HealthCards: Exploring Text-to-Image Generation as Visual Aids for Healthcare Knowledge Democratizing and Education

Anonymous ACL submission

Abstract

The evolution of text-to-image (T2I) generation techniques has brought new capability for information visualization, and this advancement could have the potential to boost knowledge democratization and educational equity. In this paper, we envision these technologies as powerful tools to promote accessible healthcare knowledge education, which could not only serve the public but also more beneficial for communities in underserved regions and people with specific disabilities such as reading ability and attention limitations. We first explore how to harness recent T2I models to generate health knowledge flashcards, which are educational aids that aggregate knowledge with visually appealing and concise presentations in an image. Then, we curated a diverse and high quality healthcare knowledge flashcards datasets with 2034 samples from credible knowledge resources. We also validate the effectiveness of fine-tuning open-sourced models with our dataset to serve as a promising health flashcards generator. Our code is available at Anonymous Github: https://anonymous.4open.science/r/HealthCards

1 Introduction

Text-to-image (T2I) generation advancements, such as Stable Diffusion (Rombach et al., 2022), Flux (Labs, 2024), and GPT-4o (Hurst et al., 2024), have empowered users to create content with unprecedented aesthetic expression while offering fine-grained control and personalization. This great advancement also foster valuable applications in other domains, such as design ideation (Paananen et al., 2024), K12 Education (Ali et al., 2024) and Medicine (Mai et al., 2024).

Notably, the recent model from OpenAI, *GPT-4o-Image* (OpenAI, 2025), which was first released in late March 2025, has sparked widespread global discussion and adoption for its unprecedented capability to generate highly coherent and contextually relevant visual content from natural language instructions with remarkable detail fidelity and compositional accuracy (Yan et al., 2025; Cao et al., 2025). Among its global popularity, an eventful and interesting application is "*Ghibli-style images*", a style inspired by the works of Studio Ghibli, characterized by hand-drawn aesthetics, vibrant natural landscapes, and emotionally rich characters. Users simply provide text prompts describing scenes, people, or concepts they want to visualize, and the model generates images that faithfully capture the distinctive Ghibli aesthetic—transforming ordinary descriptions into dreamlike illustrations that evoke the studio's signature warmth and fantastical elements. Social media platforms like *X* and *Instagram* have been flooded with Ghibli-style images, across personal photos transformed into this aesthetic, creative memes, reimagined movie stills, and entirely new fictional scenarios, highlighting its viral appeal (Di Placido, 2025).

This global phenomenon could reveal a profound public fascination with visually appealing, customizable content that effectively carries both information and emotion. The widespread adoption motivates us to explore more substantive applications: *(1) Can we harness text-to-image generation technology to create knowledge disseminators with more educational value? And (2) How effectively can privacy-friendly open-source models perform in these educational contexts?*

For these questions, in this work, we try to start approaching it in the field of public health education. As healthcare knowledge could sometimes remain inaccessible due to its technical complexity, especially for undereducated populations in less developed regions and people with reading disabilities all around the world (Shahid et al., 2022; Gréaux et al., 2023). We expect this kind of new technique could help popularize health knowledge and promote equity in health education.

Following this line of thinking, which kind of

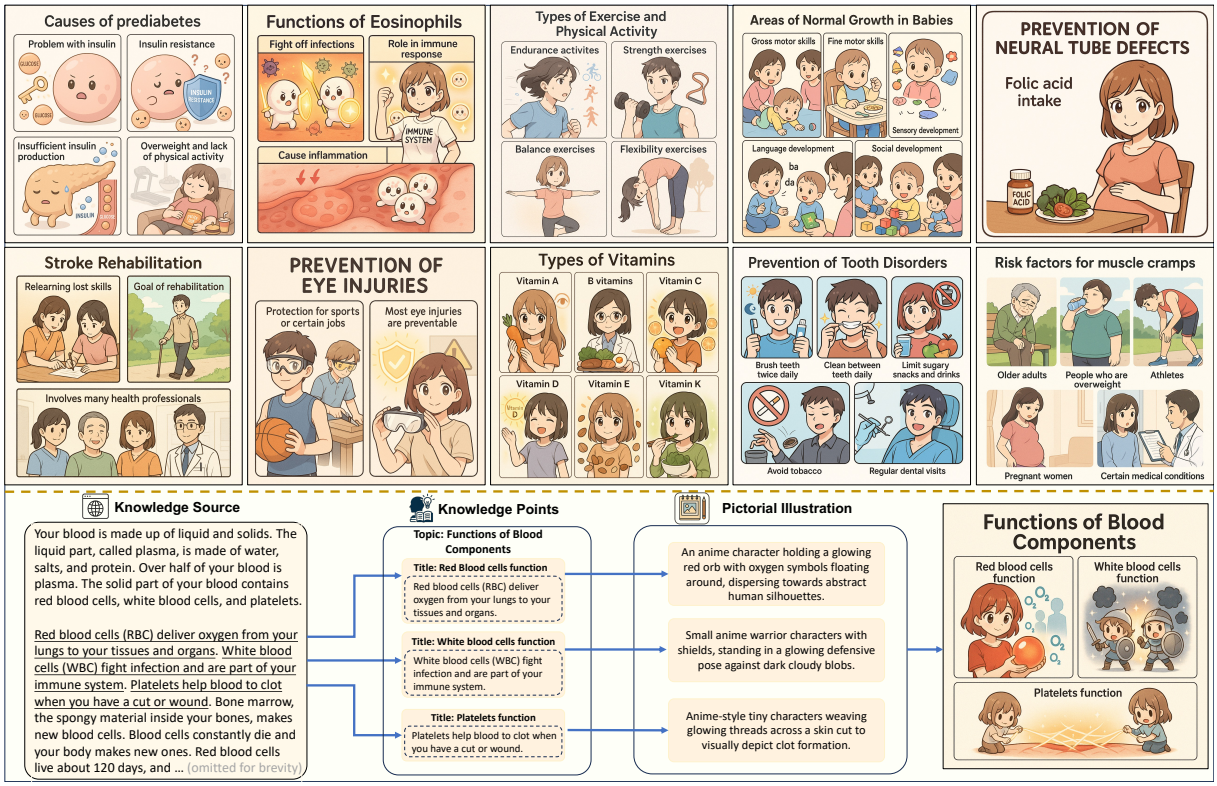

Figure 1: Samples of our *HealthCards* dataset illustrating diverse health topics with anime-style visualizations. The cards feature structured layouts with clear topics, knowledge point headings, and corresponding visual illustrations that effectively represent key health concepts in an accessible format. Below, a simplified data creation pipeline demonstrates how credible health knowledge is extracted, converted into pictorial illustrations, and composed into visually engaging educational content via a text-to-image generation model. These tiny cards carry our vision that such automatic generation of educational flashcard-like knowledge presentations could serve as a new potential for health knowledge democratization (offering more accessible and personalized health information for the public with diverse backgrounds), education (stimulating learners to actively recall knowledge and make conceptual connections), and beyond.

visual media with what designs are desired for this research and could be valuable for the community? After investigation, a classic learning tool, *Flash-Cards (Wikipedia contributors, 2025; Logopsycom, 2020)* come to our vision. Flashcards are efficient learning tools that organize information in a way that prompts active recall and memorization. They present concise, focused content that can usually combine text with visual elements such as images and diagrams for better comprehension and retention.

Long recognized as an effective educational approach, flashcard-based learning facilitates independent study and reinforces understanding of key concepts. This versatility makes flashcards widely adopted by educators, students, and parents across various disciplines (Santika et al., 2023; Tirtayani et al., 2017). For instance, flashcards are particularly effective in vocabulary learning, as demonstrated by educational tools like *Duolingo*(Teske, 2017) and *4-pics-1-word*(Fithriani, 2023), which incorporate relevant or deliberately confusing images to aid topic retention (Loewen et al., 2019). In healthcare education, flashcards are similarly valuable. Tools like *ANKI* (Anki, 2025) are widely used by medical students to create personalized flashcards summarizing complex health knowledge points, such as drug recommendations and precautions of disease. Studies show that these tools enhance learning outcomes and knowledge retention (Lu et al., 2021; Harris et al., 2022). Building on this foundation, we envision that automating the design of flashcards using text-to-image generation could further facilitate this process and democratize healthcare knowledge, particularly for underserved communities and individuals with reading disabilities or attention limitations.

In this work, our contributions could be summarized as follows:

(1) To the best of our knowledge, we are the first

to systematically explore the potential of text-to-image generation models for creating educational health flashcards. We devise problem formulations and design customized pipelines, together with detailed evaluation methodologies for this task. This research is especially timely given the recent surge in interest around GPT-4o's image generation capabilities and the viral spread of Ghibli-style imagery, demonstrating both the technical feasibility and public enthusiasm for engaging visual knowledge representation that could make complex healthcare knowledge more accessible to diverse populations.

(2) We curated *HealthCards*, a high-quality health knowledge flashcard dataset comprising 2,034 samples, each carefully selected and verified by medical experts. This dataset encompasses diverse health topics sourced from credible healthcare resources and features varied knowledge types, with differing numbers of knowledge points per sample. This diversity makes it a valuable resource for developing AI systems capable of generating accurate visual health education content.

(3) Our extensive experiments reveal that open-source text-to-image models can be effectively fine-tuned with our dataset to generate high-quality health education flashcards. The performance is validated by detailed automatic metrics and human evaluation. Importantly, we identify an *evenness bias* among existing T2I models, where they preferentially generate an even number of subfigures for symmetrical visual aesthetics. Models fine-tuned with our dataset demonstrate significant mitigation of this bias and achieve precise subfigure generation that accurately follows the content requirement.

