Document-level Simplification and Illustration Generation: Multimodal Coherence

Yuhang Liu[1]* **Mo Zhang**[1]* **Zhaoyi Cheng**[1] **Sarah Ebling**[1]†
[1]University of Zurich
{yuhang.liu3, mo.zhang, zhaoyi.cheng}@uzh.ch, ebling@cl.uzh.ch

Abstract

We present a novel multimodal system for document-level text simplification and automatic illustration generation, aimed at enhancing information accessibility for individuals with cognitive impairments. While prior research has primarily focused on sentence- or paragraph-level simplification, and text-to-image generation for narrative contexts, this work addresses the unique challenges of simplifying long-form documents and generating semantically aligned visuals. The pipeline consists of four stages: (1) Document-Level Text Simplification, (2) discourse-aware segmentation using large language models, (3) visually grounded description generation via abstraction, and (4) controlled image synthesis using state-of-the-art diffusion models, including DALL·E 3 and FLUX.1-dev. We further incorporate stylistic constraints to ensure visual coherence, and we conduct a human evaluation measuring comprehension, semantic alignment, and visual clarity. Experimental results demonstrate that our system effectively combines simplified text and visual content.

1 Introduction

Simplified language aims to enhance information accessibility for individuals with cognitive impairments, learning disabilities, and others who may have difficulty comprehending standard-language written texts (Bredel and Maaß, 2016). Existing research primarily focuses on transforming complex sentences or paragraphs into more comprehensible variants, particularly for readers with low literacy skills and non-native speakers (Al-Thanyyan and Azmi, 2021). While automatic text simplification technologies have made significant progress in recent years, the role of visual aids in supporting textual comprehension is comparatively underresearched. Studies have shown that incorporating illustrations into simplified texts can further enhance understanding among individuals with cognitive disabilities (Lin et al., 2009; Winberg and Saletta, 2018; Sutherland and Isherwood, 2016). Most existing work has focused on sentence- or paragraph-level simplification and accompanying image generation (Zhang et al., 2024; Shou et al., 2023; Anschütz et al., 2024). With the growing contextual understanding capabilities of large language models (LLMs), document-level simplification and illustration generation for longer texts become possible. Our goal is to leverage these capabilities to build a pipeline for document-level text simplification and corresponding illustration generation, aiming to improve information comprehension for persons with cognitive impairments.

2 Related Work

In recent years, advances in natural language processing have significantly accelerated the development of automatic text simplification technologies. Concurrently, researchers have begun to explore multimodal approaches to further enhance the comprehensibility of simplified texts, especially through image generation. Illustrating text with images is an effective strategy to support comprehension. Visual elements not only help capture the reader's attention but also concretize abstract concepts, thereby reducing cognitive load (Glenberg and Langston, 1992; Guo et al., 2020; Wang and Zewen, 2023). The recent progress in text-to-image generation (T2I) has made the automated realization of this idea increasingly feasible. Diffusion-based models (Ho et al., 2020) have emerged as the dominant paradigm in T2I and have achieved remarkable breakthroughs. State-of-the-art models such as OpenAI's DALL·E 3 (Betker et al., 2023), Google's Imagen 3 (Baldridge et al., 2024), and Stability AI's Stable Diffusion 3 (Esser et al., 2024) can generate high-resolution, semantically relevant,

*Equal contribution.
†Corresponding author. **Email:** ebling@cl.uzh.ch

Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), pages 19–35
November 4-9, 2025

and visually creative images from complex textual descriptions. Existing research has demonstrated the use of DALL·E 3 and Stable Diffusion 3, to generate high-quality visual content that is semantically well-aligned with Easy-to-Read (E2R) textual materials (Anschütz et al., 2024). These advances are largely attributed to the models' pretraining on massive text-image paired datasets, which enable them to learn nuanced mappings between textual semantics and visual representations.

Our work builds upon these powerful diffusion models as the foundation for visual generation. While most existing systems are still confined to sentence- or paragraph-level T2I generation, generating contextually appropriate illustrations from long-form documents requires document-level understanding and scene planning capabilities. With the rapid evolution of LLMs, new frameworks are emerging that allow for these models to manage long-document processing and orchestrate image generation. In such settings, LLMs serve as directors or scriptwriters that structure the narrative and guide visual synthesis (Gado et al., 2025; Leandro et al., 2024). However, these systems are primarily tailored for narrative storytelling. In contrast, generating illustrations for informational documents demands greater factual accuracy and lower semantic ambiguity. Our research aims to bridge this gap by developing a multimodal system that integrates an LLM with a diffusion-based image generation model, enabling more accurate document-level text simplification and illustration. The proposed system features a structure-aware text simplification module and a semantically aligned image generator. Through semantic optimization and cross-modal feedback mechanisms, our method enhances the coherence between text and images and improves cognitive accessibility.

3 Method

To generate visual content for complex documents, we propose a multi-stage generative pipeline. This pipeline first decomposes a document into semantically coherent units, then translates these units into visually grounded descriptions, and finally renders them into images. Our approach integrates the advanced capabilities of LLMs for complex text processing with state-of-the-art T2I models for visual synthesis. The entire framework consists of four key stages: (1) Document-Level Text Simplification, (2) semantic document segmentation, (3) visually grounded description generation, and (4) controlled image generation.

3.1 Document-Level Text Simplification

For each discourse segment, GPT-4o rewrites the passage into Easy-to-Read German under fidelity constraints: preserve named entities, numbers, and domain terms; avoid deletions that remove obligations or eligibility; keep sentences short and syntax simple; and prohibit invented facts. Outputs are returned in JSON. We enforce tokenizer-aware limits (less than 20k characters per call) and run sanity checks for numeric consistency and entity preservation. Evaluation uses expert ratings on Simplicity (Q4), Semantic Adequacy (Q5), and Fluency (Q6) described in Section 5.

3.2 Semantic Document Segmentation

Real-world documents are rarely monolithic in topic; rather, they typically exhibit inherent discourse structures involving shifts in themes, scenes, or arguments (Grosz and Sidner, 1986). Long documents, often exceeding several thousand words, pose challenges for direct image generation, resulting in overgeneralization or omission of critical details in the images. To address this, our first step involves re-segmenting the source document into shorter sub-paragraphs, each expressing a single idea or thematic unit. Each segment is then paired with a corresponding illustration. Prior work in document-level text simplification has similarly emphasized the importance of managing information hierarchy and discourse structure, often through explicit structural analysis or summarization (Cripwell et al., 2023; Fang et al., 2025; Blinova et al., 2023). Our method automates this decomposition process, forming the foundation for downstream visual generation.

