OpenAI. 2024. Hello gpt-4o.

OpenAI. 2025. Gpt-5 system card.

Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. 2024. VALOR-EVAL: holistic coverage and faithfulness evaluation of large vision-language models. In *Proceedings of Findings of the Association for Computational Linguistics*, pages 1783–1805.

QwenTeam. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. Scifibench: Benchmarking large multimodal models for scientific figure interpretation.

In *Proceedings of Advances in Neural Information Processing Systems*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.

Michael Saxon, Fatima Jahara, Mahsa Khoshnoodi, Yujie Lu, Aditya Sharma, and William Yang Wang. 2024. Who evaluates the evaluations? objectively scoring text-to-image prompt coherence metrics with t2iscorescore (TS2). In *Proceedings of Advances in Neural Information Processing Systems*.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, and Nanqing Dong. 2025. Many heads are better than one: Improved scientific idea generation by A llm-based multi-agent system. In *Proceedings of the Annual Meeting of the Association for Computational Linguistic*, pages 28201–28240.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024. Generative multimodal models are in-context learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14398–14409. IEEE.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Ultralytics. 2023. Yolov8. https://github.com/ultralytics/ultralytics. Accessed: 2026-04-15.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. In *Proceedings of Advances in Neural Information Processing Systems*.

Yue Wu, Jiaming Liu, Maoguo Gong, Peiran Gong, Xiaolong Fan, A. Kai Qin, Qiguang Miao, and Wenping Ma. 2024. Self-supervised intra-modal and cross-modal contrastive learning for point cloud understanding. *IEEE Trans. Multim.*, 26:1626–1638.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2025. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13618–13628.

Zhijian Xu, Yilun Zhao, Manasi Patwardhan, Lovekesh Vig, and Arman Cohan. 2025. Can llms identify critical limitations within scientific research? A systematic evaluation on AI research papers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistic*, pages 20652–20706.

Cheng Yang, Chufan Shi, Yaxin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran Xu, Xinyu Zhu, Siheng Li, Yuxiang Zhang, Gongye Liu, Xiaomei Nie, Deng Cai, and Yujiu Yang. 2025. Chartmimic: Evaluating lmm's cross-modal reasoning capability via chart-to-code generation. In *Proceedings of the International Conference on Learning Representations*.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2025. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models.

Jerrold H. Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Leixin Zhang, Steffen Eger, Yinjie Cheng, Weihe Zhai, Jonas Belouadi, Fahimeh Moafian, and Zhixue Zhao. 2025a. Scimage: How good are multimodal large language models at scientific text-to-image generation? In *Proceedings of the International Conference on Learning Representations*.

Zicheng Zhang, Tengchuan Kou, Shushi Wang, Chunyi Li, Wei Sun, Wei Wang, Xiaoyu Li, Zongyu Wang, Xuezhi Cao, Xiongkuo Min, Xiaohong Liu, and Guangtao Zhai. 2025b. Q-eval-100k: Evaluating visual quality and alignment level for text-to-vision content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10621–10631.

Yilun Zhao, Weiyuan Chen, Zhijian Xu, Manasi Patwardhan, Chengye Wang, Yixin Liu, Lovekesh Vig, and Arman Cohan. 2025. Abgen: Evaluating large language models in ablation study design and evaluation for scientific research. In *Proceedings of the Annual Meeting of the Association for Computational Linguistic*, pages 12479–12491.

Bocheng Zou, Mu Cai, Jianrui Zhang, and Yong Jae Lee. 2024. Vgbench: Evaluating large language models on vector graphics understanding and generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3647–3659.

The appendix is structured as follows:

- Comparison between real captions and synthetic prompts in Section A.
- Details of related work in Section B.
  - Scientific image datasets in Section B.1.
  - Evaluation metrics in Section B.2.
- Details of training data in Section C.
- More details of HE4AFG in Section D.
  - More Examples of academic figures in Section D.1.
  - Data annotation guidelines in Section D.2.
- Details of experimental settings in Section E.
  - SRCC and statistical significance in Section E.1.
  - Computational cost in Section E.2.
  - Model versions in Section E.3.
  - Hyperparameter in Section E.4.
  - Inference details in Section E.5.
- Details of compared methods in Section F.
- Broader impact in Section G.