2 Related Works

2.1 Text-to-image Generation

Text-to-image (T2I) generation techniques have achieved unprecedented progress in recent years through pioneering works like *Stable Diffusion*(Rombach et al., 2022) series, *FLUX*(Labs, 2024) series, DALLE (Ramesh et al., 2021), Midjourney and the more recent GPT-4o-Image (Hurst et al., 2024). Research efforts have focused on enhancing and evaluating T2I models across multiple dimensions, including visual-textual alignment (Huang et al., 2025), variability (Tang et al., 2024), generalizability (Cao et al., 2024), and specialized capabilities such as text rendering (Zhao et al., 2025) and scientific knowledge alignment (Li et al., 2025). These advancements have enabled valuable healthcare applications, from visualizing anatomical structures (Noel, 2024) and generating realistic surgical images (Nwoye et al., 2025) to creating visual aids that bridge communication gaps and enhance patient understanding of medical conditions (Goparaju, 2024).

2.2 Visual aids for Health Education

Visual aids are critical in health education, enhancing understanding and retention, particularly for individuals with low literacy. Studies show that visual aids improve health literacy, medication adherence, and comprehension (Mbanda et al., 2021). The AHRQ (Agency for Healthcare Research and Quality, 2020) recommends visuals to simplify complex information, such as using graphics to clarify portion sizes or numerical data. A review (Lee and Nathan-Roberts, 2021) confirms that visual aids positively impact adherence, comprehension, and recall when supplementing written or spoken medical instructions. Additionally, visuals benefit diverse demographics by enhancing accessibility, with WebMD Ignite (Ignite, 2022) highlighting how they promote health equity through inclusive information.

Flashcards are widely used in medical education for active recall and spaced repetition. Platforms like Anki and Brainscape offer extensive medical flashcard collections (Anki, n.d.; Brainscape, n.d.). Research demonstrates their effectiveness, with studies showing improved engagement and knowledge retention (Sun et al., 2021).

3 Problem Formulation

Drawing from real-world flashcard design principles (31Memorize, 2024) and multimedia learning (Mayer, 2005), we try to formalize our task of HealthCards as a specialized text-to-image generation problem. Specifically, we model a *health flashcard* HC as a structure with four components:

$$HC = (\tau, K, \Phi, I) \tag{1}$$

where $\tau \in \mathcal{T}$ denotes the topic of the flashcard as text (e.g., *"Prevention of Hypertension"*). $K = \{k_i\}_{i=1}^{m} \subset \mathcal{K}$ represents a knowledge set where each element $k_i = (h_i, c_i, p_i) \in \mathcal{H} \times \mathcal{C} \times \mathcal{P}$ consists of a concise heading $h_i \in \mathcal{H}$, a knowledge content $c_i \in \mathcal{C}$, and a pictorial illustration description $p_i \in \mathcal{P}$ for the knowledge point. $\Phi = \phi(\tau, K)$ constitutes a synthesized generation prompt that guides

a text-to-image generation model $g : \Phi \rightarrow \mathcal{I}$ in producing a healthcard image $I = g(\Phi) \in \mathcal{I}$. The symbolic representation described above is illustrated in Figure 2(A).

The generated health card image I is expected to satisfy the following properties:

Property 1 (Text Rendering)**.** *Let* $\mathsf{T} : \mathcal{I} \rightarrow 2^{\mathcal{V}}$ *be a function that extracts all text elements from an image, where* $\mathcal{V}$ *is the set of all possible text strings. For any valid health card image* I*:*

$$\mathsf{T}(I) = \{\tau\} \cup \{h_i \mid i \in \{1, \ldots, m\}\} \quad (2)$$

That is, the rendered text must precisely consist of the topic τ *and all knowledge headings* $\{h_i\}_{i=1}^{m}$*.*

Property 2 (Subfigure Counting)**.** *Let* $|\mathcal{S}(I)|$ *denote the number of subfigures in image* I*, and* $|K|$ *denote the number of knowledge points. A valid health card image must satisfy:*

$$|\mathcal{S}(I)| = |K| = m \quad (3)$$

ensuring the number of generated subfigures exactly matches the number of knowledge points required.

Property 3 (Subfigure Content Alignment)**.** *For a health card image* I *with* m *identified subfigures* $\mathcal{S}(I) = \{s_1, s_2, \ldots, s_m\}$ *and knowledge set* $K = \{k_i = (h_i, c_i, p_i)\}_{i=1}^{m}$*, there should exist a correspondence between subfigures and knowledge points such that each subfigure visually represents the content described in its corresponding pictorial description. This visual-semantic alignment is necessary for effective knowledge presentation through the health flashcard.*

4 Healthcards Dataset

4.1 Construction Pipeline

We sourced public health knowledge from the NIH *MedlinePlus* website[1], a public platform supported by the NIH that provides health information across various categories including Body Location, Disorders, Conditions, Diagnosis, Therapy, and Demographic Groups. It serves as a credible health knowledge resource for the public. As of our data collection on April 5th, 2025, the platform contained 1,017 health topics. The website provides packed resources for research purposes and free public access.

[1]https://medlineplus.gov/

Our construction pipeline can be formalized as a sequence of transformations:

$$E \xrightarrow[extract]{f(prompt1;E)} (\tau, K) \xrightarrow[compose]{\phi(\tau,K)} \Phi \xrightarrow[generate]{g(\Phi)} I$$

where E represents the raw essay from MedlinePlus, f is our knowledge extraction function implemented using *GPT-4o* (Hurst et al., 2024) with few-shot examples that produces the topic τ and knowledge set K (the *prompt* 1 is presented in Appendix Table 2). After that, $\phi(\tau, K)$ synthesizes these elements into a generation prompt Φ, a demonstration of this process is shown below:

Prompt Composition: $\phi(\tau, K) \rightarrow \Phi$

Design a clean, 1:1 aspect-ratio health flashcard with anime-style illustrations for **{m}** knowledge points. Use a **{layout type}** layout (**{description}**). Each sub-figure must contain a concise title and an illustration cue, focusing on clarity, readability, and an approachable style. Flashcard topic: "**{τ}**".

For each $k_i \in K$ ($i = 1, \ldots, m$):
Subfigure **{i}** (**{position}**): "**{$k_i.h$}**"
- Illustration: **{$k_i.p$}**.

The elements in **{$*$}** denote the contents in knowledge set K that is extracted by the previous step. m is the number of knowledge points in the set. Those in **{$*$}** represent layout type (e.g. *two-layer grid layout*), layout description (e.g. *two above, two below*) and subfigure position (e.g. *bottom left*). We provide some examples in Appendix Table 3.

Finally, g represents the text-to-image model (specifically *GPT-4o-Image*) that leverages the composed prompt Φ to generate a healthcard image I. Through this process, we generated 2,693 images for our initial dataset, reserving 861 prompts for testing purposes.

4.2 Experts Verification

Even though *GPT-4o-Image* has achieved unprecedented text-to-image generation capability and could be the most powerful model at the time of this manuscript submission, we observe that its performance in generating HealthCards remains unstable. We categorize some common shortcomings: incorrect layouts, inappropriate visual presentations, and inaccurate medical knowledge. Examples of these issues are provided in Appendix Figure 6. Therefore, conducting meticulous quality reviews is crucial to ensure high-quality data collection, which can subsequently serve as a valuable fine-tuning resource for local models.

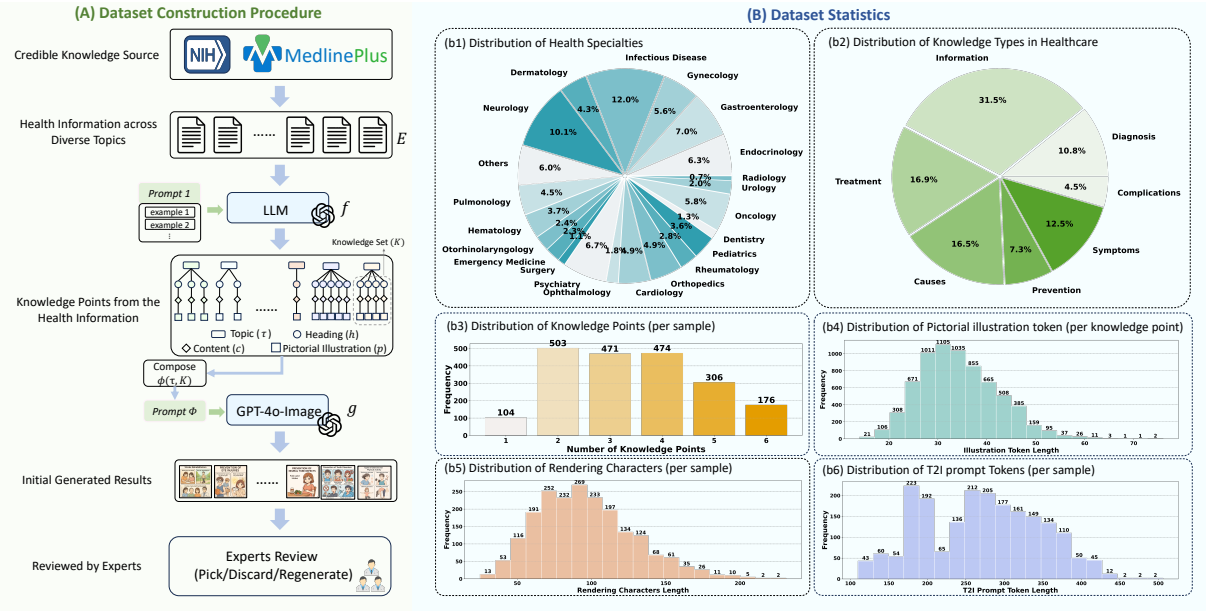

Figure 2: The Construction Procedure and Statistics of our *HealthCards* Dataset.

