Presentation Slide Translation and Layout Error Correction by LLMs

Futo Kajita[†], **Nobuyori Nishimura**[†], **Takehito Utsuro**[†],
Naoki Muto[‡], **Chee Siang Leow**[‡], **Hiromitsu Nishizaki**[‡]

[†]The Graduate School of Science and Technology, University of Tsukuba
[‡]The Graduate Faculty of Interdisciplinary Research, University of Yamanashi

{s2620757, s2620807}@u.tsukuba.ac.jp utsuro@iit.tsukuba.ac.jp
naoki_m@alps-lab.org {leow, hnishi}@yamanashi.ac.jp

Abstract

We propose a novel approach to translating Japanese slides into English and to correcting their layout errors by utilizing multimodal LLMs with slide images and XML structures. Existing translation tools often suffer from layout errors after translation due to text expansion during the translation process, causing text to overlap with figures or other items in slides and thereby reducing readability. To overcome this issue, our proposed framework introduces two steps consisting of (i) translating text fragments within the slide, and (ii) correcting layout errors by optimizing layout placement based on visual consistency. In step (ii), we empirically show that few-shot prompts are quite effective in layout error correction. Given that the optimal combination of steps (i) and (ii) varies depending on the slide layout, our method generates eight different layout candidates. Consequently, we introduce a third step that automatically selects the optimal output from these eight candidates. The experimental results showed that the proposed method outperformed baselines and achieved 4.1% layout error rate and over 80% model selection success rate.

1 Introduction

With the rapid increase of globalization, Japanese universities accept more and more international students. However, many of these students lack sufficient Japanese proficiency, suffering from insufficient comprehension in classes taught in Japanese language. To address this, many universities have adopted measures such as offering courses in English.

Considering this background, we aim at developing a "language barrier free lecture support system" that integrates AI technology with smart glasses (Figure 1). In lectures in classrooms, "lecture slides" serve as an information source with importance equal to the instructor's speech. Thus, their translation from Japanese to English is equally important as the translation of the instructor's speech from Japanese to English. However, the issue of translating slides involves inherent problems that cannot be solved by text translation alone. Due to differences in text volume between languages, just by simply translating text contents from one language into another, layout issues such as text overlapping with other items frequently occur, failing to preserve the intended layout of the original slide after translation across languages.

To address this challenge, simply applying existing content generation or recognition technologies is insufficient. For instance, recent studies such as PPTBench (Huang et al., 2025) and AutoPresent (Ge et al., 2025) primarily focus on generating slides from scratch based on textual instructions. In those slide generation tasks, layout of generated objects can be optimized without any restriction, while in slide translation task, layout restriction within the source language slide does exist even after text content translation from the source language into the target language. Unlike slide generation tasks, the slide translation system must accommodate expanded translated text within the rigid geometric constraints of the source language slide. The primary challenge is not creating a new aesthetic layout, but fitting dynamic text into static containers without violating the geometrical constraint of the object layout within the source language slide. Furthermore, in lecture slides, multiple items shown in each slide often have strong dependencies among each other (e.g., a caption placed precisely next to a specific data point in a chart). On the contrary, although previously studied layout generation models like LayoutDM (Inoue et al., 2023) excel at synthesizing as a whole aesthetic layouts from scratch, they are ill-suited for the task of slide translation due to the loss of dependencies among items shown in each slide. Therefore, instead of regenerating the global lay-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 326–339

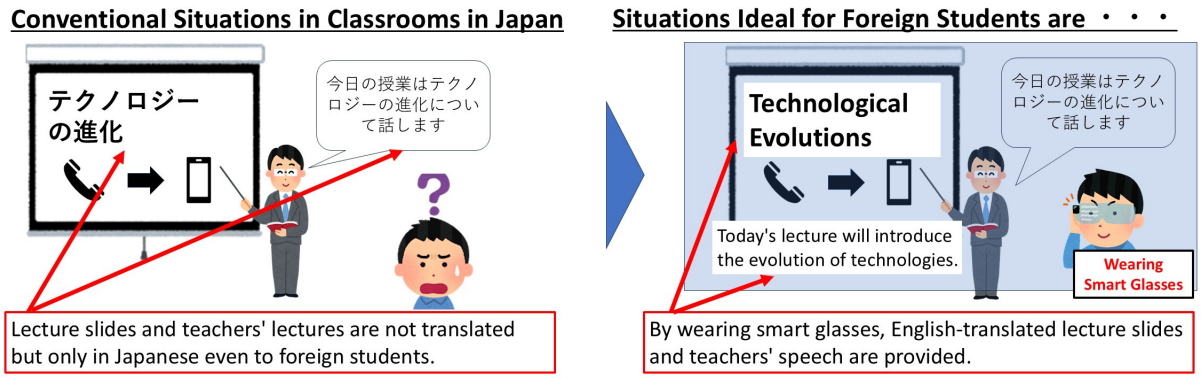

Figure 1: Proposed System of Translating Lectures given in Japanese into English

out from scratch, our approach adopts a strategy of integrating the sequence of minimal layout modification, each of which is performed by applying localized edits to the XML structure only where necessary to resolve overlaps caused by the slide translation operations, thereby strictly preserving the semantic integrity of the original slide layout.

More concretely, this paper describes a method for translating slides by directly parsing and editing the internal XML structure (ECMA International, 2021). This approach ensures that even items with complex layouts are translated without decreasing visibility and the internal XML structures also remain in their editable formats through the whole slide translation operation. Figure 2 presents the three steps in the overall framework of our proposed approach. Our approach consists of a pipeline that utilizes LLMs to generate eight candidates based on the combinations of several configurations in the variants of inputs[1] and subsequently selects the optimal one. To overcome layout issues caused by text expansion, we introduce two steps consisting of (i) translating text fragments shown in slides, and (ii) correcting layout errors by optimizing layout placement based on visual consistency. The first step performs text translation alongside the first operation for correcting layout errors, while the second step executes the layout edit operations for the remaining layout errors. In this second step (ii) of layout edit, we incorporate few-shot prompting, as shown in Figure 4, to instruct specific edit operations such as "font resizing", "text box reshaping", "text box decomposing". Those explicit few-shot promptings are provided to instruct the LLM on the intended edit, thereby enhancing its ability to resolve complex layout errors. We prepared eight variants of the combinations of (i) and (ii), considering that the optimal strategy may vary according to layout variations. Consequently, we further introduce an LLM-based third step of automatically selecting an optimal translated English slide among those eight candidates, enabling the delivery of high quality translated slides that maintain international consistency. The contributions of this paper are summarized by addressing the following two research questions (RQs):

RQ 1 : Which framework is the most suitable for solving the two tasks of slide text translation and layout error correction that might have certain (and not small) interferences on each other?