A Real Captions vs Synthetic Prompts

The dataset most related to HE4AFG is ScImage, which provides expert-level annotations to assess scientific text-to-image generation. However, HE4AFG uses real captions as input instructions, which originate from real scientific papers, while ScImage uses synthetic prompts as input instructions, which are created by simple and limited templates. We further compare these two types of input instructions in Table 9, suggesting the benefits of using real scientific captions.

B Details of Related Work

Scientific images are often represented in various forms, including visualization figures (Roberts et al., 2024) (e.g., line plots, flowcharts, and model diagrams), raster figures (Li et al., 2025) (consisting of pixels), and vector graphics (consisting of basic geometric elements such as points, lines, and curves) (Yang et al., 2025). For simplification, we will not strictly distinguish these types and uniformly refer to them as "scientific images".

B.1 Scientific Image Datasets

Scientific Image Captioning Datasets. Various caption-image benchmarks have been developed aimed at scientific domains. SciCap (Hsu et al., 2021) contains more than 2M figures extracted from more than 290,000 computer science arXiv articles published between 2010 and 2020. Multimodal ArXiv (Li et al., 2024), consisting of ArXivCap and ArXivQA, aims to improve the academic understanding of LMMs. ArXivCap is a caption-figure dataset comprising 6.4M figures and 3.9M captions, sourced from 572K ArXiv papers spanning various academic domains. Drawing from ArXivCap, ArXivQA, a question-answering dataset, is generated by prompting GPT-4V based on academic figures. SciFIBench (Roberts et al., 2024) is an academic figure interpretation benchmark consisting of 2,000 questions split between two tasks (question answering and figure captioning). The questions are curated from arXiv paper figures and captions, using adversarial filtering to find hard negatives and human verification for quality control. CharXiv (Wang et al., 2024) is a comprehensive evaluation suite that consists of 2,323 natural, challenging, and diverse charts from academic articles. Each chart is paired with two types of questions: (1) descriptive questions about examining basic chart elements and (2) reasoning questions that require synthesizing information across complex visual elements in the chart.

Scientific Text-to-Image Datasets. General-purpose text-to-image models have made significant strides in scientific domains. Belouadi *et al.* (Belouadi et al., 2024) proposed the use of TikZ, a well-known abstract graphics language that can be compiled into vector graphics, as an intermediate representation of scientific figures. Thus, they introduced DaTikZ, the first large-scale TikZ dataset consisting of 120k TikZ drawings aligned with captions. Vector graphics (VG), on the other hand, offer a textual representation of visual content that can be more concise and powerful for content. Zou *et al.* proposed VGBench (Zou et al., 2024), a comprehensive benchmark for LLMs on handling vector graphics through various aspects, including

Aspect	Real Captions	Synthetic Prompts
Semantic Accuracy	(a) Written by humans, accurately describing actual content; (b) More precise text-image alignment; (c) Reduces "hallucination" problems.	(a) May contain model-inferred or fabricated information; (b) May have semantic deviations; (c) Prone to generating inaccurate descriptions.
Language Quality	(a) More natural expression, following human daily language habits; (b) Human-reviewed and fewer grammatical errors; (c) Richer context, including implicit information such as academic culture, research objectives, etc.	(a) May inherit biases from training data; (b) Involves semantically ambiguous templates or style descriptions; (c) Tends to ignore diversity.
Detail Description	Usually contains specific object names and attributes, as well as accurate spatial relationship descriptions.	May use vague vocabulary leading to overgeneralization and ignore key visual details leading to information loss.

Table 9: Comparison between real captions and synthetic prompts.

visual understanding and generation, evaluation of various vector graphics formats, diverse question types, etc. ScImage is a benchmark designed to evaluate the multimodal capabilities of LLMs in generating academic figures from textual descriptions. ScImage (Zhang et al., 2025a) assesses three key dimensions of understanding: spatial, numeric, and attribute comprehension, as well as their combinations. Science-T2I (Li et al., 2025) is an expert-annotated adversarial dataset comprising adversarial 20k figure pairs with 9k prompts, covering many different categories of academic knowledge. Using Science-T2I, the authors presented SciScore, an end-to-end reward model that assesses hallucinations of academic knowledge in generated figures.