According to our problem formulation in Section 3, we need to manually review whether three properties (*text rendering*, *Subfigure Counting*, and *Subfigure Content Alignment*) are satisfied. For this purpose, two medical postgraduate students manually verified the generated images and provided next-action labels (*Accept*, *Regenerate*, *Discard*, *Prompt Modification*, *CheckNeeded*) for each image. After that, all the results were again reviewed by a senior doctor with PhD degree. All the annotation procedures were performed on our specially developed platform, as shown in Appendix Figure 7. We provide more detailed description such as instructions and usage of the platform in Appendix A. During this phase, a final number of 2034 images has been verified which constitute our *HealthCards* dataset.

4.3 Data Statistics

Our HealthCards dataset demonstrates considerable diversity across multiple dimensions. Figure 2 (b1) illustrates the distribution of medical specialties, encompassing 22 distinct categories, which collectively span most healthcare domains. Figure 2 (b2) depicts the corresponding knowledge types (such as Treatment and Diagnosis) represented within the dataset.

In Figure 2 (b3), we present the distribution of knowledge points per sample, which ranges from 1 to 6 points—a range sufficient for most healthcare educational contexts. Samples containing more knowledge points were strategically segmented to ensure all entries fall within this optimal range. Figure 2 (b4) displays the token length distribution of pictorial illustrations per knowledge point, with a maximum constraint of 77 tokens to maintain compatibility with standard text encoders (e.g., CLIP (Radford et al., 2021)).

The distribution of composed text-to-image prompts is shown in Figure 2 (b6), with lengths carefully controlled to remain under 512 tokens, ensuring compatibility with some text encoders such as T5 (Raffel et al., 2020). Finally, we present the character count distribution for each sample's rendered text in (b5), calculated as the combined length of the topic and all associated headings, which provides insight into the textual density of our HealthCards.

5 Finetuning and Inference

Flow matching models based on diffusion transformer (DiT), such as Stable Diffusion 3.5 (Rombach et al., 2022) and Flux (Labs, 2024), have significantly outperformed traditional diffusion models in terms of prompt adherence, visual quality and sampling speed. Based on these advantages, we select this class of conditional Flow matching models as a base model and describe its fine-tuning and inference strategies[2].

Training objective. We adopt conditional flow-matching (CFM) (Lipman et al., 2022, 2024). For

[2]For these concepts, we provide a more comprehensive recapitulation in the Appendix B.

$t \sim \mathcal{U}[0,1]$, prompt y, and samples $x_0 \sim p(\cdot \mid y)$, $x_1 \sim q(\cdot \mid y)$, the loss can be expressed as:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,x_0,x_1}\Big[\big\|v_\theta^t(x_t \mid y) - (x_1 - x_0)\big\|_2^2\Big]$$

where $x_t = (1-t)\,x_0 + t\,x_1$ and v_θ^t is a neural velocity field with trainable parameters θ.

Inference with classifier-free guidance (CFG). Following Zheng et al. (2023); Ho and Salimans (2022), during inference phase, conditional and unconditional velocities are linearly mixed:

$$\widehat{v}_\theta^t(x \mid y) = (1-\omega)\,v_\theta^t(x \mid \varnothing) + \omega\,v_\theta^t(x \mid y), \omega \geq 1, \tag{4}$$

where ω is the guidance scale ($\omega = 1$: no guidance; $\omega > 1$: stronger adherence to y).

6 Experiments

6.1 Experiment and Evaluation Settings

Aligned with our problem formulation in Section 3, the experiments were designed to evaluate each HealthCard image generated by a T2I model from the following aspects:

(1) Text Rendering Score. We harness an OCR model (PaddleOCR V4 (Contributors, 2025)) to extract all identifiable words and characters. Then, all words are first grouped according to their layout positions, and in each group, they are reordered with human reading sequence (from left to right, from up to down). After that, we obtain all processed detected strings. They are compared with the expected detected string, which is a set of Flashcard topic τ plus all subheadings h. Then, following the common practice of NLP, we compute the Word Error Rate (WER) and Character Error Rate (CER) for each paired detected string and ground truth string. We provide a more detailed explanation of this score in Appendix C.1.

(2) Subfigure Counting Score. we measure whether the number of generated subfigures in a generated image aligns with the number of knowledge points for that case. For this purpose, we harness a powerful vision language model, *Qwen2.5-VL-32B* (Bai et al., 2025). It has an innate layout detection capability which is capable of this task without further training, and this helps reduce the risk of overfitting to a certain layout which could be unfair for some models. We compute the absolute subfigure count gap (Gap) and subfigure count error rate for this score. More details are illustrated in Appendix C.2.

(3) Subfigure Alignmnet Score. We evaluate the semantic alignment between each subfigure and its corresponding pictorial illustration description. This provides fine-grained assessment of content representation quality. For this dimension, we employ CLIPScore (Hessel et al., 2021) and VQAScore (Lin et al., 2024) as measurement tools. Comprehensive evaluation procedures are detailed in Appendix C.3.

General Scores: Beyond the three dimensions directly addressing our problem formulation, we also assess general image quality and aesthetic properties using *Q-align*. We also computed the global image and T2I prompt alignment score with the VQAScore. (As token lengths of our T2I prompts exceed 77, as reflected in Figure 2 (b6), metrics like CLIPScore are not feasible here).

6.2 Implementation Details

All open-source models were trained and tested on a server equipped with 2 H100 GPUs. The code and annotation platform are available via the anonymous link provided in the Abstract for further details. Our experiments employed multiple state-of-the-art models: Stable Diffusion 3.5 Medium & Large (Rombach et al., 2022), FLUX.1-dev (Labs, 2024), Lumina (Qin et al., 2025), and HiDream-I1-Full (HiDream-ai, 2025). All fine-tuning was performed using LoRA (Hu et al., 2022) with rank 128 unless otherwise specified, trained for 10,000 steps with a batch size of 4. Classifier-free Guidance scale was set to 5 during inference by default. Commercial models including Flux.1 [Pro] (Black Forest Labs, 2025), DALLE3 (Ramesh et al., 2021), and GPT4o-Image (OpenAI, 2025) were accessed through their official APIs.

6.3 Results

We present the quantitative results in Table 1 and qualitative comparisons in Figure 5. GPT4o-Image demonstrates superior performance in *text rendering* and *Subfigure Alignment*, yet exhibits a significant limitation in accurately generating the requested number of subfigures (approximately 14% of total cases). As shown in Figure 5 (a-5), the generated healthcard fails to match the expected 5 knowledge points. Further investigation reveals an "Evenness Bias" in GPT4o-Image. The confusion matrix in Figure 3 illustrates that subfigure counting errors predominantly occur in healthcard generation tasks with odd numbers of subfigures

Model	Text Rendering		Subfigure Counting		Subfigure Alignment		General		
	WER↓	CER↓	Gap↓	Error Rate↓	CLIPScore$_s$↑	VQA$_s$↑	Aesthetic↑	Quality↑	VQA↑
DALLE3	0.827	0.603	0.552	0.387	0.236	0.705	3.520	4.568	0.786
SD3.5 Medium	1.044	0.794	0.626	0.474	0.221	0.696	3.103	3.966	0.761
SD3.5 Large	0.694	0.485	0.585	0.407	0.221	0.630	3.547	4.525	0.791
Lumina2.0	1.309	1.029	0.569	0.425	0.209	0.699	3.144	4.150	0.674
HiDream-I1	0.429	0.342	0.681	0.499	0.237	0.647	3.383	4.671	0.821
Flux.1[Dev]	0.640	0.529	0.352	0.348	0.251	0.758	3.533	**4.674**	0.785
Flux.1[Pro]	0.833	0.665	0.626	0.479	0.249	0.753	**3.654**	4.669	0.774
GPT4o-Image	**0.067**	**0.052**	0.143	0.141	**0.270**	**0.903**	3.357	4.527	**0.838**
SD3.5 Medium*	0.853	0.545	0.172	0.170	0.227	0.731	3.376	4.399	0.808
SD3.5 Large*	0.387	0.222	0.033	0.027	0.241	0.760	3.441	4.565	0.832
Lumina2.0*	1.140	0.915	0.511	0.442	0.211	0.741	3.395	4.444	0.590
HiDream-I1*	0.378	0.288	0.121	0.100	0.229	0.672	3.469	4.631	0.821
Flux.1[Dev]*	0.118	0.064	**0.010**	**0.008**	0.265	0.861	3.429	4.636	0.838

Table 1: Quantitative results on test prompts. *indicates models fine-tuned with our *HeathCards* dataset. **Bold** indicates the best performance, and underlined the second-best.

(e.g., 3 and 5), with results often defaulting to even numbers (4 and 6). This suggests an inappropriate preference for visual symmetry at the expense of content accuracy. While this bias is also observed in the raw Flux.1[Dev] model (Figure 5 (c3 and c5)), our fine-tuned version significantly mitigates this issue. This could indicate the potential value of our dataset for improving healthcare educational materials, where accurate representation of medical knowledge points is critical for effective learning and patient education.

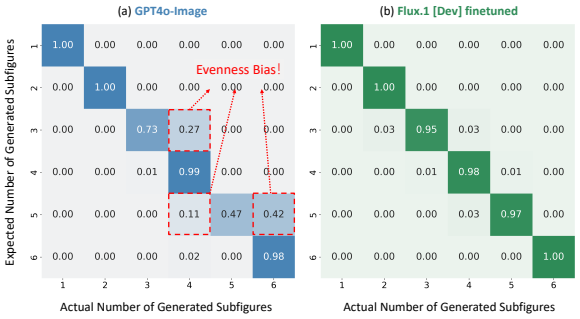

Figure 3: Confusion Matrices Comparing Expected and Actual Number of Generated Subfigures. (a) Confusion matrix for GPT-4o-Image, with red dotted boxes highlighting the *evenness bias*, where the model tends to preferentially generate an even number of subfigures. (b) Confusion matrix for Flux.1[Dev], fine-tuned on our *HealthCards* dataset, demonstrating a significant reduction in confusion compared to GPT-4o-Image.