Contribution 1 : We established an architecture that distinctly separates the roles of the LLM into two steps: translating text contents, and subsequently correcting layout errors caused by the text expansion during translation.

RQ 2 : How can we qualify and quantify the human perceived visual quality of the results of performing the two tasks of slide text translation and layout error correction, where no existing metric has been studied previously?

Contribution 2 :

Layout error rate : As the third step coming after previously mentioned first step of slide translation and the second step of layout error edit, this paper demonstrated that the optimal slide selection from 8 candidates performed extremely well achieving low layout error rate of 4.1%.

Model selection success rate : As the result of comparing the outputs from baselines and the proposed method and then selecting the optimal one, the outputs by the proposed method were selected

[1]In this paper, the term *layout error correction* is defined as representing the notion of correcting layout error, while the term *layout edit* is defined as representing the operation of editing the layout where the layout may include errors and henceforth the layout edit operation may correct those errors.

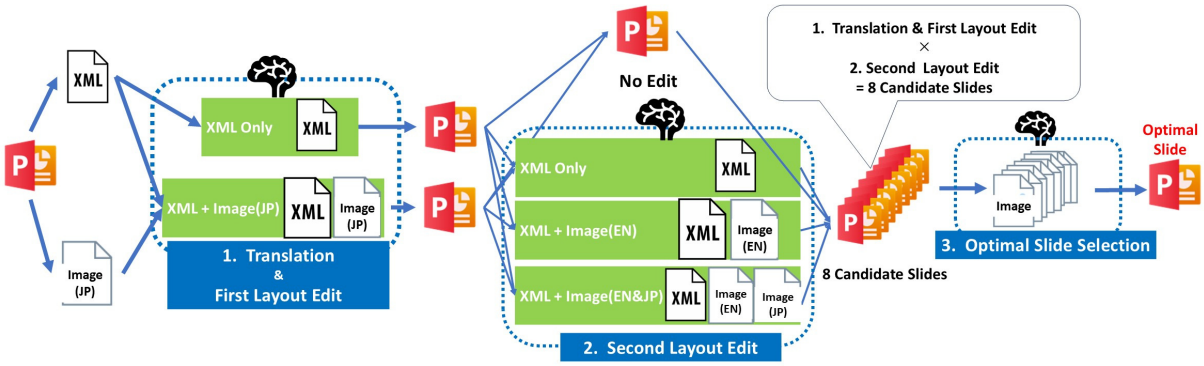

Figure 2: Optimal Slide Selection Aggregating the Results from "1. Translation + First Layout Edit" and "2. Second Layout Edit"

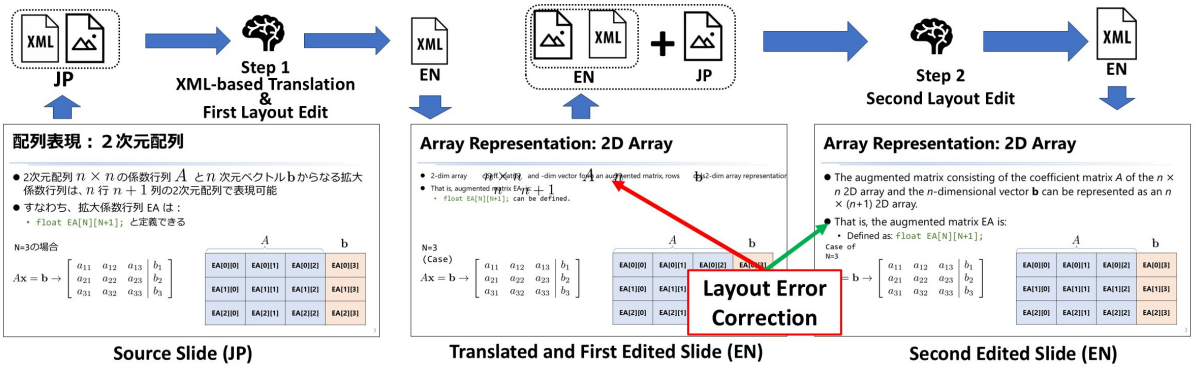

Figure 3: Overview of the Proposed Slide Translation and Layout Edit Pipeline (The Japanese source slide (left) is translated into English and layout edited (first time) (middle), and again layout edited (second time) (right), which is the final output English translated slide without any layout error.)

in over 80.0% of the evaluation.

2 Related Work

Methods for slide generation based document summarization (Bandyopadhyay et al., 2024; Sun et al., 2021; Mondal et al., 2024) and advanced generation frameworks utilizing multimodal agents (Xu et al., 2025; Zheng et al., 2025; Tang et al., 2025) have been established. These approaches excel at generating visually appealing slides from scratch based on textual or visual instructions. However, these methods assume slide generation from scratch and do not address the requirements essential for translation tasks such as replacing and adjusting text contents while preserving the strict XML structure given with the original source language slide.

In the field of computer vision, models such as LayoutGPT (Feng et al., 2023), as well as understanding and recognition models like Monkey (Li et al., 2024), TrOCR (Li et al., 2023), and PIX2STRUCT (Zhu et al., 2024), have proposed techniques of document structure analysis and high precision information extraction. However, the former is not designed to preserve the semantic arrangement of existing items in each source language slide, while the latter cannot output editable PowerPoint files (XML). Therefore, these technologies cannot be directly applied to the objective of this study, which is the layout edit of existing files in the course of slide translation and layout error correction.

BOOM (Beyond Only One Modality) (Koneru et al., 2025), which specializes in slide translation, is an approach that leverages visual information to improve the quality of translated text. However, it does not possess a mechanism to resolve, at the structural level, issues such as item overlap and layout error caused by the increase in text volume during translation.

