Attend to what I say: Highlighting relevant content on slides

Megha Mariam K M[0009−0001−7188−9818] and
C. V. Jawahar[0000−0001−6767−7057]

CVIT, International Institute of Information Technology, Hyderabad, India
{megha.km@research.,jawahar@}iiit.ac.in

Abstract. Imagine sitting in a presentation, trying to follow the speaker while simultaneously scanning the slides for relevant information. While the entire slide is visible, identifying the relevant regions can be challenging. As you focus on one part of the slide, the speaker moves on to a new sentence, leaving you scrambling to catch up visually. This constant back-and-forth creates a disconnect between what's being said and the most important visual elements, making it hard to absorb key details—especially in fast-paced or content-heavy presentations, as in a conference talk. This will require an understanding of slides, especially text, graphics, and layout. We introduce a method that automatically identifies and highlights the most relevant slide regions based on the speaker's narrative. By analyzing spoken content and matching it with textual or graphical elements in the slides, our approach ensures better synchronization between what you hear and what you need to attend to. We explore different ways of solving this problem and assess their success and failure cases. Analyzing multimedia documents is emerging as a key requirement for seamless understanding of content-rich videos, such as educational videos and conference talks, by reducing cognitive strain and improving comprehension. Code and dataset available at: https://github.com/meghamariamkm2002/Slide_Highlight

Keywords: OCR · ASR · Multimodal Documents

1 Introduction

Attending a multimedia presentation often feels like a balancing act—listening to the speaker while simultaneously scanning the slides for relevant information. The entire slide is available, yet identifying the most critical regions can be challenging. As the speaker moves through different aspects of the content, audience members must quickly locate the corresponding visual elements to fully grasp the explanation. However, spoken language is fleeting—by the time they find the right section, the discussion has already moved forward. This misalignment forces attendees to split their attention between processing speech and searching

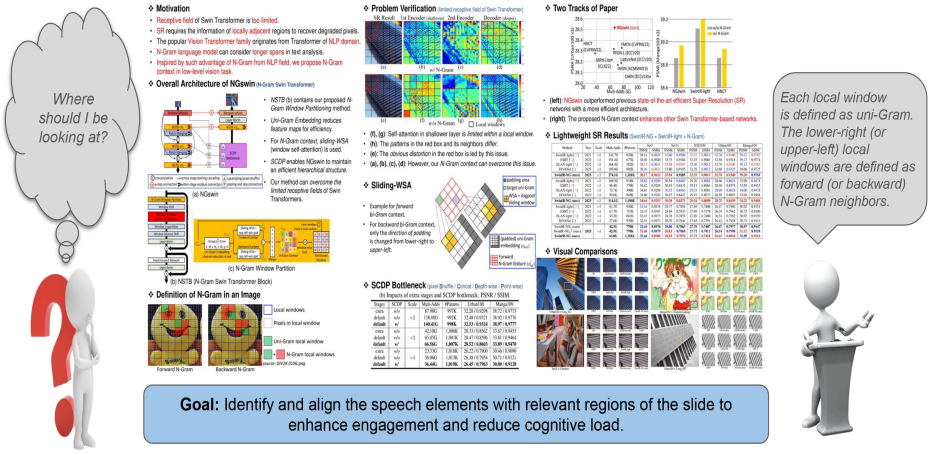


Fig. 1. This illustration highlights the challenge audiences face during presentations in simultaneously locating relevant content on slides while paying attention to the speaker. By synchronizing spoken information with corresponding visual elements, cognitive load is reduced, engagement is improved, and comprehension and retention of key insights are enhanced.

for visual cues, increasing cognitive effort and making it hard to absorb information effectively. Over time, this struggle can lead to information loss, reduced engagement, and a diminished overall learning experience. Figure 1 illustrates this challenge.

How do we address this problem? One simple strategy should be to highlight the relevant regions. However, this is not trivial as we need to work together with speech and slides. When speakers reference multiple slide elements at once or fail to visually emphasize key content in sync with their narration, the difficulty becomes even greater. This lack of coordination places an additional cognitive burden on the audience, requiring them to process auditory information while searching for relevant visuals—ultimately impacting comprehension and engagement. Research on cognitive load in multimedia learning suggests that an increased working memory demand can negatively affect learning outcomes [1]. Furthermore, individual differences in how people process multimedia content play a crucial role—when the presentation format does not align with their preferred learning style, the struggle intensifies [17].