While all fine-tuned models significantly outperform their raw counterparts, Flux.1[Dev] emerges as the most promising among all open-source models evaluated. However, it still falls short in text rendering and subfigure alignment, as shown in Table 1. To better understand human preferences regarding these shortcomings and to assess the validity of our evaluation methods, we conducted a human evaluation study.

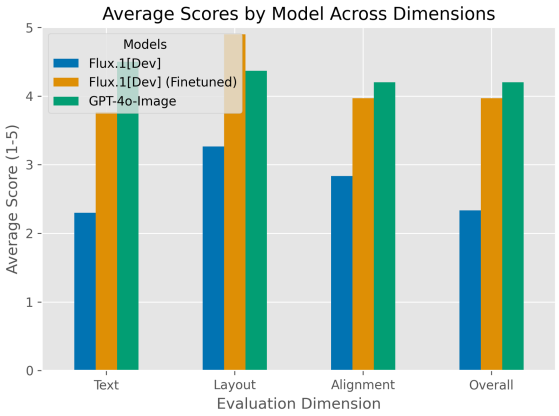

Figure 4: The results of Human Evaluation Study

6.4 Human Evaluation Results

To validate our automatic evaluation metrics and gain deeper insights into model performance, we conducted a comprehensive human evaluation study with four participants who have backgrounds in healthcare education. Each evaluator independently assessed 90 images (30 distinct image pairs from Flux.1[Dev], our fine-tuned Flux.1[Dev], and GPT4o respectively). Participants rated each image on a 5-point Likert scale (1=poor, 5=excellent) across four dimensions: Text Readability, Layout Organization, Text-Subfigure Alignment, and Overall Quality. Details are provided in Appendix D. The results are shown in Figure 4.

GPT-4o-Image achieves better performance in

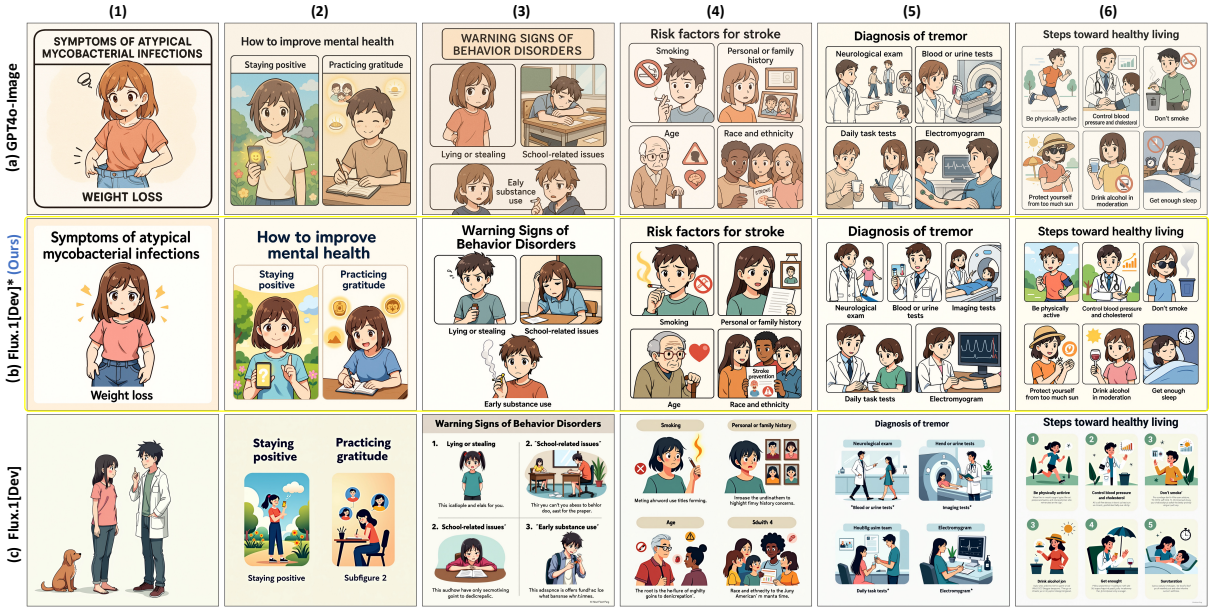

Figure 5: Qualitative Demonstration of Our Models. The Healthcards shown are expected to include knowledge points covering a range from 1 to 6, arranged sequentially from left to right. Results from three selected models are presented: **GPT-4o-Image** (top row), **fine-tuned Flux.1[Dev]** using our dataset (middle row), and **baseline Flux.1[Dev]** (bottom row). T2I prompts for these cases are provided in the Appendix Table 3.

Text Readability, Alignment, and Overall Quality, while fine-tuned Flux.1[Dev] shows significantly better results in Layout Organization. This aligns with our automatic evaluation outcomes in Table 1. Notably, our fine-tuned model achieves an average overall score of 4 (good), while its raw counterpart receives a negative overall score, demonstrating the value and effectiveness of fine-tuning on our dataset for real-world applications.

We also examined how human evaluation correlates with our automatic metrics. The Spearman correlation between Text Readability and Word Accuracy and Character Accuracy are 0.526 and 0.530 respectively ($p<0.05$). The Layout Organization score correlates with the accuracy of subfigure counting at 0.796 ($p<0.001$). The Text-Subfigure Alignment score correlates with $CLIPScore_s$ at 0.315 and with VQA_s at 0.368 . The Overall Assessment correlates with VQA at 0.337, with accuracy of subfigure count at 0.567 ($p<0.01$), and with word accuracy at 0.505 ($p<0.01$). These correlations reveal that subfigure count accuracy and word accuracy most strongly influence evaluators' overall impressions, highlighting the critical importance of accurate text rendering and layout generation in health educational materials. This aligns with intuitive human review patterns for knowledge cards—when evaluators notice incorrect knowledge point counts or inaccurate titles, they tend to assign lower scores without needing to examine the image content in detail. This finding underscores the fundamental importance of structural and textual accuracy as gatekeeping criteria in the evaluation of educational health materials.

6.5 More experimental Explorations.

We provide more experimental explorations such as different Lora ranks, data ratios, and guidance scales in Appendix E.

7 Conclusion

We introduced HealthCards, the first comprehensive exploration of text-to-image models for healthcare knowledge visualization through educational flashcards. Our contributions include a high-quality dataset of expert-verified healthcare flashcards sourced from credible medical knowledge resources. We identified an "evenness bias" in current models like GPT-4o-Image, where they prioritize visual symmetry over content accuracy, and demonstrated that fine-tuning with our dataset effectively mitigates this issue while significantly improving overall quality. Our experiments show that fine-tuned open-source models can achieve good ratings from healthcare educators, with particularly strong improvements in layout organization and subfigure counting accuracy.

Limitations

(1) Our current models still face challenges in highly specialized medical contexts, which is a limitation inherent to the foundational knowledge of the underlying models—a constraint that affects even GPT-4o-Image. These knowledge limitations led us to focus our initial application on public health education rather than specialized medical training. The models perform well in conveying general health concepts but may struggle with nuanced medical terminology or highly technical concepts that require expert-level domain knowledge.

(2) While our fine-tuned models successfully mitigate the evenness bias and demonstrate significant improvements, they still fall short of GPT-4o-Image in text rendering and content alignment. At the time of this submission, the architecture, parameters, and methodologies behind GPT-4o-Image remain undisclosed, creating uncertainty for our alignment efforts. GPT-4o-Image may employ fundamentally different approaches such as autoregressive generation or agent-based systems, which could explain the performance gap and present challenges for future improvements using purely open-source methods.

(3) As pioneers in this emerging research area, we were unable to conduct long-term educational effectiveness studies within our project timeline. While our human evaluations with healthcare educators provide valuable initial validation, more extensive studies on knowledge retention and learning outcomes with diverse learner populations would strengthen claims about educational benefits. We plan to release our weights and dataset after the anonymous review period, enabling interested communities to conduct educational trials in various domains and contribute to this promising research direction.

References

31Memorize. 2024. Guiding principles of effective flashcard creation. Accessed: 2025-05-18.

Agency for Healthcare Research and Quality. 2020. Topic: Use of visual aids. Accessed: 2025-05-18.

Safinah Ali, Prerna Ravi, Katherine Moore, Hal Abelson, and Cynthia Breazeal. 2024. A picture is worth a thousand words: Co-designing text-to-image generation learning materials for k-12 with educators. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23260–23267.

Anki. 2025. Anki - powerful, intelligent flashcards.

Anki. n.d. Anki for first year. Accessed: 2025-05-18.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

Black Forest Labs. 2025. FLUX1.1 Pro: Best in class image generation.

Brainscape. n.d. The best flashcards for medical and nursing students. Accessed: 2025-05-18.

Jingtao Cao, Zhang Zheng, Hongru Wang, and Kam-Fai Wong. 2024. Vleu: a method for automatic evaluation for generalizability of text-to-image models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11034–11049.

Pu Cao, Feng Zhou, Junyi Ji, Qingye Kong, Zhixiang Lv, Mingjian Zhang, Xuekun Zhao, Siqi Wu, Yinghui Lin, Qing Song, and 1 others. 2025. Preliminary explorations with gpt-4o (mni) native image generation. *arXiv preprint arXiv:2505.05501*.