3 Language Barrier Free Lecture Support System

This paper aims at constructing a language barrier free classroom environment where international students can attend lectures taught in Japanese

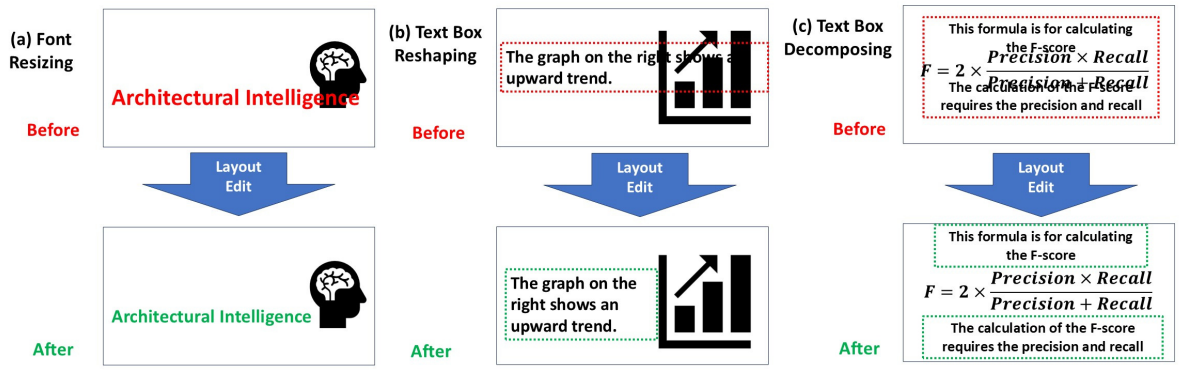

Figure 4: Few-shot Examples of Layout Edit Operations of 3 Types

(or other languages) while understanding them through more international languages such as English by leveraging state-of-the-art artificial intelligence (AI) technologies. This system utilizes smart glasses (e.g., Apple Vision Pro[2]) to present information within a mixed reality (MR) space (Figure 1), thereby addressing challenges that conventional translation technologies alone cannot cover, such as the understanding of technical terminology and the translation of visual materials. The technical components of the system primarily consist of three sub-systems: (1) the advancement of speech/image recognition and translation technologies capable of handling technical terminology; (2) the design of MR user interfaces using smart glasses; and (3) research on sensing the degree of comprehension on the international students' side. By wearing smart glasses, the instructor's speech and the lecture slides in front of the international students are converted into English (or whatever languages) and displayed in the MR space in real-time. This allows international students to seamlessly attend lectures without perceiving any language barrier, while sharing the physical classroom space with Japanese students. The slide translation method proposed in this paper is positioned as a core technology responsible for the translation of visual materials within this system.

4 Overview of the Proposed Framework

Figure 2 presents the three steps in the overall framework of our proposed system. We introduce two steps consisting of (i) translating text fragments shown in slides, and (ii) correcting layout errors by optimizing layout placement based on visual consistency (Figure 3). We further introduce an LLM-based third step of automatically selecting an optimal translated English slide among those eight candidates.

5 Slide Translation

In this section, we describe our proposed method for translating slide content and correcting layout errors caused by text expansion using LLM. Focusing on the fact that a PowerPoint file (.pptx) is essentially a zip archive composed of multiple XML files, we adopt an approach where the LLM directly parses and edits the XML data containing the slide's structural information. This approach allows for manipulation of items shown in each slide while perfectly preserving the underlying structure of text and vector objects.

In the slide translation step, rather than applying translation tools such as Google Lens[3] and DeepL[4], we employ an LLM (Gemini-3-Pro (Google, 2025)) to perform translation and preliminary layout edit (denoted as "first layout edit" hereafter) simultaneously. Regarding the validation of the decision above of covering the whole procedures of slide translation and layout edit by LLMs, Appendix A discusses comparison with another baseline of integrating text translation and separate layout edit by LLMs, where such a baseline underperforms the proposed approach presented above.

LLMs other than Gemini-3-Pro such as GPT-5.2[5], Claude Sonnet 4.5[6], and Qwen3-VL[7] mostly failed in generating translated slides, so that we concluded that Gemini-3-Pro as the LLM to be used in this study[8]. The failure analysis regard-

[2] https://www.apple.com/apple-vision-pro/

[3] https://lens.google/

[4] https://www.deepl.com/

[5] https://platform.openai.com/docs/models

[6] https://www.anthropic.com/claude/sonnet

[7] https://github.com/QwenLM/Qwen3-VL

[8] Also in the tasks of layout edit presented in section 6 and of optimal slide selection presented in section 7, Gemini-3-Pro

ing the exclusion of other LLMs is detailed in Appendix B. Specifically, we instruct the LLM to target only the Japanese text contained within the <a:t>...</a:t> tags in the XML, ensuring that the translation is executed without corrupting the XML tag structure. We established two configurations for the input modality as below and experimentally compared their performance:

1. **XML only:** An approach that takes only XML, which contains the slide's structural and text information, as input. Since it lacks visual information, the LLM must infer the layout solely from the information within the XML tags.

2. **XML + source slide image (JP):** An approach that inputs the image of the original Japanese slide alongside the XML. By leveraging visual information, this method aims to achieve translation and layout edits that better reflect the slide's context (e.g., the correspondence between figures and captions).

6 Layout Edit

In this section, we describe the specific design of layout edit in our proposed method. Section 6.1 describes the prompt tuning applied commonly to both the first and the second steps. Subsequently, section 6.2 describes the three input modalities and the few-shot prompting introduced in the second step.

6.1 Prompt Tuning of First Layout Edit

To ensure that the LLM executes appropriate layout edits while preserving the integrity of the XML constituting the PowerPoint file, we designed the prompt based on two perspectives of visual layout edit and XML structure preservation.

First, we present the constraints regarding specific layout modifications.

- **font resizing:** Given the tendency for text volume to increase in Japanese-to-English translation, we instructed the model to reduce the font size attribute (sz) if the translated text becomes too long.

- **text box decomposing and text wrapping:** We allowed structural changes to text boxes when overlapping items were predicted. Additionally, in slides where the automatic wrapping feature is disabled ("wrap=none"), a single line of translated text may become extremely long. To prevent this, we included a constraint to change the wrap attribute within the <a:bodyPr> tag to "square" (wrapping enabled).

- **layout preservation:** We instructed the model to leave sections unchanged where layout errors did not occur due to translation, and to rewrite the XML only for parts requiring edit. This reduced the risk of unintended alteration caused by LLM hallucinations.

To further ensure that the edited XML functions correctly as a PowerPoint file, we present the following constraints regarding structural integrity.

- **preservation of non-text items:** We instructed the model to maintain the status quo for items such as images (p:pic), connectors (p:cxnSp), and tables (p:graphicFrame), as modifying them carelessly can easily degrade visibility due to hallucinations. In particular, we strictly prohibited changing ID references such as the r:embed attribute, as this leads directly to file corruption.

- **escape character processing:** Since special characters such as & and < cause syntax errors in XML, we instructed the model to ensure that they are properly escaped.