We employ GPT-4o as a zero-shot, discourse-aware segmenter. This decision is motivated by the emergent capabilities of LLMs to perform complex structural tasks without task specific fine-tuning. Unlike traditional unsupervised approaches that rely on shallow lexical cohesion signals, LLMs can exploit deep semantic and world knowledge to detect more nuanced topic boundaries (Mu et al., 2024). Using carefully designed prompts, we guide the model to function as an advanced textual analyzer. Because GPT-4o exhibits degraded performance when processing input text with long contexts, its effectiveness decreases as the length of the input increases (Karpinska et al., 2024;

Ma et al., 2024). To accommodate this context-length constraint, we implement a tokenizer-aware, sentence-preserving segmentation procedure that limits each model input to fewer than 20,000 characters. Specifically, we first compute the input length using a byte-pair–encoding (BPE) (Sennrich et al., 2016) tokenizer consistent with the model's tokenization scheme; if the character count exceeds 20,000, we partition the document into sub-documents. For sentence boundary detection, we adopt "Segment any Text" (Frohmann et al., 2024). To ensure machine-readability and robust integration with downstream components, we enforce output in JSON format. Under our tokenizer-aware JSON prompting, the model produced consistent discourse segments with low formatting error rates, which was sufficient for downstream components.

3.3 Visually Grounded Description Generation

Narrative language in documents often differs significantly from the concrete, descriptive phrasing required by T2I models to generate high-quality images (Saharia et al., 2022). Using raw text snippets as prompts frequently results in vague or abstract outputs. To bridge this semantic-to-visual gap, we introduce an intermediate transformation step, which we conceptualize as cross-modal abstraction. The goal is to distill the essential, visually representable elements from each text segment. This aligns with the broader goals of multimodal learning, where shared representation spaces enable meaningful alignment between textual and visual modalities.

We again leverage GPT-4o, configuring it as a text-to-text transformation agent for this task. Using few-shot prompting, we embed examples that guide the model to learn the desired input-output mapping without parameter updates, a practical benefit in our setting (few-shot prompting without task-specific finetuning) (Zhang and Xu, 2024). The prompt explicitly decomposes the task into two steps: (1) internal summarization to extract key entities and actions, and (2) translation of this summary into a visual scene description. We adopt a two-step prompt (key-entity summary $\rightarrow$ scene description) to make the transformation explicit; this yielded clearer, more actionable descriptions for image generation in our pilot settings (Wei et al., 2022).

A crucial component of our prompting strategy is the imposition of faithfulness constraints. We explicitly instruct the model to avoid hallucinations—i.e., adding objects, attributes, or details not present in the source text. By including directives such as "Do not invent or alter details not mentioned in the original text," we aim to minimize semantic drift and ensure that the final image is a faithful visual rendering of the textual document content.

3.4 Controlled Image Generation

Once visually grounded descriptions are prepared, the final step is image synthesis. The choice of T2I model architecture and the stylization of prompts play pivotal roles in determining the visual clarity, aesthetic quality, and suitability of the generated content for the target audience. For instance, illustrations for children's storybooks require an entirely different visual style than those in technical manuals. Recent advances in T2I, particularly diffusion models and their Transformer-based successors, have enabled unprecedented levels of photorealism and fine-grained style control (Betker et al., 2023).

In this study, we experiment with two state-of-the-art T2I models to explore architectural diversity: OpenAI's DALL·E 3 and Black Forest Labs's FLUX.1-dev. DALL·E 3 is renowned for its high fidelity to complex prompts, attributed largely to its use of a powerful language model to preprocess and enrich textual input prior to image generation (Betker et al., 2023). In contrast, FLUX.1-dev represents a new generation of diffusion/Transformer hybrids, making it a valuable comparative baseline.

To ensure stylistic consistency and interpretability across illustrations generated for the same document, we apply a stylistic modifier to each prompt. As an initial case study, we target the generation of "clear and concise cartoon-style illustrations." This stylistic choice is widely recognized for its ability to convey information clearly by abstracting over details, making it particularly suitable for educational and simplified communication contexts. The final prompt structure for image generation is a simple concatenation:"A clear and concise cartoon-style illustration depicting: <VISUAL DESCRIPTION>".

4 Experiments

4.1 Materials

We utilized the WebCorpus dataset (Battisti and Ebling, 2020), which is specifically designed for

automatic readability assessment and text simplification tasks in the German language. The corpus comprises approximately 6,200 documents and nearly 211,000 sentences, collected from the websites of governmental bodies, professional institutions, and non-profit organizations across Germany, Austria, and Switzerland, covering a total of 92 distinct domains. The data includes both HTML webpages and PDF files, with content dated between late 2018 and early 2019. In addition to providing parallel corpora and monolingual simplified German texts, the dataset is characterized by the preservation of text structure, typographic information, and embedded image content, which were structurally extracted using an HTML parser and PDFlib tools.

For our experiments, we constructed a specialized evaluation subset from the WebCorpus to test and assess our proposed method. We selected twelve parallel documents available in PDF format based on the following criteria:

Human-Generated Simplifications with Illustrations These documents not only contain simplified versions created by human experts but also include illustrations accompanying specific paragraphs in the simplified texts. As illustrated in Figure 1, these images are designed to visually explain or complement the core ideas of the corresponding textual segments.

Topical and Stylistic Diversity To ensure the generalizability of our evaluation, we deliberately selected documents covering a range of topics such as legal aid, public health guidelines, and social welfare application procedures. These documents also exhibit considerable variation in both the complexity of the source texts and the visual styles of the illustrations.

Although modest in scale, this twelve-document subset offers high-quality human annotations and rich internal diversity. It provides a rigorous and controlled experimental setting for end-to-end evaluation, allowing us to verify the full pipeline, from text segmentation and visual description generation to final image synthesis, and to conduct direct comparisons with the original simplified documents.