PaddlePaddle Contributors. 2025. Paddleocr: Awesome multilingual ocr toolkits based on paddlepaddle. `https://github.com/PaddlePaddle/PaddleOCR`. Accessed: 2025-05-18.

Dani Di Placido. 2025. The AI-generated Studio Ghibli trend, explained.

Weichen Fan, Amber Yijia Zheng, Raymond A Yeh, and Ziwei Liu. 2025. Cfg-zero*: Improved classifier-free guidance for flow matching models. *arXiv preprint arXiv:2503.18886*.

Rahmah Fithriani. 2023. Gamified efl learning: Utilizing 4 pics 1 word application in vocabulary class. *Journal of English Language Studies*, 8(1):1–18.

Niharika Goparaju. 2024. Picture this: Text-to-image models transforming pediatric emergency medicine. *Annals of Emergency Medicine*, 84(6):651–657.

Mélanie Gréaux, Maria Francesca Moro, Kaloyan Kamenov, Amy M Russell, Darryl Barrett, and Alarcos Cieza. 2023. Health equity for persons with disabilities: a global scoping review on barriers and interventions in healthcare services. *International Journal for Equity in Health*, 22(1):236.

David M Harris, Michael Chiang, and MIchael Chiang. 2022. An analysis of anki usage and strategy of first-year medical students in a structure and function course. *Cureus*, 14(3).

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

HiDream-ai. 2025. Hidream-i1: A new open-source image generative model. https://github.com/HiDream-ai/HiDream-I1.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2025. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

WebMD Ignite. 2022. Visuals in health education. Accessed: 2025-05-18.

Black Forest Labs. 2024. Flux. https://github.com/black-forest-labs/flux.

Katherine Lee and Dan Nathan-Roberts. 2021. Using visual aids to supplement medical instructions, health education, and medical device instructions. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, 10(1):257–262.

Jialuo Li, Wenhao Chai, Xingyu Fu, Haiyang Xu, and Saining Xie. 2025. Science-t2i: Addressing scientific illusions in image synthesis. *Preprint*, arXiv:2504.13129.

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 *European Conference on Computer Vision*, pages 366–384. Springer.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky TQ Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. 2024. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*.

Shawn Loewen, Dustin Crowther, Daniel R Isbell, Kathy Minhye Kim, Jeffrey Maloney, Zachary F Miller, and Hima Rawal. 2019. Mobile-assisted language learning: A duolingo case study. *ReCALL*, 31(3):293–311.

Logopsycom. 2020. Flashcards: A tool to facilitate learning. Accessed: 2025-05-02.

Matthew Lu, John H Farhat, and Gary L Beck Dallaghan. 2021. Enhanced learning and retention of medical knowledge using the mobile flash card application anki. *Medical science educator*, 31(6):1975–1981.

Michelle Mai, Fatima N Mirza, and Christopher DiMarco. 2024. Application of text-to-image translation algorithms in medicine: a systematic review. *JAAD Reviews*.

Richard E Mayer. 2005. *The Cambridge handbook of multimedia learning*. Cambridge university press.

Njabulo Mbanda, Shakila Dada, Kirsty Bastable, Gimbler-Berglund Ingalill, and 1 others. 2021. A scoping review of the use of visual aids in health education materials for persons with low-literacy levels. *Patient education and counseling*, 104(5):998–1017.

Geoffroy PJC Noel. 2024. Evaluating ai-powered text-to-image generators for anatomical illustration: A comparative study. *Anatomical sciences education*, 17(5):979–983.

Chinedu Innocent Nwoye, Rupak Bose, Kareem Elgohary, Lorenzo Arboit, Giorgio Carlino, Joël L Lavanchy, Pietro Mascagni, and Nicolas Padoy. 2025. Surgical text-to-image generation. *Pattern Recognition Letters*.

OpenAI. 2025. Introducing 4o image generation.

Ville Paananen, Jonas Oppenlaender, and Aku Visuri. 2024. Using text-to-image generation for architectural design ideation. *International Journal of Architectural Computing*, 22(3):458–474.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Will Beddow, Erwann Millon, Wenhai Wang Victor Perez, Yu Qiao, Bo Zhang, Xiaohong Liu, Hongsheng Li, Chang Xu, and Peng Gao. 2025. Lumina-image 2.0: A unified and efficient image generative framework. *Preprint*, arXiv:2503.21758.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Rani Santika, Rafi Farizki, and Hanif Nurcholish Adiantika. 2023. English language learning with flashcard media and educational posters to improve language education. *Journal of Social Science*, 4(3):706–711.

Rabia Shahid, Muhammad Shoker, Luan Manh Chu, Ryan Frehlick, Heather Ward, and Punam Pahwa. 2022. Impact of low health literacy on patients' health outcomes: a multicenter cohort study. *BMC health services research*, 22(1):1148.

Streamlit. 2025. Streamlit: A faster way to build and share data apps. https://github.com/streamlit/streamlit.

Michael Sun, Shelun Tsai, Deborah L Engle, and Shelley Holmer. 2021. Spaced repetition flashcards for teaching medical students psychiatry. *Medical science educator*, 31:1125–1131.

Raphael Tang, Xinyu Zhang, Lixinyu Xu, Yao Lu, Wenyan Li, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. 2024. Words worth a thousand pictures: Measuring and understanding perceptual variability in text-to-image generation. *arXiv preprint arXiv:2406.08482*.

Kaitlyn Teske. 2017. Duolingo. *Calico Journal*, 34(3):393–401.

Luh Ayu Tirtayani, Mutiara Magta, and Ni Gusti Ayu Made Yeni Lestari. 2017. Teacher friendly e-flashcard: a development of bilingual learning media for young learners. *Journal of Education Technology*, 1(1):18–29.

Wikipedia contributors. 2025. Flashcard — Wikipedia, the free encyclopedia. [Online; accessed 2-May-2025].

Zhiyuan Yan, Junyan Ye, Weijia Li, Zilong Huang, Shenghai Yuan, Xiangyang He, Kaiqing Lin, Jun He, Conghui He, and Li Yuan. 2025. Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation. *arXiv preprint arXiv:2504.02782*.

Shitian Zhao, Qilong Wu, Xinyue Li, Bo Zhang, Ming Li, Qi Qin, Dongyang Liu, Kaipeng Zhang, Hongsheng Li, Yu Qiao, and 1 others. 2025. Lex-art: Rethinking text generation via scalable high-quality data synthesis. *arXiv preprint arXiv:2503.21749*.

Qinqing Zheng, Matt Le, Neta Shaul, Yaron Lipman, Aditya Grover, and Ricky TQ Chen. 2023. Guided flows for generative modeling and decision making. *arXiv preprint arXiv:2311.13443*.

A Dataset Construction and Annotation Process

In this section, we provide more details about our dataset construction and annotation Process. We present prompt 1 in box 2, which guides the LLM f to perform knowledge extraction from essays, as described in 4.1.

Prompt 1 (for Knowledge Extraction)

You are a helpful health education assistant. Your task is to extract ALL knowledge points from the provided context and organize them into appropriate health flashcards. For each knowledge point: 1. Extract the original content from the text. 2. Assign a concise but descriptive title. 3. Create a brief description for an illustration that would help visualize the concept.
Focus on extracting knowledge related to: - Symptoms - Causes - Treatments - Prevention methods - Diagnostic procedures - Risk factors - Complications - Any other health relevant information
Guidelines: - Extract ALL knowledge points present in the text, being comprehensive - Group related knowledge points under appropriate topics - The content should be directly from the original text - do not infer or add information - If the context lacks useful information (e.g., only contains external links), you may return an empty list - Make titles concise but informative, including key information - For each knowledge point, include an illustration description that would help visualize the concept
Here's an example of the expected format: {example1} {example2}
Now, extract ALL knowledge points from the following context and organize them into appropriate Health FlashCards:

Table 2: *prompt 1* used for extracting knowledge from healthcare essays.

After structured knowledge points being extracted, we assemble them into text-to-image generation prompts. This composing process has been defined as ϕ in Section 4.1. Here we provide some complete examples in Table 3. Note that the maximum length is controlled below 512 tokens.

After all the T2I prompts has been prepared, we fed them into GPT-4o-Image for healthcard image generation. Following which we performed a dataset review with experts on our specialized developed platform. A screenshot of the platform is presented in Figure 7. It is built on Streamlit and could be easily deployed for shared annotation (Streamlit, 2025).