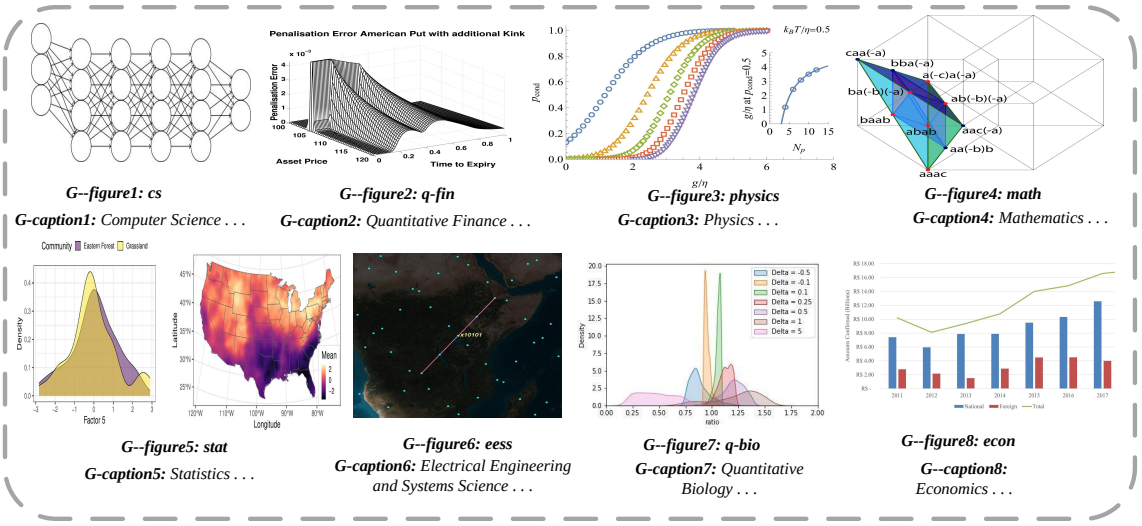
B.2 Evaluation Metrics

LMM-as-a-Judge. The emergence of LMMs has triggered extensive research in model evaluation on diverse aspects such as text-figure alignment, figure quality, complex reasoning, evaluation efficiency, and explainability, etc. Roughly most LMMs can take interleaved text and figures as input and therefore be used to evaluate text-to-figure and figure captioning tasks. Using GPT-4 as the technology behind the visual capabilities, GPT-4V (OpenAI, 2023a) is fine-tuned with additional data, using an algorithm called reinforcement learning from human feedback. GPT-4o (OpenAI, 2024) is also especially better at vision and audio comprehension compared to existing models. The Gemini family (Anil et al., 2023; Reid et al., 2024) exhibits remarkable capabilities in understanding figures, audio, video, and text, suitable for applications ranging from complex reasoning tasks to use-cases requiring limited memory on-devices. Using Qwen-LM as the foundation, Qwen-VL (Bai et al., 2023) is endowed with visual capacity by introducing a 3-stage training pipeline. With simple modifications to LLaVA, LLaVA-1.5 (Liu et al., 2024) achieves stronger baseline performance by using CLIP-ViT-L-336px and adding VQA data geared to scientific tasks.

Task-specific Metrics. Specific metrics have been proposed for text-to-image generation or image captioning tasks. For example, VALOR-Eval (Qiu et al., 2024) measures whether the generated caption is faithful to the visual content by detecting hallucinations of objects, attributes, and relationships. TIFA (Hu et al., 2023) measures the faithfulness of a generated image to its text input via visual question answering (VQA). Given a text input, it can automatically generate question-answer pairs and calculate the faithfulness of the figures by checking the accuracy of existing VQA models. Yescieval (D'Souza et al., 2025) uses LLMs as a judge to evaluate the precision of scientific inquiry tasks such as scientific QA. However, it cannot be applied directly to evaluate scientific figure generation due to the disparity between tasks.