- **XML format check and restoration:** Since the output from the LLM must be in a complete XML format, we prohibited errors such as unclosed tags. This is to prevent file corruption when the generated XML is recompressed and restored as a .pptx file.

Note that while the prompt tuning described above is applied in the first step, we do not employ the few-shot prompting to be described in the next section, i.e., executing the process in a zero-shot manner. This is because the first step performs not only the layout edit but also the translation task. At this step, where specific layout errors caused by translation are not yet determined, providing error examples via few-shot prompting raises concerns that the model might be unduly influenced by the example patterns or distracted from the translation

is employed as the LLM as with the task of slide translation, where other LLMs failed in generating layout error corrected slides or significantly underperformed Gemini-3-Pro.

task. Consequently, we designed the system to apply few-shot prompting exclusively in the second step, where the model can focus solely on layout edit based on the finalized English translated text contents. The prompt employed in this first step is shown in Appendix F.1.

6.2 Second Layout Edit

6.2.1 Three Input Modalities

In the second step, we adopted the following three input modalities for layout edit.

1. **XML only:** An approach that inputs only the XML containing the structural information of the translated slide, which may have layout issues. Since it lacks visual information, the LLM must infer layout interference solely from the numerical information within the XML tags.

2. **XML + translated slide image (EN):** An approach that inputs the image of the translated English slide in addition to the XML. Since the LLM can refer to the "actual appearance," this is expected to improve the accuracy of whitespace.

3. **XML + translated slide image (EN) + source slide image (JP):** In addition to the above, this approach also provides the original Japanese slide image. This allows the LLM to understand the original layout intent (such as the spatial relationship between figures and text) before performing edit.

6.2.2 Zero-shot and Few-shots

Regardless of the input modality, the LLM must execute appropriate XML edit based on the provided information. However, layout edit ranges from "minor adjustments" to "structural changes". Since it is difficult to appropriately address multiple types of layout errors with a single instruction, we adopt the few-shot prompting approach proposed by Brown et al. (2020) for the second layout edit. Specifically, we defined solutions and presented the model with pairs of XML and slide images before and after editing, as shown in Figure 4, to facilitate learning of specific edit intents. Note that in these few-shot examples, we limited the scope of the edit strictly to text items. We excluded layout manipulation of non-text items, such as charts and images, to avoid the risk of deviating from the original layout intent of the Japanese source slide. The specific edit operations are as follows:

(a) font resizing: Addresses text overflow and overlapping by reducing the font size attribute `sz`.

(b) text box reshaping: Addresses overlapping of adjacent items by modifying the dimensions (`cx`, `cy`) of the text box to induce text wrapping.

(c) text box decomposing: Addresses layout error caused by containing multiple sentences in a single text box by modifying the XML structure itself, such as splitting the text box to control line breaks.

Furthermore, regarding the configuration of the few-shot examples, we adopted and compared two distinct approaches based on how these edit operations are combined. We denote the set of operations applied to a single slide as (a), (b), and (c). The specific shot configurations are defined as follows:

An approach where all three operations are applied within a single slide.

- concatenated "(abc)" as 1-shot
- a sequence of two concatenated "(abc)" and "(abc)" as 2-shots
- a sequence of three concatenated "(abc)", "(abc)", and "(abc)" as 3-shots

An approach where each operation is applied to a separate slide individually.

- a sequence of three (a), (b), and (c) as 3-shots
- a sequence of six (a), (b), (c), (a), (b), and (c) as 6-shots
- a sequence of nine (a), (b), (c), (a), (b), (c), (a), (b), and (c) as 9-shots

By comparing these approaches, we clarify how the granularity of task presentation affects edit accuracy. The prompt employed in this second step is shown in Appendix F.2.

7 Optimal Slide Selection

As described in section 5 and section 6, the approaches to "translation" and "layout edit" involved multiple input conditions, resulting in multiple variations based on the combination of inputs. The

reason for establishing multiple input conditions is that the optimal edit method differs depending on the complexity and text volume of the input source language slide layout. A single method does not necessarily yield the best result for every slide. Therefore, in this section, we describe a method for selecting the optimal slide that most facilitates structure comprehension from multiple candidate slides generated by different combinations of inputs. The aforementioned approaches (2 translation methods $\times$ 4 edit methods) generate up to 8 types of candidate slides. A detailed discussion on the configurations of the input modalities is provided in Appendix D.

Here, by referencing LLM-as-a-Judge (Zheng et al., 2023), we propose a method that employs an LLM with high visual understanding capabilities as a model. Figure 2 illustrates the processing flow of 8 candidate slides outputted from the two steps into the LLM collectively and has it select the slide that is easiest for the learner to understand based on the following criteria:

1. **layout reproducibility:** Whether the original layout balance, figure structure, and use of whitespace are maintained.

2. **object reproducibility:** Whether there is any text overflow, missing items, or unnatural overlapping.

3. **overall visibility:** Whether the line breaks resulting from the English translation are appropriate and easy to read.

Based on these criteria, the LLM performs a relative evaluation among the candidate slides and automatically selects the optimal one. This enables the generation of slides with minimal layout errors, making them suitable as lecture materials. The prompt to select the optimal slide is provided in Appendix F.3.

8 Evaluation

In this section, we evaluate three perspectives to verify the quality of the slides finally generated by the proposed system. First, we compare the rate of occurrence of layout errors between the outputs of existing translation tools and our proposed method. Next, we conduct the evaluations of layout errors by both a manual oracle and an LLM on the outputs of existing tools and the proposed method to assess the overall slide quality. Finally, we measured the translation accuracy of the text obtained from the final output of the baseline and the proposed method.

8.1 Experimental Setup

As the evaluation dataset, we selected 40 slides taken from the classes taught by one (but not the first) of the authors' of this paper, covering introductory courses of machine learning and programming. Those 40 slides were selected under the criterion that they show complex objects such as mathematical formulas and charts for which layout errors easily observed during from Japanese to English translation. As baselines for comparison, we employed three methods: Google Lens[9], representing an image recognition-based translation approach; DeepL[10], representing a text-based translation approach; and the PowerPoint Translation Module of BOOM (Koneru et al., 2025)[11]. We compared the performance of our proposed method against these baselines. An annotator who is not among the authors of the paper conducted all the manual evaluation tasks.