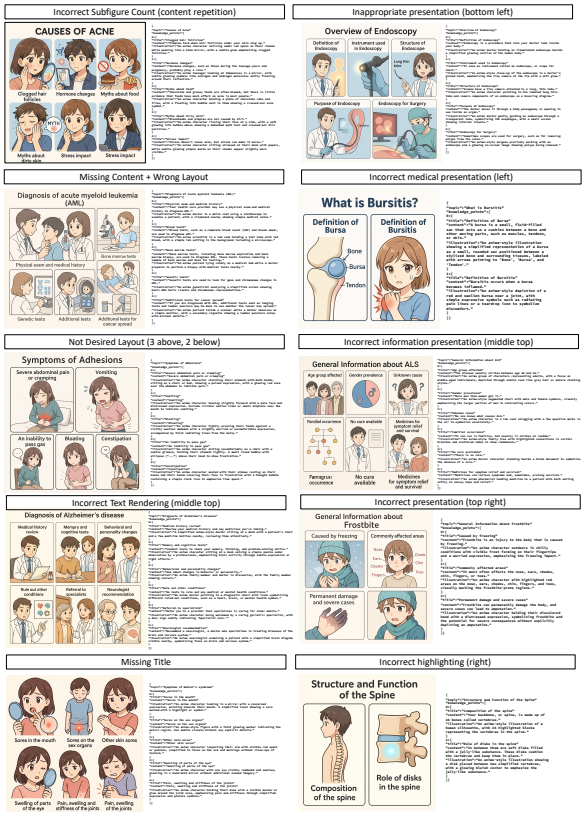

Figure 6: Some examples of discarded samples or ones need to regenerate. (Best Viewed when zoomed in)

A.1 The Annotation Platfrom

When using this platform, as shown in Figure 7, the interface displays generated images alongside the original material (the source essay from which knowledge points are extracted, shown on the left) and the structured knowledge point (shown on the right). Experts follow a systematic verification procedure: (1) They first assess whether the Health Card on the right appropriately extracts information from the original material on the left. If not, the item is labeled "Need Prompt Modification." (2) If the extraction is valid, experts proceed to evaluate the image content by verifying the correct number of subfigures, identifying expected text elements (healthcare topic and subfigure headings), and examining whether each subfigure's content aligns with its pictorial illustration. Images that pass all verification criteria are accepted into the

PROMPT FOR FIGURE 5 (1)
Design a clean, 1:1 aspect ratio medical flashcard with anime-style illustrations for the given knowledge point. The subfigure should include a clear title and illustration, emphasizing clarity, readability, and an approachable style. Topic (title) of Flashcard: 'Symptoms of atypical mycobacterial infections'. Subfigure: 'Weight loss' - Illustration: An anime character noticing their loose-fitting clothes and appearing perplexed.
PROMPT FOR FIGURE 5 (2)
Design a clean, 1:1 aspect ratio medical flashcard with anime-style illustrations for two knowledge points. Each subfigure should include a clear title and illustration, emphasizing clarity, readability, and an approachable style. Topic (title) of Flashcard: 'How to improve mental health'. Subfigure 1 (left): 'Staying positive' - Illustration: An anime character standing in a bright park with flowers nearby, smiling gently. They hold a glowing phone showing positivity while a small cloud drifts away in the background, symbolizing negative information being set aside. Subfigure 2 (right): 'Practicing gratitude' - Illustration: An anime character sitting at a simple desk, happily writing in a journal with small glowing icons of a family portrait, a meal, and a sunset floating around them.
PROMPT FOR FIGURE 5 (3)
Design a clean, 1:1 aspect ratio medical flashcard with anime-style illustrations for three knowledge points. Use a two-layer grid layout (two above, one below). Each subfigure should include a clear title and illustration, emphasizing clarity, readability, and an approachable style. Topic (title) of FlashCard: 'Warning Signs of Behavior Disorders'. Subfigure 1 (top left): 'Lying or stealing' - Illustration: An anime character holding a small object behind their back, with a nervous or guilty expression to represent dishonest behavior. Subfigure 2 (top right): 'School-related issues' - Illustration: An anime classroom setting with a character slouched on a desk, looking discouraged with scattered papers around to suggest poor academic performance. Subfigure 3 (bottom): 'Early substance use' - Illustration: An anime teenager holding a cigarette or small bottle, with a worried expression representing unhealthy habits.
PROMPT FOR FIGURE 5 (4)
Design a clean, 1:1 aspect ratio medical flashcard with anime-style illustrations for four knowledge points. Use a two-layer grid layout (two above, two below). Each subfigure should include a clear title and illustration, emphasizing clarity, readability, and an approachable style. Topic (title) of FlashCard: 'Risk factors for stroke'. Subfigure 1 (top left): 'Smoking' - Illustration: An anime character holding a cigarette with a worried expression. Include a glowing 'no-smoking' symbol in the frame for emphasis. Subfigure 2 (top right): 'Personal or family history' - Illustration: An anime character looking at framed family photos, with a single subtle chart or document visible in the background to highlight family history concerns. Subfigure 3 (bottom left): 'Age' - Illustration: An elderly anime character with a walking stick and glasses, accompanied by simplified icons like a warning sign near a heart or brain diagram. Subfigure 4 (bottom right): 'Race and ethnicity' - Illustration: An anime-style group of diverse characters looking at a stroke prevention poster or booklet held by an African American character.
PROMPT FOR FIGURE 5 (5)
Design a clean, 1:1 aspect ratio medical flashcard with anime-style illustrations for five knowledge points. Use a two-layer grid layout (three above, two below). Each subfigure should include a clear title and illustration, emphasizing clarity, readability, and an approachable style. Topic (title) of FlashCard: 'Diagnosis of tremor'. Subfigure 1 (top left): 'Neurological exam' - Illustration: An anime doctor pointing at a straight walking path while a patient walks carefully, with a small additional scene showing the doctor listening to the patient speak. Subfigure 2 (top middle): 'Blood or urine tests' - Illustration: An anime scientist handling a test tube in a laboratory setting, visually focused on analyzing the sample. Subfigure 3 (top right): 'Imaging tests' - Illustration: An anime patient inside an MRI machine with a technician observing the process through a glass control room window. Subfigure 4 (bottom left): 'Daily task tests' - Illustration: An anime patient carefully holding a cup while a doctor takes notes, demonstrating focused observation. Subfigure 5 (bottom right): 'Electromyogram' - Illustration: An anime scene showing electrodes attached to a patient's arm, connected to a monitor displaying live results.
PROMPT FOR FIGURE 5 (6)
Design a clean, 1:1 aspect ratio medical flashcard with anime-style illustrations for six knowledge points. Use a two-layer grid layout (three above, three below). Each subfigure should include a clear title and illustration, emphasizing clarity, readability, and an approachable style. Topic (title) of FlashCard: 'Steps toward healthy living'. Subfigure 1 (top left): 'Be physically active' - Illustration: An anime character jogging forward with visible energy lines and a park environment suggested by minimal tree icons. Subfigure 2 (top middle): 'Control blood pressure and cholesterol' - Illustration: An anime doctor using a blood pressure cuff with a simplified chart icon showing healthy cholesterol levels in the background. Subfigure 3 (top right): 'Don't smoke' - Illustration: An anime character throwing a cigarette into a trash bin with a smoke-free icon in the background. Subfigure 4 (bottom left): 'Protect yourself from too much sun' - Illustration: An anime character wearing a hat and sunglasses under a simplified sun icon, applying sunscreen with a beach umbrella behind. Subfigure 5 (bottom middle): 'Drink alcohol in moderation' - Illustration: An anime character holding a glass of water beside a wine glass with a moderation symbol. Subfigure 6 (bottom right): 'Get enough sleep' - Illustration: An anime character peacefully sleeping in bed with a blanket and a nearby clock showing nighttime hours.

Table 3: Examples of our Text-to-Image generation prompts.

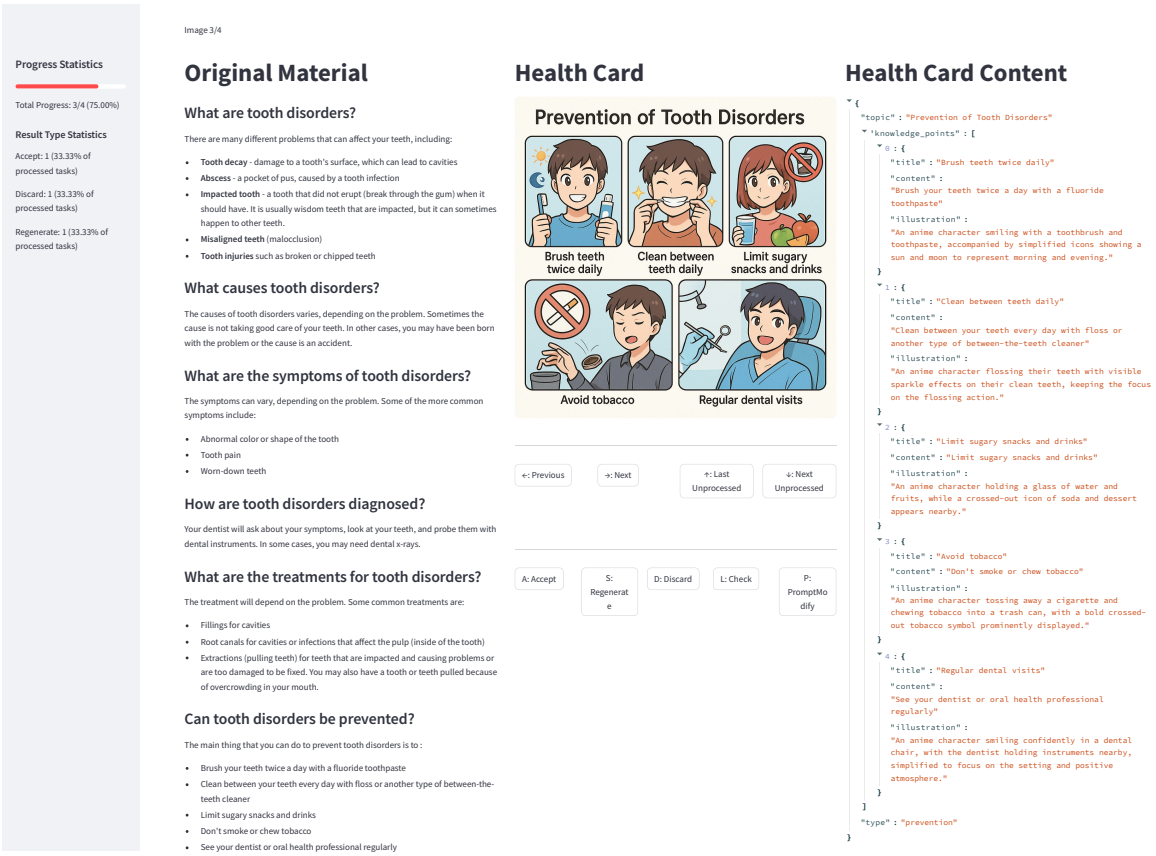

Figure 7: Screenshot of our developed platform for *HealthCards* dataset annotation.

dataset; otherwise, they require regeneration or additional review. We have released this platform in our anonymous github repository.