C Details of Training Data

Source Dataset. Starting from ArXivCap (Li et al., 2024), we generate aspect-aware positive and negative examples to train HE4AFG-E. ArXivCap (Li et al., 2024) is a caption-figure dataset comprising 6.4M figures and 3.9M captions, sourced from 572K ArXiv papers spanning 32 academic domains.
Attribute Dictionary. When constructing correctness negatives, we first generate a negative caption C_C by making a minor revision to the gold caption C_G. Specifically, we adopt the attribute dictionary D from ScImage (Zhang et al., 2025a). D defines key elements relevant to scientific figures, including objects (e.g., square and circle), attributes (e.g., color and size), spatial relations (e.g., left, right,



G--figure1: cs
G-caption1: *Computer Science . . .*

G--figure2: q-fin
G-caption2: *Quantitative Finance . . .*

G--figure3: physics
G-caption3: *Physics . . .*

G--figure4: math
G-caption4: *Mathematics . . .*

G--figure5: stat
G-caption5: *Statistics . . .*

G--figure6: eess
G-caption6: *Electrical Engineering and Systems Science . . .*

G--figure7: q-bio
G-caption7: *Quantitative Biology . . .*

G--figure8: econ
G--caption8: *Economics . . .*

Figure 7: Figure examples for academic domains used in HE4AFG.

between), and numeric values (e.g., three, five, two more). When selecting a descriptive word from D, we first locate the specific list corresponding to the word class and then randomly choose an item from it. We define a set of query templates, where each template is one or several sentences with one or more placeholders. These placeholders are designed to be filled with elements from D, which may include objects, attributes, or relations. To construct a minor revision to C_G, we select a query template and populate its placeholders with appropriate attributions from D. Finally, the revised C_G is recorded as the negative caption C_C.

D More Details of HE4AFG

D.1 More Examples of Academic Figures

As Fig.7 shows, we provide a figure example for each domain used in HE4AFG.

D.2 Data Annotation Guidelines

Human Scoring Criteria. As Table 10 shows, we provide detailed scoring guidelines used in HE4AFG. We also calculate the averaged weighted kappa value for all sampled instances and get a high score of 0.81, demonstrating good agreement between data specialists.

E Details of Experimental Settings

E.1 SRCC and Statistical Significance

For all scoring dimensions, we use Spearman Ranking Correlation Coefficients (SRCC) (Zar, 2005), which measures the monotonic relationship between two variables on the basis of their ranked values. Specifically, SRCC calculates how well automated scores ($x = (x_1, ..., x_n)$) correlate with human judgments ($y = (y_1, ..., y_n)$). A stronger correlation indicates greater agreement between prediction and reference.

$$\text{SRCC} = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \quad (1)$$

where $d_i = r_{x_i} - r_{y_i}$, and r_{x_i}, r_{y_i} denote the rank of x_i, y_i.

To show **statistical significance**, we also report the P-value (probability value), which is a number describing the likelihood of obtaining the observed data under the null hypothesis of a statistical test.

E.2 Computational Cost

All experiments of our models were completed within 3 hours split between training (2.5 hours) and inference (0.5 hours) on 2 NVIDIA A100 40GB GPUs.

E.3 Model Versions

As Table 11 shows, we use the specific versions or Hugging Face checkpoints of all experimental models.

E.4 Hyperparameter

We provide an overview of hyper-parameter settings used during training. Key parameters, including batch size, learning rate, and optimizer configurations, are summarized in Table 12.

Score	Description
	Topic Relevance
5	The figure contains no redundant objects (e.g., geometric shapes, OCR, and data points, etc).
4	The figure contains a few redundant objects but remains highly relevant to the text's requirements.
3	The figure contains some redundant objects and some required elements.
2	The figure contains more redundant objects than required elements.
1	The overall figure is not relevant to the requirements.
	Attribute Correctness
5	The figure fully meets all the requirements (objects, attributes, relations, numbers) with no mistakes.
4	The figure meets the key requirements (objects, attributes, relations, numbers), with only minor mistakes.
3	The figure meets some or half of the requirements (objects, attributes, relations, numbers), with some mistakes.
2	The figure meets only a few of the requirements (objects, attributes, relations, numbers) and contains serious mistakes.
1	The figure fails to meet the requirements (objects, attributes, relations, numbers) of the text.
	Scientific Aesthetic
5	The figure fully meets all the requirements (clean, clear, vector-style) with no mistakes.
4	The figure meets the key requirements (clean, clear, vector-styles), with only minor mistakes.
3	The figure meets some or half of the requirements (clean, clear, vector-style), with some mistakes.
2	The figure meets only a few of the requirements (clean, clear, vector-style) and contains serious mistakes.
1	The figure fails to meet the requirements (clean, clear, vector-style) of the text.

Table 10: Human scoring guidelines for data annotation of HE4AFG.