4.2 Experimental Procedure

The experimental procedure of this study adheres to the three-stage generation framework defined in Section 3, with the aim of validating the effectiveness of our proposed method in an end-to-end manner. The implementation consists of the following steps: We first extracted the plain text (.txt) versions of twelve selected PDF documents from the WebCorpus dataset as the original input texts. Following the method outlined in Section 3.1, we employed GPT-4o as a document segmentation module to automatically decompose each document into a series of semantically coherent sub-paragraphs. For each sub-paragraph, we invoked the GPT-4o model again based on the visual anchoring description strategy described in Section 3.2. Through chain-of-thought reasoning, the model distilled the abstract narrative of each sub-paragraph into concrete, renderable visual scene descriptions. These descriptions were generated under faithfulness constraints to prevent hallucinations or factual distortions. The generated visual descriptions were fed into both the DALL·E 3 and FLUX.1-dev image generation models. To ensure stylistic consistency and interpretability of illustrations throughout the document, we prefixed each description with a prompt specifying "a clear and concise cartoon-style illustration," as specified in Section 3.3. This approach yielded stylistically aligned images corresponding to each textual segment.

4.3 Evaluation

To conduct a comprehensive and reliable evaluation of the generated outputs, we employed expert human assessment. Four domain experts with backgrounds in simplified language were recruited to participate in the study. We developed a structured online questionnaire in which experts rated and qualitatively assessed the generated outputs. Original human-designed illustrations from the source documents were shown only as qualitative context when licensing permitted and were not included in the quantitative analysis. The evaluation focused on four key dimensions, including support for comprehension, semantic alignment with the input text, visual coherence, and style match. The primary goal of this assessment was to quantify the effectiveness of our method in terms of faithfulness, clarity, and aesthetic quality. The evaluation involved four text passages, each paired with two images generated by DALL·E 3 and FLUX.1-dev, resulting in eight image-text combinations. Three experimental conditions were considered: text only, text + image from FLUX.1-dev, and text + image from DALL·E 3. To ensure balanced exposure across conditions, we adopted a Latin square de-

sign.

Each evaluator completed a three-part evaluation. First, a brief pre-questionnaire collected background information such as years of professional experience and domain expertise. In the main evaluation, participants reviewed all conditions and rated four dimensions for each sample: (1) support for comprehension, (2) semantic alignment, (3) visual coherence, and (4) style match. the detailed definitions of the four dimensions in the Appendix A. Comprehension accuracy was additionally measured via multiple-choice questions as an objective check separate from the four subjective ratings.

Finally, a short post-questionnaire confirmed evaluators' understanding of the task and the rating criteria. Responses were collected using 5-point Likert scales for subjective measures (e.g., alignment and simplicity) and accuracy scores for comprehension questions. Each evaluator was required to complete a total of 99 questions. These included 80 questions pertaining to the evaluation of eight images, 12 questions assessing simplified texts, three questions concerning evaluators' background information, and four open-ended questions eliciting overall evaluations and feedback. The full survey design is provided in the appendix.

5 Results

5.1 Text Simplification Evaluation

We collected Likert ratings on a five-point scale from four expert evaluators along three criteria: Simplicity (Q4), Semantic Adequacy (Q5), and Fluency (Q6). Non-numeric "Other" entries were treated as missing for averaging and are reported separately in the distribution table. Table 1 summarizes, for each evaluator, the overall distribution of assigned scores across all texts and criteria, including a separate count for "Other" and the total numeric score. Table 2 reports per-criterion means with the effective sample size for each evaluator, together with the total numeric score and the overall mean across all available numeric ratings.

Fluency shows consistently high evaluations. Evaluator 4 attains the highest fluency mean and also the highest total score and overall mean. Simplicity varies more strongly by evaluator. Evaluator 3 tends to assign higher simplicity with a mean of 4.00, while Evaluator 2 assigns lower values with a mean of 2.25, which suggests different expectations for ease of reading. Semantic adequacy concentrates around mid to high values, with Evaluator 4 again providing the most favorable adequacy judgments. "Other" responses appear only for Evaluator 3 on Q6 and are excluded from all mean calculations by design. In aggregate, the results indicate robust perceived fluency, moderate to high semantic adequacy, and evaluator-dependent variation in perceived simplicity.

ID	1	2	3	4	5	Other	Total score
1	0	3	3	5	1	0	40
2	3	1	2	6	0	0	35
3	0	2	3	4	0	3	29
4	0	3	0	4	5	0	47

Table 1: Per-rater distribution of ratings across all texts and all three criteria. Columns "1–5" give counts of numeric ratings. "Other" counts non-numeric entries. "Total score" sums all numeric ratings for the rater.

ID	Q4 mean	Q5 mean	Q6 mean	Total score	Overall mean
1	2.75	3.00	4.25	40	3.33
2	2.25	2.50	4.00	35	2.92
3	4.00	2.50	3.00	29	3.22
4	3.00	4.00	4.75	47	3.92

Table 2: Per-rater means by criterion (Q4–Q6), along with each rater's total score (sum of all numeric ratings over Q4–Q6 and all texts) and overall mean (average of all available numeric ratings).

5.2 Illustration Generation Evaluation

We collected 64 image-related Likert responses per rater. Covering semantic alignment, support for comprehension, visual coherence, and stylistic appropriateness, under two models (FLUX.1-dev and DALL•E 3). Table 3 summarizes the per-rater tallies by score category, including non-numeric Others. Ratings for both systems concentrated in the 3–4 range, indicating that most illustrations were perceived as broadly supportive yet rarely exceptional; 5s were occasional, while 1s were rare. Raters 1–2 slightly preferred FLUX.1-dev (total differences within six points), whereas Raters 3–4 favored DALL•E 3, with Rater 4 assigning thirteen maximum scores to DALL•E 3.

Across all evaluators, ratings for both image generation systems were concentrated in the 3–4 range, with score 4 being the most frequently assigned. This indicates that the majority of generated images were perceived as broadly supportive of text comprehension and semantically adequate, but rarely outstanding. Scores of 5 were assigned only occasionally, reflecting the fact that few images were

ID	Model	Total	1	2	3	4	5	Others	Mean
1	flux.1-dev	92	6	6	7	12	1	0	3.29
	DALL·E 3	90	4	9	9	9	1	0	3.26
2	flux.1-dev	74	6	7	6	9	0	4	2.96
	DALL·E 3	69	8	9	5	7	0	3	2.84
3	flux.1-dev	58	11	5	7	4	0	5	2.39
	DALL·E 3	69	7	9	8	5	0	3	2.91
4	flux.1-dev	105	0	7	7	10	6	2	3.75
	DALL·E 3	116	2	9	1	7	13	0	3.97

Table 3: Aggregated Likert ratings (1–5) and non-numeric responses (*Others*) for image evaluation tasks across four evaluators, including the mean numeric score per model.

judged as fully satisfactory across all evaluative dimensions. Conversely, scores of 1 were rare, suggesting that completely inadequate outputs occurred only sporadically. Differences across evaluators were evident. Evaluators 1 and 2 awarded higher totals to flux.1-dev, the score difference between the two models was relatively small, with the total discrepancy remaining within six points. In contrast, Evaluators 3 and 4 assigned higher totals to DALL·E 3, with Evaluator 4 giving thirteen maximum ratings (5), far more than for flux.1-dev.