Solutions to problems like this require comprehensive solutions for multimodal document understanding. At this stage, the modalities are not aligned at a fine-grained level. There have been many attempts in the past to align modalities such as movies and scripts, speech and transcripts, and books and texts [28,5]. In this case, we need to align spoken content with relevant slide elements to create a seamless and intuitive presentation experience. To validate the method, we introduce a dataset of conference presentations. The dataset contains video segments with speech annotated with corresponding speech tran-

scripts with timestamps, slide layouts, and the text presented on the slide. We explore multiple methods for aligning the speaker's narrative with both textual and visual components of the slides, ensuring that key information is effortlessly accessible. We hope that research on problems like this will help in a better understanding of multi-modal documents, leading to many potential applications in the real world.

2 Related Work

2.1 Human Document Interaction

We as humans increasingly interact with documents today; we have gone well beyond "reading" books and historical manuscripts. We "watch" videos, such as educational videos, and learn concepts rather than "reading" books and handwritten notes as in the traditional educational system [23]. Such visual documents are rich in content, speech, text, graphics, and annotations. From a document understanding point of view, one might want to pass the visual stream through the layout and textual content. Extensive research in OCR, with appropriate adaptation, has been increasingly used. Speech can also be converted to text with timestamps using Automatic Speech Recognition (ASR) tools [18,3]. In recent years, methods like Whisper have provided high speech-recognition capability in all languages. However, combining these two modalities is still challenging. Audio-visual processing of videos has been receiving attention in recent years [20,21,16,15].

Though the accuracies of speech recognition have been rapidly improving with better and better performance, technical presentations could pose newer challenges to commercial ASR systems, as the script could contain technical terms that are unique to a domain. A typical international conference today also sees diverse accents, including those that commercial systems are not well trained for. This is also true for OCR systems that can be used for parsing slides. Though OCRs and ASRs have been commercially available with good accuracies, there is a need to improve their accuracies with domain-specific techniques such as post-processing. In the past, researchers have used lexicons and LLMs for post-processing recognition outputs [13,12,14].

2.2 Understanding Slide Documents

Presentations are essential for education and information sharing, with slides complementing the talk. Research has focused on analyzing slides to improve their structure and accessibility. Competitions such as the Page Segmentation Competition [2] and ICDAR 2017 RDCL [7] have set benchmarks for layout analysis, while the ICDAR 2023 competition on structured text extraction [27] emphasized the challenge of extracting meaningful text from complex documents, particularly under zero-shot and few-shot conditions. A key aspect of slide analysis is layout understanding, which involves identifying distinct regions like text, images, and charts while maintaining semantic coherence. Deep learning-based

approaches have significantly advanced this field. Yang and Hsu (2020) [26] introduced SLLD, a CNN-based model that approaches segmentation as an object recognition task, outperforming traditional methods. Expanding on this, Yang and Hsu (2022) [25] developed TRDLU, a transformer-based model that classifies slide content more accurately using an encoder-decoder framework. Beyond digital slides, real-world applications have been explored. WiSe [10] enables segmentation from live slide photos, allowing structured information extraction even in dynamic settings such as conferences and classrooms. Information retrieval techniques further enhance slide accessibility. OCR-based slide retrieval [9] achieves accuracy levels comparable to traditional electronic document search. Additionally, video-based frameworks dynamically analyze slide content. The SlideVQA dataset [22], consisting of over 2,600 slide decks and 52,000 images, supports complex reasoning tasks. A document VQA model integrating evidence selection and question answering has also surpassed previous approaches, though it still lags behind human comprehension. With advancements in deep learning, multimodal processing, and enriched datasets, research continues to refine slide analysis techniques, bringing us closer to machine-level document understanding.

2.3 Alignment in Document Understanding

Aligning different modalities has been a fundamental tool in various document understanding tasks. One successful application is word-spotting, where text is matched with word images using Dynamic Time Warping (DTW) and Hidden Markov Models (HMM) [19]. Alignments have also been used to create annotated data by linking document images with paragraphs and pages of text. Such alignments enable scalable and cost-effective annotation of words, facilitating the rapid development of recognition algorithms with minimal human supervision.