Model	**Version / HF Checkpoint**
	Closed-source LMMs
GPT-4V (OpenAI, 2023a)	*gpt-4-vision-preview*
GPT-4 Turbo (OpenAI, 2023b)	*gpt-4-turbo-2024-04-09*
GPT-4o (OpenAI, 2024)	*gpt-4o-2024-05-13*
Gemini Pro Vision (Anil et al., 2023)	*gemini-pro-vision*
Gemini 1.5 Pro (Reid et al., 2024)	gemini-1.5-pro-001
Claude 3 Haiku (Anthropic, 2024)	*claude-3-haiku@20240307*
Claude 3 Opus (Anthropic, 2024)	*claude-3-opus-20240229*
	Open-source LMMs
IDEFICS-9b-Instruct (Laurençon et al., 2023)	*huggingfaceM4/idefics2-8b*
IDEFICS-80b-Instruct (Laurençon et al., 2023)	*huggingfaceM4/idefics2-80b*
Emu2 (Sun et al., 2024)	*BAAI/Emu2*
InternLM-XComposer-7b (Team, 2023)	*InternLM/InternLM - XComposer - 7b*
InstructBLIP-FlanT5-xl (Dai et al., 2023)	*salesforce/instructblip- flan-t5-xl*
InstructBLIP-Vicuna-7b (Dai et al., 2023)	*salesforce/instructblip- vicuna-7b*
InstructBLIP-Vicuna-13b (Dai et al., 2023)	*salesforce/instructblip-vicuna-13b*
Qwen-VL-Chat (Bai et al., 2023)	*Qwen/Qwen-VL-Chat*
	Our Models
HE4AFG-E (SCLIP)	*openai/clip-vit-base-patch32*
HE4AFG-R (SFT)	*mplugOwl3-8B*

Table 11: Running configurations for all models.

Hyper-parameters	HE4AFG-E
batch size	64
learning rate	2 x 10^{-6}
learning rate schedule	cosine
weight decay	0.3
training steps	500
warmup steps	150
optimizer	AdamW

Table 12: Hyper-parameter settings.

E.5 Inference Details

Diverse score scales. For convenience in human scoring, we use the 5-Point Likert scale ({1, 2, 3, 4, 5}) for data annotation. However, when examining LMMs as evaluators, we require them to produce a score between 0 and 1, so that the predicted results can exhibit more subtle differences. Therefore, we first normalize all human scores in the range [0, 1] using the Min-Max method (Jain et al., 2005) and then calculate SRCC using normalized human scores and direct model scores.

Diverse model outputs. Each LMM is asked to output a relevance score, a correctness score, and an aesthetic score. However, some metrics (e.g., CLIPScore, VQAScore, BLIP2Score, and TIFA) can only generate a total faithfulness score. As usual, we use these merged scores as a hypothetical score regarding some dimension (TR, AC, SA) to calculate SRCC.

F Details of Compared Methods

In this paper, we compare HE4AFG-E with 19 competing metrics, which can be divided into the following categories.

(1) Closed-Source LMMs. Using GPT-4 as the technology behind the visual capabilities, GPT-4V (OpenAI, 2023a) is fine-tuned with additional data, using an algorithm called reinforcement learning from human feedback (RLHF) to produce outputs that are preferred by human trainers. GPT-4 Turbo (OpenAI, 2023b) supports a 128K context window and new multimodal capabilities on the platform, including vision, figure creation, and text-to-speech (TTS). GPT-4o (OpenAI, 2024) matches the performance of GPT-4 Turbo in English and code text, while being much faster and cheaper in the API. GPT-4o is especially better at vision and audio comprehension compared to existing models. The Gemini family (Anil et al., 2023) exhibits remarkable capabilities in understanding figures, audio, video, and text, suitable for applications ranging from complex reasoning tasks to use-cases requiring limited memory on-devices. In particular, Gemini 1.5 Pro (Reid et al., 2024) represents the next generation of highly compute-efficient multimodal models capable of recalling and reasoning over fine-grained information from millions of tokens of context, including multiple long documents and hours of video and audio. The Claude 3 series includes three advanced models, each designed to excel in various cognitive tasks and applications. For example, Claude 3 Opus (Anthropic, 2024) demonstrates near-human comprehension and fluency in complex tasks, particularly effective in nuanced content creation, code generation, and multi-language conversations, while Claude 3 Haiku (Anthropic, 2024) is the fastest and most cost-effective in its category, making it ideal for rapid information retrieval and analysis.