The questionnaire design allowed ratings to capture multiple facets of image quality. We conducted a closer comparison of evaluators' ratings across these different dimensions and found consistent patterns. On average, both systems received the highest scores on supporting text comprehension and semantic alignment with the text, where the majority of judgments fell between 3 and 4. By contrast, lower ratings (1–2) were more frequently observed in dimensions such as visual coherence and stylistic appropriateness, reflecting instances where images were perceived as misaligned in style or insufficiently coherent, even if they captured the general semantics of the text. This distribution indicates that the models were more successful in generating images that conveyed the intended meaning than in ensuring stylistic naturalness and visual consistency. In analyzing the responses to the question "Do you think this image was created by AI or manually?", we found that most evaluators misclassified the images generated by the FLUX.1-dev model as manually created rather than AI-generated. By contrast, the majority of evaluators correctly identified the images produced by DALL·E 3 as AI-generated. We also found a consistent issue emerged across specific combinations of the images with many texts, as shown in Figure 1. In these cases, all four experts independently highlighted the same concern: When the generated images contained too much text or overly intricate visual patterns, they became cognitively overwhelming.

6 Discussion and Conclusion

Our study indicates that automatic simplification produced texts that readers judged as fluent and largely faithful to source meaning, while perceived ease-of-reading is more sensitive to individual evaluator standards. On the visual side, Both image generation models were generally effective in providing semantically supportive illustrations, though their perceived utility varied across evaluators and dimensions. Ratings tended to cluster around the mid-scale (3–4), suggesting that while AI-generated visuals achieved a baseline adequacy, they rarely reached the level of high-quality human-created illustrations. This implies that the models captured textual meaning with reasonable reliability but seldom produced outputs regarded as exemplary in terms of clarity, stylistic appropriateness, or overall communicative effectiveness. Divergent preferences among evaluators, some favoring FLUX.1-dev and others DALL·E 3 highlight that judgments of quality are not solely determined by semantic accuracy, but are also shaped by individual aesthetic expectations and tolerance for stylistic variation. Particularly noteworthy was the tendency of participants to misclassify outputs from FLUX.1-dev as manually created, suggesting that its visual naturalness may enhance perceived authenticity. While such naturalness is promising for accessibility and engagement, it also raises potential concerns about transparency and user trust in contexts where it is important to distinguish human- from machine-generated content.

Qualitative comments about images containing too much text point to avoidable extraneous cognitive load. Images that incorporated excessive textual content or visually dense layouts were often judged as distracting, thereby diminishing rather than enhancing comprehension. This observation underscores the importance of maintaining visual simplicity, especially when designing for audi-

Figure 1: Examples of undesirable illustrations, where images contain excessive textual elements that make them confusing for the viewer.

ences with cognitive impairments. For such users, the trade-off between semantic fidelity and visual clarity becomes particularly critical; cluttered or overly detailed images may undermine the intended benefits of simplified language materials. These findings point to the necessity of introducing explicit design constraints in text-to-image workflows. Mechanisms such as filtering strategies, prompt engineering techniques that enforce minimalism, or post-processing methods to eliminate superfluous elements could help align outputs more effectively with accessibility goals.

The pattern of high fluency and adequate semantics in text, combined with mid-scale visual ratings, points to a practical synthesis: when textual simplification reliably preserves meaning and flow, illustrations function best as lightweight scaffolds rather than dense carriers of information. Simplified text can shoulder the primary communicative load, while images should reinforce key entities, relations, or processes without introducing visual clutter. This aligns with our finding that evaluators penalize visually dense layouts: if text is already fluent and semantically adequate, adding heavy captioning or intricate scene details may yield diminishing returns or even harm comprehension. Therefore, downstream design should prioritize (i) simplicity-first visual layouts, (ii) restrained use of textual overlays inside images, and (iii) explicit alignment between each image and a small set of core propositions in the simplified text.

Future research should extend beyond expert-based evaluations to incorporate direct feedback from end users, particularly individuals with cognitive impairments, in order to ensure that the generated visuals truly enhance comprehension and accessibility.

7 Limitations

While this study provides valuable evidence for the role of AI-generated images in supporting simplified text comprehension, several limitations remain. (1) The language scope of this study was limited to German, and all texts were drawn from a single corpus of simplified expository materials. This may constrain the generalizability of our findings to other languages, genres, or cultural contexts. Expanding to multilingual or narrative datasets could uncover additional design considerations. (2) Visual complexity was identified as a recurring issue, but our analysis relied on qualitative judgments rather than formal cognitive load metrics. The absence of behavioral or physiological measures (e.g., comprehension scores, reading time, or gaze data) limits our ability to precisely quantify the cognitive effects of visual detail. (3) The image generation models used (DALL·E 3 and FLUX.1-dev) were not fine-tuned for the simplified language setting or for accessibility-related constraints. As a result, the outputs may occasionally include dense textual overlays or unnecessary visual embellishments. Future research should explore prompt engineering

techniques and post-processing methods to explicitly control for simplicity and semantic salience.

8 Lay Summary

This paper presents a practical method to make long and complex documents easier to understand, especially for readers with cognitive impairments. Unlike most tools that simplify single sentences, our approach operates at the document level and adds supportive illustrations. First, a large language model (GPT-4o) divides a document into short, coherent segments and rewrites them in simpler language while preserving meaning. Next, the model drafts faithful visual descriptions for each segment, and state-of-the-art image generators (DALL·E 3 and FLUX.1-dev) produce clear, consistent cartoon-style illustrations that align with the simplified text.

We evaluated the pipeline on a curated subset of real German public-information documents and asked four experts in simplified language to review the outputs. Their ratings clustered around the middle of the 5-point scale, indicating that the images generally helped comprehension but were not uniformly excellent. Reviewers also noted that illustrations overloaded with on-image text can increase cognitive load and reduce clarity, underscoring the value of minimal, consistent visuals.