8.2 Evaluation of Layout Errors

As an evaluation metric, this study defines the "layout error rate (%)", representing the ratio of items exhibiting layout errors to the total number of manually identified visual items per slide. The detailed formula for this metric is described in Appendix E. A lower layout error rate (%) indicates fewer layout errors. 2-fold cross validation was performed on the 40 slides described in Section 8.1. In the first fold, the 20 slides were first used for the tuning task to be described below, where the remaining another 20 slides were then used for evaluation. In the second fold, on the other hand, the tuning task was performed on the subset belonging to the second fold, while the evaluation was performed on the 20 slides belonging to the first fold. Finally, the evaluation result is obtained by averaging over the whole dataset of the overall 40 slides.

Following the tuning task performed through the 2-fold cross validation presented above, out of the zero-shot and few-shots presented in section 6.2.2, optimal configurations are selected for each of the 2-folds. Then, the overall evaluation results of layout error rates were obtained by averaging over the whole dataset of the overall 40 slides as shown in

[9] https://lens.google/

[10] https://www.deepl.com/

[11] https://github.com/saikoneru/image-translator

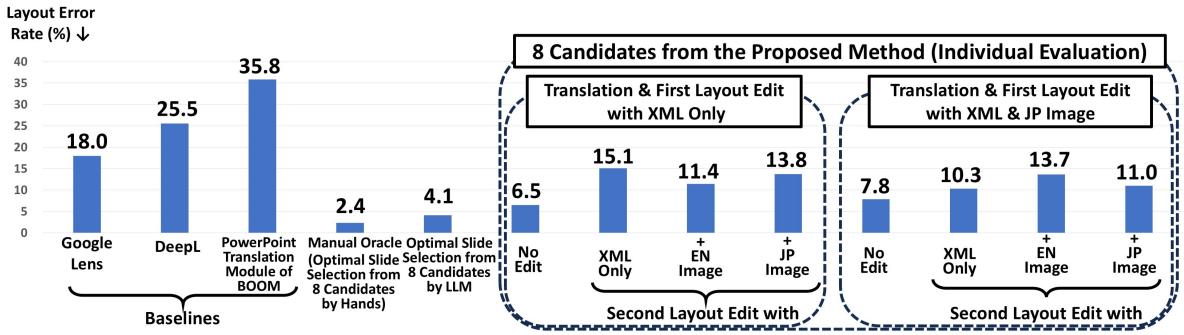

Figure 5: Layout Error Rate (%) of Baselines and Proposed Methods

evaluation by manual / LLM	baselines			proposed (optimal selection from 8 candidates)	total
	Google Lens	DeepL	PowerPoint Translation Module of BOOM		
manual oracle	10.0 (4 / 40)	7.5 (3 / 40)	0 (0 / 40)	**82.5 (33 / 40)**	100 (40 / 40)
LLM (Gemini-3-Pro)	2.5 (1 / 40)	2.5 (1 / 40)	0 (0 / 40)	**95.0 (38 / 40)**	100 (40 / 40)

Table 1: Results of Evaluating Model Selection against Baselines (Baselines: Google Lens, DeepL, PowerPoint Translation Module of BOOM, Proposed Method: optimal slide selection from 8 candidates, Metric: rates (%) of those selected as optimal out of 20 slides, Comparison: by manual oracle and by an LLM (Gemini-3-Pro))

Metric	baselines			proposed (optimal selection from 8 candidates)
	Google Lens	DeepL	PowerPoint Translation Module of BOOM	
BLEU	56.45	**66.86**	45.24	50.94
COMET	0.8335	**0.8513**	0.8297	0.8194

Table 2: Evaluation Results of Translation Accuracy

Figure 5[12]. The final output of the proposed system achieved an error rate of 4.1%, which outperformed baselines including Google Lens at 18.0%, DeepL at 25.5%, and the BOOM translation module at 35.8%, closely approaching the 2.36% of the manual oracle. Furthermore, the optimal selections at the third step by the LLM matched the manual selection by oracle in 80% of the overall 40 slides for evaluation. This optimal final output was selected by the LLM from the eight candidate outputs generated by the proposed method, where, as shown in the right half of Figure 5, their error rates are ranging from 6.5% to 15.1%. Among those individual candidates, the approach exclusively utilizing XML without the second layout edit achieved the lowest average of 6.5%, confirming that the first layout edit alone is sufficient without the second layout edit. However, as shown in the layout error rate of 4.1% by "the optimal slide selection from 8 candidates by LLM", out of the overall 40 slides for evaluation, there exist several slides for which the results of the second layout edit outperformed those without the second layout edit.

8.3 Evaluation of Optimal Model Selection

Finally, we evaluated the results of optimal model selection. Against the four candidates consisting of the output by the proposed method and those by the three baselines, the human annotator as the manual oracle as well as an LLM (Gemini-3-Pro) select an optimal one according to the criterion presented below. Here, the three baselines are Google Lens, DeepL, and the PowerPoint Translation Module of BOOM. Given the source Japanese slide, the criterion is based on the judgment regarding i) layout reproducibility of the source Japanese slide, ii) object reproducibility of the source Japanese slide, and iii) overall visibility of the target English slide itself. Layout reproducibility is measured by con-

[12]The first fold is optimized through "a sequence of three (a), (b), and (c) as 3-shots", while the second fold is as such through "a sequence of six (a), (b), (c), (a), (b), and (c) as 6-shots". This result reveals that concatenated few-shots relatively underperformed few-shots of (sequence of) separate individual layout edit operations within a single slide. We discuss this matter of comparison of few-shots of concatenated versus separate individual slides in Appendix C.

sidering how much the original layout dependency among objects within the source Japanese slide is reproduced. Object reproducibility, on the other hand, is measured by considering how much objects themselves within the source Japanese slide is reproduced.

Table 1 presents the percentage of cases where each method was selected as the optimal slide (defined as "model selection success rate"). In the evaluation by the manual oracle, the proposed method was deemed optimal in 82.5% of the cases. Conversely, the baselines of Google Lens (10.0%) and DeepL (7.5%) constituted the remaining selections and received low ratings due to issues such as inappropriate font sizes and overlaps between expanded text and figures. In the evaluation by an LLM, the proposed method achieved a significantly higher model selection success rate of 95.0%, indicating a tendency for the LLM to rate the proposed method even higher than the manual oracle. These results demonstrate that the proposed method is capable of generating translated slides of higher quality than existing methods, particularly for lecture slides with complex layouts.