B Recapitulation for Flow Matching and Classifier-free Guidance

Following the explanations as in previous works (Lipman et al., 2022, 2024; Fan et al., 2025), we provide a revisit of basic concepts of *Flow Matching* and *Classifier-free Guidance* techniques. Our notation follows the conditional setting, yet the ideas hold for the unconditional case as well.

B.1 Conditional Flow Matching (CFM)

Let the *source* density be $p(x \mid y)$ and the *target* density be $q(x \mid y)$, where y is an external conditioning signal (e.g. a class label). CFM introduces a continuum of intermediate distributions $\{p_t(x \mid y)\}_{t\in[0,1]}$ that smoothly morph the source into the target, satisfying $p_0(x \mid y) = p(x \mid y)$ and $p_1(x \mid y) = q(x \mid y)$. A convenient choice is the linear interpolation

$$p_t(x \mid y) \;:=\; (1-t)\,p(x \mid y) + t\,q(x \mid y), \quad (5)$$

implying that a sample drawn at time t can be written as $x_t = (1-t)\,x_0 + t\,x_1$ with $x_0 \sim p(x \mid y)$ and $x_1 \sim q(x \mid y)$.

To describe the evolution of x_t, we fit a time-dependent velocity field $\dot{x}_t = v_t^\theta(x \mid y)$, where v_t^θ is a neural network parameterized by θ. The parameters are optimized through the mean-squared error objective

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,x_0,x_1}\left\| v_t^\theta(x_t \mid y) - (x_1 - x_0) \right\|_2^2. \quad (6)$$

After training, new samples from $q(\cdot \mid y)$ are obtained by numerically integrating the learned ODE with a solver.

B.2 Classifier-Free Guidance (CFG)

Classifier-free guidance (Ho and Salimans, 2022) enhances conditional generation by biasing the trajectory toward the conditioning signal without a separate classifier. During training, the same neural field v_t^θ is asked to produce both conditional and *unconditional* velocities by occasionally replacing y with the null token $\varnothing$. At inference time, the

guided field is obtained by the linear blend

$$\hat{v}_t^\theta(x \mid y) \;:=\; (1-\omega)\, v_t^\theta(x \mid y = \varnothing) + \omega\, v_t^\theta(x \mid y), \tag{7}$$

where $\omega \geq 0$ controls the guidance strength. Setting $\omega = 1$ recovers the purely conditional sampler, while $\omega > 1$ amplifies the influence of y.

C More Detailed Explanation of Evaluation Setting

C.1 Text Rendering Score

To evaluate the quality of text rendering in generated images, we develop a pipeline that assesses how accurately the text elements are rendered and can be detected by OCR systems. We harness PaddleOCR (Contributors, 2025) to extract all identifiable words from the images.

As shown in Figure 8, while word-level identification and localization is achieved, these words require further processing to be reorganized into readable sentences that match the intended text layout of the flash cards.

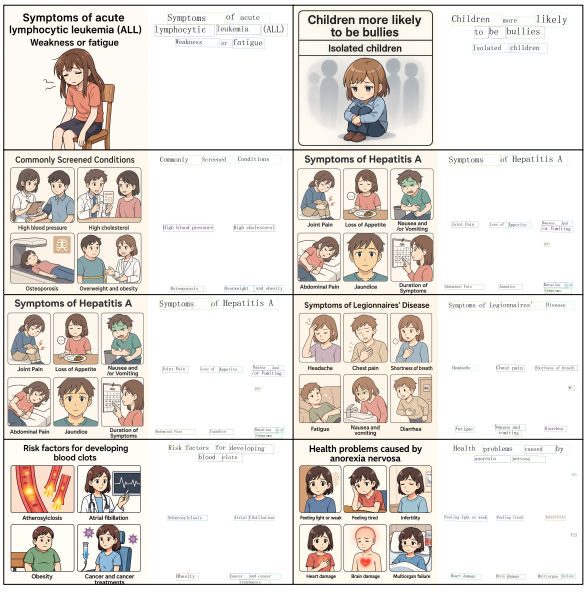

Figure 8: Examples of OCR text detection results showing word-level identification.

Our text rendering evaluation pipeline consists of the following key steps:

C.1.1 Text Grouping and Ordering

We first perform an iterative grouping process to cluster detected text elements that belong together. This is accomplished through a two-step approach:

We merge overlapping bounding boxes to handle cases where OCR detects parts of the same text element as separate entities.

We iteratively group text elements whose distances are below a predefined threshold (20 pixels in our implementation), continuing until no further merging occurs.

C.1.2 Reading Order Determination

After grouping, we determine the proper reading order of text elements within each group to reconstruct coherent text. Our approach follows natural human reading patterns (left-to-right, top-to-bottom):

We first identify line breaks within each text group by clustering the vertical (y-coordinate) positions of text elements. Text elements with vertical positions differing by less than one-third of the average text height are considered to be on the same line.

For each line, we sort text elements by their horizontal (x-coordinate) position to establish left-to-right reading order.

Finally, lines are arranged in top-to-bottom order to produce the complete reading sequence for each text group. A demo of this grouping and ordering process is visualized in Figure 9, where each group is enclosed by a dashed outline and numbers indicate the determined reading sequence.

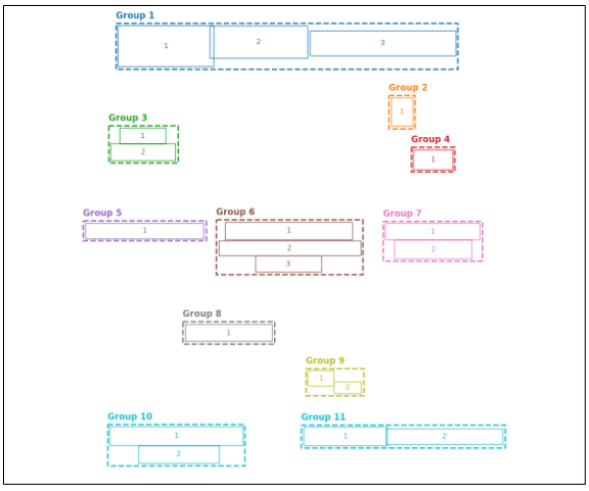

Figure 9: Visualization of text grouping and reading order determination. Each colored rectangle represents an individual text box detected by OCR. Text boxes are grouped (indicated by dashed outlines) based on proximity and overlap. Numbers within each box indicate the reading order sequence determined by our algorithm, following natural reading patterns (left-to-right within lines, top-to-bottom across lines).

C.1.3 Text Matching and Evaluation

To evaluate text rendering quality, we compare OCR-extracted text against the expected text from

the original flash card through the following process: (1) normalizing both text sets using case-folding and standardization; (2) constructing a cost matrix of Levenshtein distances between extracted and expected text elements; (3) applying the Hungarian algorithm to determine optimal text matching that minimizes overall edit distance; and (4) computing Character Error Rate (CER) and Word Error Rate (WER) metrics between matched pairs. This approach enables objective measurement of text rendering fidelity independent of spatial arrangement variations.

C.2 Subfigure Counting Score

To evaluate the structural fidelity of generated flash cards, we implement a subfigure counting mechanism that assesses whether the correct number of visual elements are present. This metric is particularly important for multi-topic flash cards where each knowledge point should be accompanied by a corresponding illustration.

We employ the Qwen2.5-VL-32B model to detect and localize subfigures within the generated images. This multimodal vision-language model is capable of understanding complex visual layouts and identifying distinct visual elements. The model receives a prompt requesting detection of all subtopics or subfigures, and returns bounding box coordinates along with associated labels.

The prompt template used for this task is:

> This image has several subtopics (subfigures) under the title. Please detect all subtopics (subfigures) in the image and return their locations in the form of coordinates. The bounding box should include the whole subfigure (including the subtitle and corresponding drawing).

As demonstrated in Figure 10, the model effectively localizes individual subfigures across various medical flash cards with different layouts and topics. For each generated flash card, we compare the number of detected subfigures against the expected count from the original flash card specification. We compute two key metrics:

1. **Mean Gap**: The absolute difference between the detected subfigure count and the expected count, averaged across all samples.

2. **Error Rate**: The proportion of samples where the detected count does not match the expected count.

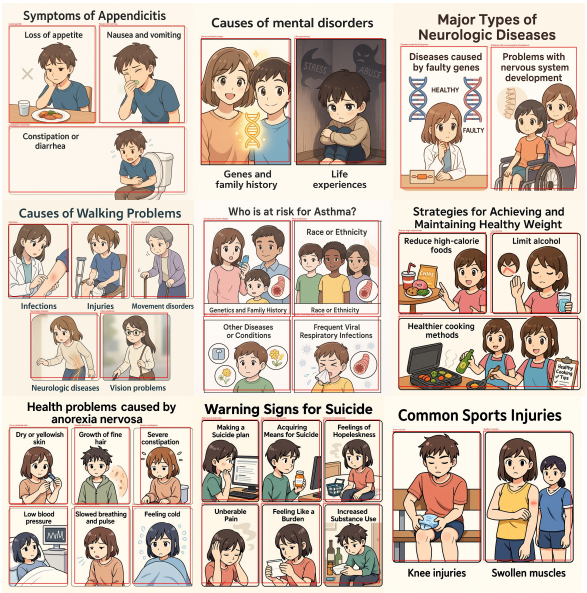

Figure 10: Illustration of subfigure detection using Qwen2.5-VL-32B model. Red bounding boxes indicate detected subfigures across various medical flash cards. The model successfully identifies distinct subtopics within each flash card, enabling accurate counting of knowledge points and their associated visual elements.