Pairing document-level simplification with faithful, stylistically coherent illustrations appears promising for making public-facing materials, such as health guidance or social-service instructions more accessible. Future work will expand user studies with target populations and further constrain visual design to keep images simple, readable, and trustworthy.

References

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: A survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Miriam Anschütz, Tringa Sylaj, and Georg Groh. 2024. Images speak volumes: User-centric assessment of image generation for accessible communication. *arXiv preprint arXiv:2410.03430*.

Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluis Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, and 1 others. 2024. Imagen 3. *arXiv preprint arXiv:2408.07009*.

Alessia Battisti and Sarah Ebling. 2020. A corpus for automatic readability assessment and text simplification of german. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 3306–3314, Marseille, France. European Language Resources Association.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Sofia Blinova, Xinyu Zhou, Martin Jaggi, Carsten Eickhoff, and Seyed Ali Bahrainian. 2023. Simsum: Document-level text simplification via simultaneous summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9927–9944.

Ursula Bredel and Christiane Maaß. 2016. *Leichte Sprache: Theoretische Grundlagen? Orientierung für die Praxis*. Duden.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Document-level planning for text simplification. In *17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006. Association for Computational Linguistics.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR.

Dengzhao Fang, Jipeng Qiang, Yi Zhu, Yunhao Yuan, Wei Li, and Yan Liu. 2025. Progressive document-level text simplification via large language models. *arXiv preprint arXiv:2501.03857*.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. *arXiv preprint arXiv:2406.16678*.

Mohamed Gado, Towhid Taliee, Muhammad Memon, Dmitry Ignatov, and Radu Timofte. 2025. Vist-gpt: Ushering in the era of visual storytelling with llms? *arXiv preprint arXiv:2504.19267*.

A. Glenberg and William E. Langston. 1992. Comprehension of illustrated text: Pictures help to build mental models. *Journal of Memory and Language*, 31:129–151.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

Daibao Guo, Shuai Zhang, Katherine Landau Wright, and E. McTigue. 2020. Do you get the picture? a meta-analysis of the effect of graphics on reading comprehension. *AERA Open*, 6.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One thousand and one pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jorge Leandro, Sudha Rao, Michael Xu, Weijia Xu, Nebojsa Jojic, Chris Brockett, and Bill Dolan. 2024. Geneva: Generating and visualizing branching narratives using llms. In *2024 IEEE Conference on Games (CoG)*, pages 1–5. IEEE.

Y. Lin, Ting-Fang Wu, Ya-Hui Tasi, Hui-Ching Chen, and Ming-Chung Chen. 2009. The effect of different representations on reading digital text for students with cognitive disabilities. *Br. J. Educ. Technol.*, 40:764–770.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiao wen Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *ArXiv*, abs/2407.01523.

Yida Mu, Chun Dong, Kalina Bontcheva, and Xingyi Song. 2024. Large language models offer an alternative to the traditional approach of topic modelling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 10160–10171.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, and 1 others. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Mike Zheng Shou, Di Zhang, Chunhua Shen, Zhongyuan Wang, Lele Cheng, Yan Li, Weijia Wu, Yefei He, Tingting Gao, and Zhuang Li. 2023. Paragraph-to-image generation with information-enriched diffusion model. *ArXiv*, abs/2311.14284.

R. Sutherland and T. Isherwood. 2016. The evidence for easy-read for people with intellectual disabilities: A systematic literature review. *Journal of Policy and Practice in Intellectual Disabilities*, 13:297–310.

Yuanzhe Wang and Zheng Zewen. 2023. An eye-tracking based study: The role of images and explanatory texts in reading comprehension. *International Journal of Frontiers in Sociology*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jennifer Winberg and Meredith Saletta. 2018. Leveled texts for adults with intellectual or developmental disabilities: A pilot study. *Focus on Autism and Other Developmental Disabilities*, 34:118 – 127.

Chao Zhang and Shangqing Xu. 2024. Misconfidence-based demonstration selection for llm in-context learning. *ArXiv*, abs/2401.06301.

Yangqianhui Zhang, Qinghao Hu, Liang Zhao, Chunjiang Fu, Tengtuo Chen, and Pingda Huang. 2024. Multi-sentence complementarily generation for text-to-image synthesis. *IEEE Transactions on Multimedia*, 26:8323–8332.

A Operational Definitions.

We use the following operational definitions for rated dimensions: *support for comprehension* = perceived helpfulness of the image for understanding the passage's main message; *semantic alignment* = fidelity of depicted entities, attributes, and relations to the source passage; *visual coherence* = absence of distracting artifacts or clutter and a legible, uncluttered composition within the image; *style match* = appropriateness and consistency of the visual style with the document's genre and Easy-to-Read conventions.

B Prompt Templates

We include here the prompts used for text simplification, text segmentation and visual description generation. All prompts are designed to enforce output constraints (JSON formatting) and minimize semantic drift in multimodal generation.

B.1 Prompt for Text Simplification

`You are a professional editor specializing in document-level text simplification for broader accessibility.`
Your goals are: (1) preserve meaning and factual correctness, (2) increase readability and accessibility, and (3) maintain discourse-level coherence across sections.

Follow the steps and constraints below exactly. Do not hallucinate or omit essential information.

0) Controls (adjust these before running)
Target audience: general adult readers without domain expertise.
Readability target: approximately B1–B2 (plain language); avoid jargon unless defined.
Lexical simplicity: prefer high-frequency, concrete words; define any necessary technical terms briefly.
Syntactic simplicity: prefer simple main clauses; split long/complex sentences (>25–30 words).
Style: neutral, clear, consistent; no rhetorical questions; active voice where appropriate.

1) Plan at the Document Level (no output yet)
Produce a hidden plan to guide rewriting (do not include the plan in the final answer):
Section map: list sections/paragraphs and their main points.
Entity & timeline register: people, organizations, quantities, dates, and events; ensure consistency of names/abbreviations across the document.
Discourse links: for each section, note how it connects to the previous one (cause $\rightarrow$ effect, problem $\rightarrow$ solution, comparison, sequence, contrast).
Risk items: legal/medical/financial claims; numbers, percentages, dates, and units that must remain exact.