8.4 Evaluation of Translation Accuracy

To measure translation accuracy, text was extracted from the Japanese slides and the respective output results. As the proposed method, DeepL and the PowerPoint Module of BOOM produce pptx files, texts enclosed within `<a:t>...</a:t>` tags in the XML were extracted. As Google Lens performs image translation, texts were extracted using character recognition with Gemini-3-Pro. The evaluation metrics used were BLEU and COMET. Furthermore, for the evaluation, two reference sentences were generated using GPT-5.2 and Claude Sonnet 4.5, and the average of the results was calculated. The results are shown in Table 2. In the BLEU results, the proposed method underperformed compared to DeepL and Google Lens. In particular, a clear difference was observed between DeepL (66.86) and the proposed method (50.94). This difference in scores is thought to be because, whereas DeepL produces output that is close to a literal translation of the text, the proposed method performs translation with the assumption that the layout will be adjusted. On the other hand, the COMET results showed no considerable differences between the various outputs. This suggests that the proposed method is capable of producing translations that retain their meaning even when layout adjustments are required.

9 Conclusion

We proposed a structure preserving slide translation framework utilizing direct XML manipulation that outperformed the baselines in both the layout error rate and the model selection success rate. Future work prioritizes processing acceleration for the real time integration detailed in Section 3.

10 Acknowledgments

This work was supported in part by JSPS KAKENHI Grant Number 25H00566 and 2025 Grant from Hoso Bunka Foundation.

Limitations

While our proposed method demonstrates high performance of layout error correction, there exist several limitations to be addressed in future work.

First, our evaluation is currently limited to a single translation direction from Japanese to English. However, the employed architecture of manipulating language independent XML structures by LLMs should be generalizable to any translation direction of any language pair. Conducting empirical validation across diverse language pairs remains for future work.

Second, regarding the evaluation dataset, we tested our method on 40 slides from the field of machine learning. To further prove our approach to be generalizable and robust, we plan to construct and evaluate on a larger dataset covering multiple academic domains that consist of over 100 slides.

Third, the prompts used to evaluate optimal slide selection and optimal model selection performed by LLMs are not fully optimized. Current prompts of the evaluation by LLMs may limit the full potential of LLMs to accurately assess complex visual inputs. Therefore, to conduct more accurate and robust evaluations, it is necessary to concentrate on prompt tuning of evaluating optimal slide selection and optimal model selection.

Fourth, the manual evaluation was performed by a single annotator. However, the lack of inter-annotator agreement undermines the reliability of the results. Since layout evaluation involves subjective judgments, it is difficult to standardize the criteria. Therefore, in this study, we focused on establishing a foundational framework for introducing new evaluation metrics. In the future, we plan

to conduct evaluations using multiple annotators and report on inter-annotator agreement.

Finally, our current system explicitly excludes the manipulation of non text items. Therefore, within inserted images that are not native PowerPoint objects, it does not support translation of text and layout adjustment of objects. Specifically, it is essential to incorporate any existing image translation technology that involves text extraction using optical character recognition (OCR), followed by text translation and the direct overlay of the translated text onto the original image.

References

Sambaran Bandyopadhyay, Himanshu Maheshwari, Anandhavelu Natarajan, and Apoorv Saxena. 2024. Enhancing presentation slide generation by LLMs with a multi-staged end-to-end approach. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 222–229, Tokyo, Japan. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 33rd NeurIPS*, pages 1877–1901. Curran Associates, Inc.

ECMA International. 2021. Standard ECMA-376: Office open XML file formats. https://ecma-international.org/publications-and-standards/standards/ecma-376/.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. Layoutgpt: compositional visual planning and generation with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Jiaxin Ge, Zora Zhiruo Wang, Xuhui Zhou, Yi-Hao Peng, Sanjay Subramanian, Qinyue Tan, Maarten Sap, Alane Suhr, Daniel Fried, Graham Neubig, and Trevor Darrell. 2025. AutoPresent: Designing structured visuals from scratch. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2902–2911.

Google. 2025. Gemini 3 Pro frontier safety framework report. Technical report, Google DeepMind.

Zheng Huang, Xukai Liu, Tianyu Hu, Kai Zhang, and Ye Liu. 2025. PPTBench: Towards holistic evaluation of large language models for powerpoint layout and design understanding. https://arxiv.org/abs/2512.02624. *Preprint*, arXiv:2512.02624.

Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. 2023. LayoutDM: Discrete diffusion model for controllable layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10167–10176.

Sai Koneru, Fabian Retkowski, Christian Huber, Lukas Hilgert, Seymanur Akti, Enes Yavuz Ugan, Alexander Waibel, and Jan Niehues. 2025. BOOM: Beyond only one modality KIT's multimodal multilingual lecture companion. https://arxiv.org/abs/2512.02817. *Preprint*, arXiv:2512.02817.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. TrOCR: transformer-based optical character recognition with pre-trained models. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26763–26773.

Ishani Mondal, Shwetha S, Anandhavelu Natarajan, Aparna Garimella, Sambaran Bandyopadhyay, and Jordan Boyd-Graber. 2024. Presentations by the humans and for the humans: Harnessing LLMs for generating persona-aware slides from documents. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2664–2684, St. Julian's, Malta. Association for Computational Linguistics.

Edward Sun, Yufang Hou, Dakuo Wang, Yunfeng Zhang, and Nancy X. R. Wang. 2021. D2S: Document-to-slide generation via query-based text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1418, Online. Association for Computational Linguistics.

Wenxin Tang, Jingyu Xiao, Wenxuan Jiang, Xi Xiao, Yuhang Wang, Xuxin Tang, Qing Li, Yuehe Ma, Junliang Liu, Shisong Tang, and Michael R. Lyu. 2025. SlideCoder: Layout-aware RAG-enhanced hierarchical slide generation from design. In *Proceedings of*

the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9015–9039, Suzhou, China. Association for Computational Linguistics.

Xiaojie Xu, Xinli Xu, Sirui Chen, Haoyu Chen, Fan Zhang, and Ying-Cong Chen. 2025. PreGenie: An agentic framework for high-quality visual presentation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3045–3063, Suzhou, China. Association for Computational Linguistics.

Hao Zheng, Xinyan Guan, Hao Kong, Wenkai Zhang, Jia Zheng, Weixiang Zhou, Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2025. PPTAgent: Generating and evaluating presentations beyond text-to-slides. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14402–14418, Suzhou, China. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the 37th NeurIPS*. Curran Associates Inc.

Wang Zhu, Alekh Agarwal, Mandar Joshi, Robin Jia, Jesse Thomason, and Kristina Toutanova. 2024. Efficient end-to-end visual document understanding with rationale distillation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8401–8424, Mexico City, Mexico. Association for Computational Linguistics.