(2) Open-Source LMMs. IDEFICS (9/80B) (Laurençon et al., 2023) is the vision and language model trained on the OBELICS dataset, an open web-scale filtered dataset of interleaved figure-text documents comprising 141 million web pages extracted from Common Crawl. Emu2 (37B) (Sun et al., 2024) is trained on large-scale multimodal sequences with a unified auto-regressive objective, showing strong multimodal in-context learning abilities. InternLM-Composer (104B) (Team, 2023) is pre-trained on a large corpora with 1.6T tokens with a multi-phase progressive process, and then fine-tuned to align with human preferences. InstructBLIP (Dai et al., 2023) conducts vision-language instruction tuning based on the pretrained BLIP-2 models and introduces the instruction-aware Query Transformer, which extracts informative features tailored to the given instruction. Using Qwen-LM as the foundation, Qwen-VL (Bai et al., 2023) is endowed with visual capacity by introducing a 3-stage training pipeline.

(3) Text-Image Alignment Metrics. We also compare with fundamental text-image alignment models. For instance, CLIPScore (Hessel et al., 2021) is a widely-used evaluation metric that measures the alignment between an figure and text prompt by calculating the cosine similarity of the figure and text embeddings. With its tight focus on figure-text compatibility, CLIPScore is complementary to existing reference-based metrics that emphasize text-text similarities. Similarly, BLIP2Score (Li et al., 2023) is based on the BLIP2 text-figure alignment. However, recent studies reveal some limitations of CLIPScore, e.g., it fails to produce reliable scores for complex prompts involving compositions of objects, attributes, and relations. To address this, VQAScore (Lin et al., 2024) is introduced by using a visual-question-answering (VQA) model to produce an alignment score by computing the probability of a "Yes" answer to a simple "Does this figure show [text]?" question. Its in-house model, CLIP-FlanT5, outperforms even the strongest baselines that make use of the proprietary GPT-4V. Specific metrics are also proposed for text-to-figure generation or figure captioning tasks. For instances, TIFA (Hu et al., 2023) measures the faithfulness of a generated figure to its text input via VQA. Specifically, given a text input, it can automatically generate question-answer pairs and calculate figure faithfulness by checking the accuracy of existing VQA models.

G Broader Impact

Theoretically, this work represents a fundamental yet emerging research direction aimed at developing automatic faithfulness evaluation metrics for academic figure generation and interpretation, which is highly significant for understanding the alignment between multimodal content and academic semantics in the fields of computer vision (CV) and natural language processing (NLP). Since human evaluation of academic caption-figure alignment is both costly and time-consuming, it is crucial to establish reliable and interpretable automatic evaluation metrics that strongly correlate with human judgments. However, existing faithfulness metrics—such as CLIPScore, TIFA, and

VALOR-EVAL—are primarily designed for natural figures and lack the specificity required for academic domains. Moreover, large multimodal models (LMMs) like GPT-4o, although powerful, are often too large and expensive for widespread use. To address these limitations, we propose HE4AFG, a high-quality evaluation benchmark that incorporates academic domain-aware training via contrastive learning and rationale generation. To our knowledge, our work first proposes a novel evaluation dataset & model specifically designed for academic figure tasks and encourages future work to improve evaluation in academic domains.

Practically, HE4AFG addresses the critical need for a domain-specific, reliable, and actionable evaluation base in academic multimodal AI. First, it provides fine-grained assessment, allowing researchers to pinpoint specific inaccuracies—such as misrepresented data trends or incorrect object attributes—in model-generated content. Second, by leveraging lightweight adapted encoders and a compact LMM for feedback generation, HE4AFG-E achieves high performance with significantly lower computational cost than large-scale evaluators such as GPT-4o, making it feasible for widespread use in resource-constrained environments. Third, extensive validation on HE4AFG shows that our model correlates more closely with human judgment than 19 competing metrics, confirming its reliability as an automatic and human-aligned evaluator. Finally, the dataset is highly flexible and scalable: it supports easy adaptation to new academic domains, integration with various backbone models, and expansion with additional training data, positioning it as a practical and sustainable solution for evolving academic multimodal tasks.