2) Rewrite Rules (apply throughout)
Meaning preservation: keep all factual statements, numbers, dates, and units; do not invent content; do not change scope or evidential hedges.
Sentence-level operations: (a) split long sentences; (b) delete redundancy and filler; (c) paraphrase rare idioms and nominalizations; (d) reorder for subject–verb proximity.
Lexical operations: replace rare words with common alternatives; define unavoidable terms in-line the first time they appear.
Coreference & cohesion: resolve ambiguous pronouns; repeat a short, clear noun phrase when needed; add explicit connectives (e.g., "However," "As a result," "In addition") to preserve coherence across sentences and sections.
Structure: keep informative headings; convert dense lists into bullets or tables where it improves clarity; keep citations/references but simplify their surrounding prose.
Safety & integrity: never remove warnings, limitations, or risk qualifiers; never alter quoted material; keep figure/table references consistent.

3) Self-Review Checklist (enforce before finalizing)
Confirm all items; if any fail, revise and re-check:
Meaning preservation: Each paragraph answers the same questions as the source; all numbers/dates/units/entities match the original.
Readability & simplicity: Average sentence length reduced; complex clauses minimized; jargon defined or replaced.
Document coherence: Section openings include bridging phrases; topic flow is consistent; pronouns are unambiguous.
Style consistency: Tone and tense are consistent; active voice is used where natural; no rhetorical filler.

Table 4: Prompt used for text simplification

B.2 Prompt for Text Segmentation

This prompt guides the model to identify thematic and semantic boundaries in expository texts and return machine-readable subparagraphs. The output format is constrained to valid JSON to ensure compatibility with downstream modules in our pipeline.

Please analyze the following text and split it into coherent subparagraphs based on thematic and semantic boundaries. Follow these rules strictly:
1. Output MUST be valid JSON format only
2. Use numbered keys starting from "1"
3. Ensure all strings are properly quoted
4. Escape any internal double quotes
5. Do NOT include any additional text or explanations
6. Maintain original content integrity

Text to process: {text}

Output format example:

{"1": "first subparagraph text", "2": "second subparagraph text"}

Table 5: Prompt used for text segmentation

B.3 Prompt for Visual Description Generation

This prompt constrains the model to produce faithful visual descriptions of the segmented subparagraphs. A key element is the explicit instruction to avoid hallucination, ensuring that no visual elements are introduced beyond the source text.

Please generate visual descriptions for each subparagraph following these steps:

1. Create a brief summary highlighting main content
2. Convert summaries into visual descriptions suitable for image generation. Ensure that the visual description faithfully represents the original text without adding or altering objects, attributes, or details not present in the source. Maintain semantic accuracy while simplifying the expression for better clarity.
3. Return ONLY JSON with subparagraph numbers and visual description. Do not include any additional text or explanations.
Output format example:

{"1": "Peaceful countryside with green fields and cottages", "2": "Busy city street with neon lights"}

Table 6: Prompt used for visual description generation

C Example of WebCorpus

As part of the WebCorpus dataset, we include authentic examples of German documents written in Leichte Sprache (easy-to-read German). One representative source is the newsletter series Bericht aus Genf [1].

[1] https://www.bodys-wissen.de/bericht-aus-genf.html

Example of simplified German text from WebCorpus
Bericht aus Genf Nr. 8 / 2014 Newsletter von Theresia Degener Mitglied im Ausschuss für den UN-Vertrag über die Rechte von Menschen mit Behinderungen Begrüßung Dieser Info-Brief ist über die 12. Sitzung von unserer Arbeits-Gruppe in Genf. Vor dem Treffen habe ich gedacht: Das ist das letzte Mal für mich. Die Mitglieder in der Arbeits-Gruppe arbeiten immer 4 Jahre mit. Und ich bin schon 4 Jahre dabei. Aber im Juni ist etwas Schönes passiert: Es waren Wahlen für die Arbeits-Gruppe. Und ich wurde wieder-gewählt. Das bedeutet: Ich darf noch einmal 4 Jahre in der Arbeits-Gruppe mitmachen. Darüber freue ich mich sehr. Seit November gibt es eine Sonder-Bericht-Erstatterin für die Rechte von Menschen mit Behinderungen. Sie arbeitet für den Menschen-Rechts-Rat bei den Vereinten Nationen. Das ist die Aufgabe von der Sonder-Bericht-Erstatterin: Sie schreibt Berichte für den Menschenrechts-Rat: Wie geht es Menschen mit Behinderungen auf der ganzen Welt. 2 Die Sonder-Bericht-Erstatterin heißt: Catalina Devantas. Sie kennt sich sehr gut aus mit Menschen-Rechten. Sie ist Rollstuhl-Fahrerin und Mutter von drei Kindern. Auf dem Bild bin ich mit Catalina Devantas und ihren Kindern und mit Silvia Judith Quan-Chang. Sie ist auch in der Arbeits-Gruppe. Bei der 12. Sitzung hatten wir einen Gast aus Mexiko. Er heißt Ricardo Adair Rubles. Er ist ein Mann mit Lern-Schwierigkeiten. Ricardo Adair Rubles ist sehr mutig. Das hat er der Arbeits-Gruppe erzählt: Er kämpft gegen ein Gesetz in Mexiko. In dem Gesetz steht: Menschen mit Lern-Schwierigkeiten dürfen nicht selbst entscheiden. Das ist schlecht. Jeder Mensch darf selbst entscheiden. Manchmal braucht ein Mensch Unterstützung. Dann muss der Mensch die Unterstützung bekommen. Ich wünsche Ihnen eine schöne Weihnachts-Zeit und ein gutes Jahr 2015! Theresia Degener. Ein Tipp zum Lesen: Der Bericht ist in Leichter Sprache. Aber manchmal stehen auch schwere Wörter in dem Bericht. Diese Wörter sind dann dick und blau. Die Wörter werden im Text erklärt.

Table 7: An example of simplified German text, extracted from the first two pages of Bericht aus Genf 8 Leichte Sprache.pdf

D Segmentation and Visual Descriptions for the WebCorpus Example

This appendix reports the output of our processing pipeline applied to the simplified German document in Appendix B.