A Integrating Text Translation and First Layout Edit by LLMs as a Baseline

To demonstrate the absolute necessity of integrating the first layout edit with text translation, this section examines the limitations of a decoupled approach that relies solely on text translation tools such as DeepL during the first step. Specifically, the method evaluated in this study translated only the text information extracted from XML, overwrote the data without performing any layout edit in the first step, and transitioned the output to the second step for subsequent layout edit.

Experimental results revealed that ignoring layout constraints during the initial translation phase caused severe layout errors, such as extreme text overlaps and structural breakdowns. Consequently, even after applying the subsequent second layout edit, it was difficult for the model to fully recover the original design intent, and the final layout quality was inferior to our proposed method. This finding demonstrates that integrating text translation and preliminary layout edit within the first step as in our proposed method is effective in preventing catastrophic layout degradation that cannot be repaired downstream. Note that the text translation and all layout edit tasks that make up these evaluated approaches are performed by an LLM (Gemini-3-Pro).

B Failure Analysis of Other LLMs

We discuss the issues encountered with GPT-5, Claude Sonnet 4.5, and Qwen3-VL, which were excluded from this study. The XML files that make up PowerPoint have a strictly defined structure; if discrepancies in tags or other issues occur, the files become impossible to parse. Therefore, LLMs must output complete XML without structural errors. However, the excluded LLMs frequently stopped outputting XML midway, resulting in numerous instances where the structure was corrupted. Furthermore, even when the XML structure was preserved, issues arose in which the LLM omitted descriptions of objects within the slides, resulting in the loss of many elements. For these reasons, this study adopted only Gemini-3-Pro.

C Discussion on Comparison of Few-shots of Concatenated versus Separate Individual Slides

Evaluation result of section 8.2 reveals that concatenated few-shots relatively underperformed few-shots of (sequence of) separate individual layout edit operations within a single slide. This indicates that, from the perspective of LLMs' performing layout edit operations, concatenated few-shots are less easy to recognize constituent layout edit operations than few-shots of (sequence of) separate individual layout edit operations within a single slide. Therefore, even if it is supposed that users of the proposed slide translation method collect diverse error cases and then intend to provide them as few-shots, it is clearly shown by the evaluation results of section 8.2 that adding those diverse error cases as few-shots simply damage the slide translation performance.

D Configurations of Input Modalities

The core operation of the proposed method is XML editing using an LLM, and images are used to assist with the editing process. In the first step, the available image input is a Japanese image. In the second step, both the translated English image and

the Japanese image become available. Therefore, depending on the combination of these available image data, a total of eight input modalities are formed. In the second step, we excluded inputs containing only Japanese images. The reason for this is that verifying current layout errors is essential for layout editing, and Japanese images serve merely as reference information for the intended layout.

E Definition of the Layout Error Rate

As an evaluation metric, this study defines the *layout error rate* (LER), formally computed as:

$$\text{LER} = \frac{1}{N} \sum_{i=1}^{N} \frac{e_i}{v_i} \times 100 \quad (\%) \qquad (1)$$

where N is the number of slides, v_i is the total number of manually identified visual items (text boxes, charts, formula objects, etc.) in the i-th slide, and e_i is the number of items exhibiting layout errors (e.g., overlaps between items, overflow beyond borders, or disappearance) in the i-th slide.

F Detailed Prompts for Slide Translation and Layout Edit

This appendix provides the specific system prompt used for the LLM in the first step of our proposed method. To avoid redundancy, the prompt is presented as a single template. The behavior conditionally changes based on the input modality (XML Only vs. XML + Source Slide Image), indicated by the *{If ...}* blocks. The prompt was originally constructed in Japanese to instruct the model accurately. We present both the translated English version and the original Japanese version.

F.1 Prompt for the First Step (Translation & First Layout Edit)

English Translation:

{If image is provided: } Cross reference the provided "original Japanese slide image" with the "XML data", translate the Japanese text in the XML into English, and adjust the layout.
{If image is not provided: } Translate the Japanese text within the provided XML into English, and adjust the layout.
Greetings, explanations, and partial omissions are strictly prohibited.

[Translation Policy]

- Translate only the Japanese text enclosed within `<a:t>...</a:t>` tags into English.

[Edit Policy]

- 1. Apply font size (`sz`) adjustments to simple overflows or element overlaps caused by translation.
- 2. Apply bounding box size (`cx`, `cy`) adjustments or relocate text boxes for overflows or overlaps caused by line break position differences between languages.
- 3. Apply structural changes to text boxes for element overlaps or complex layout corruptions that cannot be resolved by the above steps.
- 4. If the `wrap` attribute within the `<a:bodyPr>` tag in the XML is '`none`' and the text overlaps with figures or other text, strictly change it to '`square`' to enable text wrapping.
- 5. Leave sections unchanged where no layout issues are observed after translation, returning them as they are.

[Important Notes (Preservation of XML Structure)]

- The output must be in a complete XML format.
- Maintain the structure of images (`p:pic`), tables (`p:graphicFrame`), and connectors (`p:cxnSp`) without altering their contents.
- Strictly do not change ID attributes such as `r:embed` or `uri` attributes, as this will prevent images from being displayed.
- When using '`&`' in the text, you must write '`&amp;`'.
- Unclosed tags, missing attribute quotes, and forgetting to escape special characters (&, <) within the text are not allowed.
- Maintain extension tags such as `a16:` and `p14:` without deleting them.

Original Japanese Prompt:

*{*画像が入力される場合*: }* 提供された「元の日本語スライド画像」と「XMLデータ」を照合し、XML内の日本語テキストを英語に翻訳した上で、レイアウトを調整して出力してください。
*{*画像がない場合*: }* 提供されたXML内の日本語テキストを英語に翻訳し、レイアウトを調整して出力してください。
挨拶、解説、途中での省略は一切禁止です。

[翻訳の方針**]**

- `<a:t>...</a:t>` タグ内の日本語テキストのみを英語に翻訳してください。

[修正の方針**]**

- 1. 翻訳時に生じる単純なはみ出しや要素の重なりにはフォント(sz)調整を適用する。
- 2. 言語間の違いにより生じる改行位置によるはみ出しや要素の重なりには枠サイズ(cx, cy)の調整やテキストボックスの移動を適用する。