We exclude flash cards with only a single knowledge point from this evaluation, as the subfigure counting is primarily meaningful for multi-topic cards. Lower values for both metrics indicate better structural fidelity to the original flash card design.

This evaluation approach objectively quantifies a model's ability to maintain the intended structural organization of educational flash cards, which is crucial for proper information hierarchy and visual organization in educational materials.

C.3 Subfigure Alignment Score

After localizing individual subfigures within the flash cards, we evaluate how well each subfigure aligns with its intended knowledge point. This alignment assessment is critical for educational materials, as it measures whether the visual representations accurately convey the corresponding concepts. We employ two complementary metrics to quantify this alignment: CLIP Score and VQA Score.

C.3.1 CLIP Score

We utilize the Contrastive Language-Image Pre-training (CLIP) framework (Hessel et al., 2021) to measure the semantic similarity between each subfigure and its corresponding textual description. Specifically, we employ OpenAI's ViT-L-14-336

CLIP model to compute cosine similarity between image and text embeddings. This approach provides a direct measurement of how well the generated visual elements match their intended semantic content.

The CLIP Score ranges from 0 to 1, with higher values indicating stronger alignment between the subfigure and its textual description. This metric effectively captures the general semantic correspondence but may not fully assess more complex relationships or fine-grained details.

C.3.2 VQA Score

To complement the CLIP Score with a more contextual understanding of image-text alignment, we implement a Visual Question Answering (VQA) approach using the CLIP-FlanT5-XXL model (Lin et al., 2024). This model evaluates whether the subfigure contains the necessary visual information to answer questions derived from the knowledge point description.

For each subfigure, we treat the knowledge point description as a prompt or implicit question that the image should address. The VQA Score represents the model's confidence (from 0 to 1) that the subfigure adequately illustrates the described concept. This metric provides insight into how well the subfigure serves its educational purpose by visually representing the intended knowledge.

By combining these two alignment metrics, we obtain a comprehensive assessment of how effectively each subfigure communicates its intended educational content. This evaluation is crucial for ensuring that generated flash cards not only maintain correct structure but also convey accurate and clear visual representations of the knowledge points they aim to illustrate.

D Details of Human Evaluation

Here we provide instruction to the human evaluators, in Figure 11.

E More Experimental Results and Analysis

In this section, we present additional experimental results using Flux.1[Dev], as shown in Table 4. We investigate: (1) Different LoRA ranks, ranging from 1 to 256; (2) Various scales of classifier-free guidance; (3) Fine-tuning with Flux.1[Dev]'s text-rendering specialist, *LexFlux* (Zhao et al., 2025); and (4) Different data ratios during training (while maintaining consistent total iterations). For reference, our primary model was trained with LoRA rank 128, using the complete dataset and CFG guidance of 5.

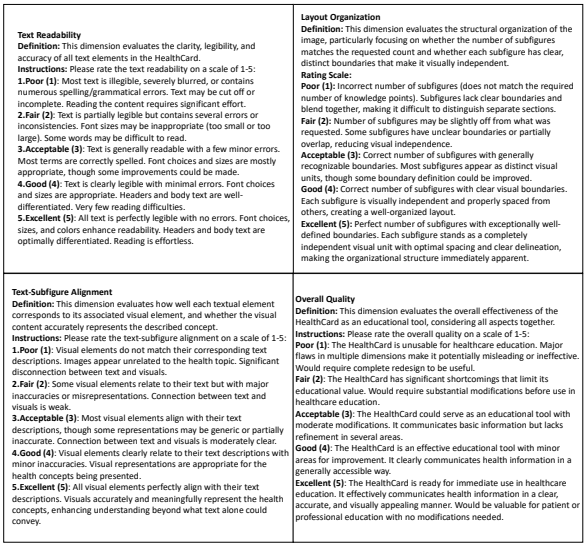

Figure 11: Instructions for human evaluation.

Results indicate that slightly increasing CFG to 7 enhances subfigure counting performance, though at the expense of text rendering quality. Both excessive and insufficient guidance lead to deterioration across all metrics. Surprisingly, utilizing a text rendering-specialized pretrained model did not yield improvements, possibly because pre-optimization introduces biases that impede learning for our specific task.

From the data ratio experiments, we observe that performance metrics generally improve with increased data usage, particularly for text rendering scores. This demonstrates the value of our dataset's scale. Finally, LoRA rank 128 appears optimal for this task, as both higher and lower ranks result in decreased performance.

F ARR Submission Checklist

F.1 Potential Risks

Despite improvements through fine-tuning, our models occasionally generate inaccurate or misleading healthcare information, which presents significant concerns in educational contexts. Incorrect visual representations of medical concepts could lead to misunderstanding among learners and potentially influence future healthcare decisions. The visually compelling nature of the generated flashcards might increase the likelihood that incorrect information is trusted and remembered, amplifying

Model	Text Rendering		Subfigure Counting		Subfigure Alignment		General		
	WER↓	CER↓	Gap↓	Error Rate↓	ClipScore$_s$↑	VQA$_s$↑	Aesthetic↑	Quality↑	VQA↑
lora r1	0.223	0.149	0.115	0.112	0.261	0.828	3.361	4.613	0.830
lora r16	0.152	0.088	0.119	0.115	0.264	0.838	3.396	4.614	0.832
lora r64	0.153	0.092	0.016	0.016	0.265	0.852	3.387	4.606	0.836
lora r256	0.155	0.097	0.055	0.051	0.265	0.859	3.404	4.635	0.835
CFG guidance3	0.191	0.109	0.014	0.010	0.261	0.850	3.377	4.611	0.830
CFG guidance7	0.191	0.209	**0.004**	**0.004**	0.262	0.844	3.352	4.618	0.835
CFG guidance10	0.204	0.119	0.012	0.006	0.260	0.841	3.337	4.604	0.831
LexFlux	0.173	0.109	0.031	0.031	0.265	0.858	3.409	4.632	0.837
20% Data	0.195	0.133	0.045	0.041	0.265	0.854	3.384	4.601	0.836
50% Data	0.154	0.096	0.072	0.072	0.266	0.852	3.406	4.628	0.835
75% Data	0.124	0.065	0.039	0.035	**0.266**	0.858	3.405	4.634	0.835
Ours	**0.118**	**0.064**	0.010	0.008	0.265	**0.861**	**3.429**	**4.636**	**0.838**

Table 4: More Experimental Study of hyperparameters and settings.

the impact of any inaccuracies. We strongly recommend human expert review before deployment in actual educational settings, and future work should explore developing robust verification mechanisms specifically for healthcare content. Addressing these safety concerns will be crucial for responsible implementation of these technologies in health education.

F.2 Use Or Create Scientific Artifacts

F.2.1 Cite Creators Of Artifacts

We sourced credible healthcare resources from the NIH MedlinePlus website. According to the website's "Linking to and Using Content from MedlinePlus" page (https://medlineplus.gov/about/using/usingcontent/), certain content is in the public domain and not copyrighted. We strictly followed their guidelines to ensure all sourced knowledge was from non-copyrighted sections, including health topic summaries, medical test information, Genetics page summaries, and illustrations with the "U.S. National Library of Medicine" watermark. For citation of creators, we include the acknowledgment "Courtesy of MedlinePlus from the National Library of Medicine" as required by the website. We also utilize GPT4o-Image (OpenAI, 2025) and have properly cited this in our paper.

F.2.2 Discuss The License For Artifacts

As mentioned in the previous subsection, our sourced knowledge information is not copyrighted. This data source was also used by previous work (Ben Abacha and Demner-Fushman, 2019), which shared their dataset under the Creative Commons Attribution 4.0 International License (CC BY).

F.2.3 Artifact Use Consistent With Intended Use

Yes, our use of artifacts strictly follows the guidelines provided by MedlinePlus. We only utilize content that is explicitly marked as being in the public domain, including health topic summaries, medical test information, and genetics information. Our artifacts will be released under the Creative Commons Attribution 4.0 International License (CC BY).

F.2.4 Data Contains Personally Identifying Info Or Offensive Content

No. No personally identifying data has been created or included. We have not observed any identifiable personal data in our dataset.

F.2.5 Statistics For Data

Please refer to Section 4.3

F.3 Computational Experiments

F.3.1 Model Size And Budget

40 H100 GPU hours are required for training a Flux.1[Dev] (Labs, 2024) which has 12B parameters.

F.3.2 Descriptive Statistics

We report results of single runs, due to computational limitations. As aforementioned, it requires 40 H100 GPU hours to finetune a model. Our results are reproducible.

F.3.3 Parameters For Packages

Yes, we have included necessary mentions of models used, such as QwenVL (Bai et al., 2025) and PaddleOCR (Contributors, 2025).

F.4 Human Subjects Including Annotators

F.4.1 Instructions Given To Participants

Yes, we have provided screenshots of annotation platforms and instructions for annotating in section A.

F.4.2 Recruitment And Payment

all annotators were recruited as volunteers and paid at rates exceeding local standards.

F.4.3 Data Consent

Yes. We have explained the purpose of our work to all annotators. They are aware and have agreed that their opinions, such as ratings and modifications of text-to-image prompts, will be used as public data. Throughout the process, no personally identifiable information was collected.

F.4.4 Ethics Review Board Approval

Yes, our data collection protocol strictly adheres to the licenses from each source. No sensitive data was collected or used, and no sensitive experiments were conducted. The ethics review board determined this project to be exempt.

F.4.5 Characteristics Of Annotators

Yes, we describe this in section 4.

F.5 Ai Assistants In Research Or Writing

F.5.1 Information About Use Of Ai Assistants

Yes, we used AI assistants like Claude and GPT4o only for polishing language use.