ID	Segment (Simplified German)	Visual Description
1	Bericht aus Genf Nr. 8 / 2014 Newsletter von Theresia Degener Mitglied im Ausschuss für den UN-Vertrag über die Rechte von Menschen mit Behinderungen Begrüßung Dieser Info-Brief ist über die 12. Sitzung von unserer Arbeits-Gruppe in Genf. Vor dem Treffen habe ich gedacht: Das ist das letzte Mal für mich. Die Mitglieder in der Arbeits-Gruppe arbeiten immer 4 Jahre mit. Und ich bin schon 4 Jahre dabei. Aber im Juni ist etwas Schönes passiert: Es waren Wahlen für die Arbeits-Gruppe. Und ich wurde wieder-gewählt. Das bedeutet: Ich darf noch einmal 4 Jahre in der Arbeits-Gruppe mitmachen. Darüber freue ich mich sehr.	Theresia Degener freut sich über ihre Wiederwahl in die Arbeits-Gruppe.
2	Seit November gibt es eine Sonder-Bericht-Erstatterin für die Rechte von Menschen mit Behinderungen. Sie arbeitet für den Menschen-Rechts-Rat bei den Vereinten Nationen. Das ist die Aufgabe von der Sonder-Bericht-Erstatterin: Sie schreibt Berichte für den Menschenrechts-Rat: Wie geht es Menschen mit Behinderungen auf der ganzen Welt. Die Sonder-Bericht-Erstatterin heißt: Catalina Devantas. Sie kennt sich sehr gut aus mit Menschen-Rechten. Sie ist Rollstuhl-Fahrerin und Mutter von drei Kindern. Auf dem Bild bin ich mit Catalina Devantas und ihren Kindern und mit Silvia Judith Quan-Chang. Sie ist auch in der Arbeits-Gruppe.	Catalina Devantas, eine Rollstuhlfahrerin und Mutter, arbeitet als Sonder-Bericht-Erstatterin.
3	Bei der 12. Sitzung hatten wir einen Gast aus Mexiko. Er heißt Ricardo Adair Rubles. Er ist ein Mann mit Lern-Schwierigkeiten. Ricardo Adair Rubles ist sehr mutig. Das hat er der Arbeits-Gruppe erzählt: Er kämpft gegen ein Gesetz in Mexiko. In dem Gesetz steht: Menschen mit Lern-Schwierigkeiten dürfen nicht selbst entscheiden. Das ist schlecht. Jeder Mensch darf selbst entscheiden. Manchmal braucht ein Mensch Unterstützung. Dann muss der Mensch die Unterstützung bekommen.	Ricardo Adair Rubles kämpft mutig gegen ein Gesetz in Mexiko.
4	Ich wünsche Ihnen eine schöne Weihnachts-Zeit und ein gutes Jahr 2015! Theresia Degener Ein Tipp zum Lesen: Der Bericht ist in Leichter Sprache. Aber manchmal stehen auch schwere Wörter in dem Bericht. Diese Wörter sind dann dick und blau. Die Wörter werden im Text erklärt.	Theresia Degener wünscht frohe Weihnachten und gibt einen Lesetipp.

Table 8: Segmentation (IDs 1–4) and automatically generated visual descriptions for the simplified German text from appendix B

E Generated Images from Visual Descriptions

This appendix presents generated images corresponding to the visual descriptions in Appendix C. We compare two text-to-image models, DALL·E 3 and FLUX.1-dev, across four descriptions (IDs 1–4).

DALL·E 3

FLUX.1-dev

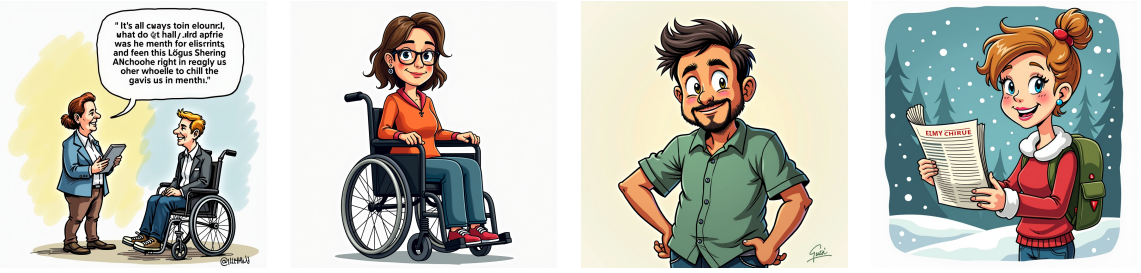

Figure 2: Comparison grid of generated images using DALL E 3 and FLUX.1-dev for four visual descriptions from Appendix C (IDs 1–4). Columns map left-to-right to description IDs 1, 2, 3, and 4.

F Questionnaire for Image Evaluation

This appendix presents the full bilingual questionnaire used to evaluate AI-generated and manually created images accompanying simplified texts. It consists of a brief pre-questionnaire about participants' background, a main questionnaire covering text comprehension, text-image alignment, and image quality, followed by a short post-study comparison and an open feedback item. Items explicitly marked for conditions (b) and (c) apply only when a text is shown with an image. Items without such a marker apply to all presentation conditions, including text-only. For check items, respondents select exactly one option unless otherwise stated; free-text fields are provided for comments.

Legend of Conditions / Legende der Bedingungen. (a) Text only / Nur Text (b) Text + AI-generated image / Text + KI-generiertes Bild (c) Text + manually created image / Text + manuell erstelltes Bild

Pre-Questionnaire / Vorbefragung

Question 1 / Frage 1. How many years of experience do you have working as a simplified language expert? / Wie viele Jahre Erfahrung haben Sie als Expert:in für vereinfachte Sprache? (select one / eine Option wählen)

- 0–1 years / 0–1 Jahr
- 1–2 years / 1–2 Jahre
- 3–5 years / 3–5 Jahre
- More than 5 years / Mehr als 5 Jahre
- Other (please specify) / Andere (bitte angeben): _____

Question 2 / Frage 2. How many years of experience do you have evaluating images as part of simplified language? / Wie viele Jahre Erfahrung haben Sie in der Bewertung von Bildern im Kontext vereinfachter Sprache? (select one / eine Option wählen)

- 0–1 years / 0–1 Jahr
- 1–2 years / 1–2 Jahre
- 3–5 years / 3–5 Jahre
- More than 5 years / Mehr als 5 Jahre
- Other (please specify) / Andere (bitte angeben): _____

Question 3 / Frage 3. What is your work setting in this field? / In welchem Arbeitsverhältnis sind Sie in diesem Bereich tätig? (select one / eine Option wählen)

- Freelancer / Freiberuflich
- Employee at a company / Angestellt in einem Unternehmen
- Employee at a research institute / Angestellt in einem Forschungsinstitut
- Employee at an association / public sector organization / Angestellt in einem Verband / einer öffentlichen Einrichtung
- Other (please specify) / Andere (bitte angeben): _____

Main Questionnaire: Evaluation of AI-Generated Images and Text-Image Combination / Hauptfragebogen: Bewertung von KI-generierten Bildern und Text-Bild-Kombinationen

Section 1 / Abschnitt 1: Overall Text Comprehension / Textverständnis (all conditions a–c / alle Bedingungen a–c) Question 4 / Frage 4. How simple is this text? / Wie einfach ist dieser Text?