- 3. 上記で調整しきれなかった要素の重なりや複雑な配置崩れには、テキストボックスの構造変更を適用する。
- 4. xml中の<a:bodyPr>タグ内のwrap属性が'none'かつテキストが図や他のテキストに被っている場合、'square'に変更して折り返しを有効にする。
- 5. 翻訳した際にスライド内の問題が見られない箇所は手を付けずに、現状維持したまま返す。

[重要事項 (XML構造の維持)]

- 出力は完全なXML形式でなければなりません。
- 画像 (p:pic) や 表 (p:graphicFrame)、コネクタ (p:cxnSp) の構造は、内容を変更せずに維持してください。
- 特に r:embed などのID属性や、uri 属性は絶対に変更しないでください。画像が表示されなくなります。
- テキスト内で '&' を使う場合は必ず '&' と書いてください。
- タグの閉じ忘れ、属性のクォート漏れ、テキスト内の特殊文字（&, <）のエスケープ忘れは許されません。
- a16: や p14: などの拡張タグも削除せずに維持してください。

F.2 Prompt for the Second Step (Second Layout Edit)

For the second step, the LLM is instructed to focus exclusively on layout editing using few-shot examples. Similar to the first step, the prompt is presented as a single template. The behavior conditionally changes based on the three input modalities described in section 6.2 (XML Only, XML + EN Image, or XML + EN Image + JP Image).

English Translation:

{If XML Only: } Learn from the provided "edit examples" and optimize the layout of the target XML.
{If XML + EN Image: } Analyze the provided "edit examples" and the "target slide image", and optimize the layout.
{If XML + EN Image + JP Image: } Analyze the provided "edit examples", the "target slide image", and the "original Japanese slide image for reference", and optimize the layout.
Greetings, explanations, and partial omissions are strictly prohibited.

[Edit Policy]

- 1. Apply font size (sz) adjustments to simple overflows or element overlaps.
- 2. Apply bounding box size (cx, cy) adjustments or relocate text boxes for overflows or overlaps caused by line break positions.
- 3. Apply structural changes to text boxes for element overlaps or complex layout corruptions that cannot be resolved by the above steps.
- 4. If the wrap attribute within the <a:bodyPr> tag in the XML is 'none' and the text overlaps with figures or other text, strictly change it to 'square' to enable text wrapping.
- 5. Leave sections unchanged where no layout issues are observed, returning them as they are.

[Important Notes (Preservation of XML Structure)]

- The output must be in a complete XML format.
- Maintain the structure of images (p:pic), tables (p:graphicFrame), and connectors (p:cxnSp) without altering their contents.
- Strictly do not change ID attributes such as r:embed or uri attributes, as this will prevent images from being displayed.
- When using '&' in the text, you must write '&'.
- Unclosed tags, missing attribute quotes, and forgetting to escape special characters (&, <) within the text are not allowed.
- Maintain extension tags such as a16: and p14: without deleting them.

Original Japanese Prompt:

{XMLのみの場合: } 提供される「修正事例」を学習し、ターゲットXMLのレイアウトを最適化してください。
{XML + 英訳画像の場合: } 提供される「修正事例」および「ターゲットスライド画像」を分析し、レイアウトを最適化してください。
{XML + 英訳画像 + 日本語画像の場合: } 提供される「修正事例」、「ターゲットスライド画像」、そして「参照用オリジナル日本語スライド画像」を分析し、レイアウトを最適化してください。
挨拶、解説、途中での省略は一切禁止です。 **[修正の方針]**

- 1. 単純なはみ出しや要素の重なりにはフォント(sz)調整を適用する。
- 2. 改行位置によるはみ出しや要素の重なりには枠サイズ(cx, cy)の調整やテキストボックスの移動を適用する。
- 3. 上記で調整しきれなかった要素の重なりや複雑な配置崩れには、テキストボックスの構造変更を適用する。
- 4. xml中の<a:bodyPr>タグ内のwrap属性が'none'かつテキストが図や他のテキストに被っている場合、'square'に変更して折り返しを有効にする。
- 5. スライド内の問題が見られない箇所は手を付けずに、現状維持したまま返す。

[重要事項 (XML構造の維持)]

- 出力は完全なXML形式でなければなりません。

- 画像 (p:pic) や 表 (p:graphicFrame)、コネクタ (p:cxnSp) の構造は、内容を変更せずに維持してください。
- 特に r:embed などのID属性や、uri 属性は絶対に変更しないでください。画像が表示されなくなります。
- テキスト内で '&' を使う場合は必ず '&' と書いてください。
- タグの閉じ忘れ、属性のクォート漏れ、テキスト内の特殊文字（&, <）のエスケープ忘れは許されません。
- a16: や p14: などの拡張タグも削除せずに維持してください。

F.3 Prompt for the Third Step (Optimal Slide Selection)

In the third step, the LLM functions as an evaluator to select the optimal slide from the eight generated candidates. The model is provided with the original Japanese slide image as a reference, along with the eight candidates, and is instructed to rank them based on predefined visual criteria and output the result in JSON format.

English Translation:

Compare and evaluate the "original Japanese image" of the PowerPoint slide against the "multiple layout candidates" translated into English. Based on the following criteria, rank them in order of how well they capture the intent of the original slide.

- **1. layout reproducibility :** Whether the original layout balance, figure structure, and use of whitespace are maintained.
- **2. object reproducibility :** Whether there is any text overflow, missing items, or unnatural overlapping.
- **3. overall visibility :** Whether the line breaks resulting from the English translation are appropriate and easy to read.

[Output Format]
Output only the following JSON format:

```
{
"best_method": "The best method (folder name)",
"ranking": ["1st method", "2nd method", ...],
"scores": {"Method A": 95, "Method B": 80, ...},
"reasoning": "Brief reasoning"
}
```

Original Japanese Prompt:

PowerPointスライドの「日本語元画像(Original)」と、それを英訳した「複数のレイアウト案(Methods)」を比較評価してください。 以下の基準に基づき、元スライドの意図を最もよく汲んでいる順にランク付けしてください。

- **1.** レイアウト再現性 **:** 元の配置バランス、図形構造、余白の使い方が維持されているか。
- **2.** 情報の完全性 **:** テキストのはみ出し、欠損、不自然な重なりがないか。
- **3.** 視認性 **:** 英語化に伴う改行位置などが適切で読みやすいか。

[出力形式**]**
以下のJSONフォーマットのみを出力してください。

```
{
"best_method": "最も良かった手法(フォルダ名)",
"ranking": ["1位の手法", "2位の手法", ...],
"scores": {"手法名": 95, "手法名": 80, ...}, // 100点満点
"reasoning": "簡潔な理由"
}
```