- 1 = Very difficult / Sehr schwierig
- 2 = Somewhat difficult / Eher schwierig
- 3 = Neutral / Neutral
- 4 = Somewhat easy / Eher einfach
- 5 = Very easy / Sehr einfach
- Other (please specify) / Andere (bitte angeben): _____

Question 5 / Frage 5. Is the text semantically adequate? / Ist der Text semantisch angemessen?

- 1 = Not at all / Überhaupt nicht
- 2 = Mostly not / Meistens nicht
- 3 = Partially / Teilweise
- 4 = Mostly / Größtenteils
- 5 = Completely / Vollständig
- Other (please specify) / Andere (bitte angeben): _____

Question 6 / Frage 6. Is the text fluent / grammatical? / Ist der Text flüssig / grammatikalisch korrekt?

- 1 = Not at all / Überhaupt nicht
- 2 = Mostly not / Meistens nicht
- 3 = Partially / Teilweise
- 4 = Mostly / Größtenteils
- 5 = Completely / Vollständig
- Other (please specify) / Andere (bitte angeben): _____

Section 2 / Abschnitt 2: Text-Image Alignment / Text-Bild-Übereinstimmung (conditions b–c / Bedingungen b–c) Question 7 / Frage 7. Does the image enhance the understanding of the text? / Unterstützt das Bild das Verständnis des Textes?

- 1 = Not at all / Überhaupt nicht
- 2 = Mostly not / Meistens nicht
- 3 = Partially / Teilweise
- 4 = Mostly / Größtenteils
- 5 = Completely / Vollständig
- Other (please specify) / Andere (bitte angeben): _____

Question 8 / Frage 8. How well does the text align with the image (meaning, message)? / Wie gut stimmt der Text in Bedeutung und Botschaft mit dem Bild überein?

- 1 = Not aligned at all / Überhaupt nicht übereinstimmend
- 2 = Mostly not aligned / Meistens nicht übereinstimmend
- 3 = Partially aligned / Teilweise übereinstimmend
- 4 = Mostly aligned / Größtenteils übereinstimmend
- 5 = Completely aligned / Vollständig übereinstimmend
- Other (please specify) / Andere (bitte angeben): _____

Question 9 / Frage 9. Which type of image do you think was used? / Was glauben Sie, welche Art von Bild verwendet wurde?

- AI-generated / KI-generiert
- Manually created / Manuell erstellt
- Unsure / Unsicher
- Other (please specify) / Andere (bitte angeben): _____

Section 3 / Abschnitt 3: Image Quality Evaluation / Bildqualitätsbewertung (conditions b–c / Bedingungen b–c) **Question 10 / Frage 10.** How visually coherent are the images? / Wie visuell kohärent ist das Bild?

- 1 = Not coherent at all / Überhaupt nicht kohärent
- 2 = Mostly not coherent / Meistens nicht kohärent
- 3 = Partially coherent / Teilweise kohärent
- 4 = Mostly coherent / Größtenteils kohärent
- 5 = Completely coherent / Vollständig kohärent
- Other (please specify) / Andere (bitte angeben): _____

Question 11 / Frage 11. What is the function of the image relative to the text? / Welche Funktion hat das Bild in Bezug auf den Text?

- "Expansion" / Expansion
- "Exemplification" / Exemplifikation
- "Explication" / Explikation
- "Condensation" / Kondensation
- Other (please specify) / Andere (bitte angeben): _____

Reference within item / Referenz im Item: see (**??**).

Question 12 / Frage 12. How well does the image fulfill this function relative to the text? / Wie gut erfüllt das Bild diese Funktion in Bezug auf den Text?

- 1 = Not at all / Überhaupt nicht
- 2 = Mostly not / Meistens nicht
- 3 = Partially / Teilweise
- 4 = Mostly / Größtenteils
- 5 = Completely / Vollständig
- Other (please specify) / Andere (bitte angeben): _____

Question 13 / Frage 13. How well does the image style match the text? / Wie gut passt der Bildstil zum Text?

- 1 = Not natural or pleasing at all / Überhaupt nicht natürlich oder ansprechend
- 2 = Mostly not natural or pleasing / Größtenteils nicht natürlich oder ansprechend
- 3 = Partially natural and pleasing / Teilweise natürlich und ansprechend
- 4 = Mostly natural and pleasing / Größtenteils natürlich und ansprechend
- 5 = Completely natural and pleasing / Vollständig natürlich und ansprechend
- Other (please specify) / Andere (bitte angeben): _____

Post-study Questionnaire / Nachbefragung

Section 4 / Abschnitt 4: Comparison of Image Conditions / Vergleich der Bildbedingungen (for b & c / für b & c) Question 14 / Frage 14. Which type of image do you find more useful? / Welche Art von Bild empfinden Sie als nützlicher?

- AI-generated images / KI-generierte Bilder
- Manually created images / Manuell erstellte Bilder
- No significant difference / Kein signifikanter Unterschied

Question 15 / Frage 15. Which type of image do you find more visually appealing? / Welche Art von Bild empfinden Sie als visuell ansprechender?

- AI-generated images / KI-generierte Bilder
- Manually created images / Manuell erstellte Bilder
- No significant difference / Kein signifikanter Unterschied

Question 16 / Frage 16. Which type of image best supports comprehension of the text? / Welche Art von Bild unterstützt das Textverständnis am besten?

- AI-generated images / KI-generierte Bilder
- Manually created images / Manuell erstellte Bilder
- No significant difference / Kein signifikanter Unterschied

Section 5 / Abschnitt 5: Open Feedback / Offenes Feedback

Question 17 / Frage 17. Do you have any comments or suggestions on the text, images, or their combination? / Haben Sie Kommentare oder Anregungen zum Text, zu den Bildern oder zu deren Kombination?

- Response / Antwort:_____