It is very common today to align videos with transcripts to create temporal annotations, aiding in retrieval and weakly supervised captioning [5]. For example, books are often aligned with movies [28]. Fine-grained alignment can help in many challenging annotation tasks, such as phoneme recognition and sign language translation [5].

Matching is also carried out across languages and modalities. Matching is also tried with the help of semantic similarities. For example, a word in one language could be matched (or aligned) with i) a word written in a different font or by a different person, ii) a synonym of the word appearing in print or handwritten form, iii) an equivalent word appearing in a different language, iv) the word appearing in speech, v) the word or associated object appearing in a video stream.

3 Methodology

This work presents a framework for aligning spoken content with presentation slides to enhance audience engagement. By establishing a clear correspondence between speech and visuals, the approach promotes better comprehension and

reduces cognitive load during presentations. Below, we describe our method in detail.

3.1 Problem Definition

Given a presentation slide and its corresponding speech, the objective is to establish a correspondence between the spoken content and specific regions of the slide. This involves mapping the speech to relevant areas within the slide. As illustrated in Figure 2, the alignment module utilizes the outputs of OCR and ASR for alignment. Once the relevant slide regions are identified, the results are passed to the highlighting module, from which a specific visualization can be selected for presentation.

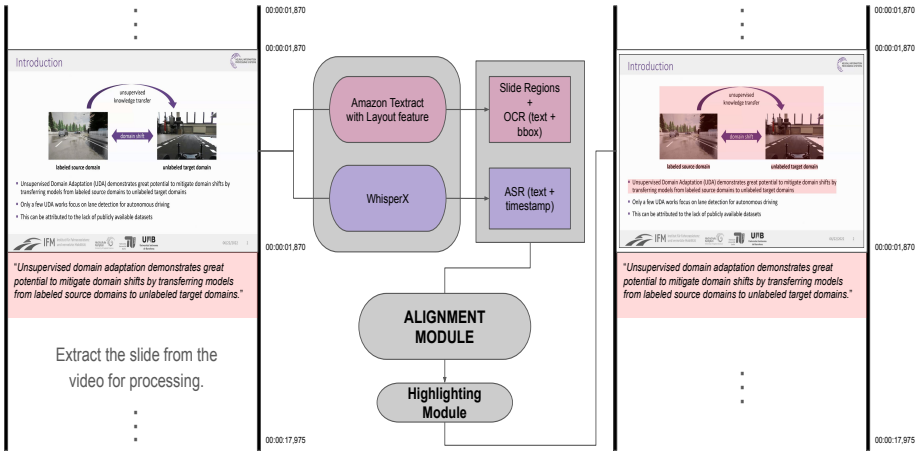


Fig. 2. This figure illustrates the pipeline of our method. The Alignment Module takes OCR-extracted text and ASR-generated transcripts as inputs to establish correspondences between spoken and visual content. These aligned results are then passed to the Highlighting Module, from which a suitable visualization can be chosen to highlight the relevant slide content.

3.2 Dataset

Data Collection: We curated a dataset comprising 14 presentation videos from NeurIPS 2022 and ICML 2023 conference presentations available on the SlidesLive platform. Each slide is paired with its corresponding audio segment and transcript. The dataset consists of 150 slides, encompassing diverse visual and textual elements such as figures, equations, tables, and text in various fonts and sizes. Figure 3 presents a selection of representative samples from our dataset, illustrating the diversity in both visual and textual content.

Statistics: The total duration of the dataset exceeds one hour. Figure 5 illustrates the distribution of presentation slides based on the duration of their corresponding audio segments, highlighting that most slides are associated with audio segments lasting between 10 and 20 seconds. The slide text corpus contains 7,466 unique alphabetical words, while the transcript text comprises 6,708 unique alphabetical words. Additionally, Figure 4 provides statistics on the audio segment durations, word count in OCR, and word count in ASR, including the minimum, maximum, average, and median values. This dataset serves as a valuable resource for evaluating the approach discussed below.

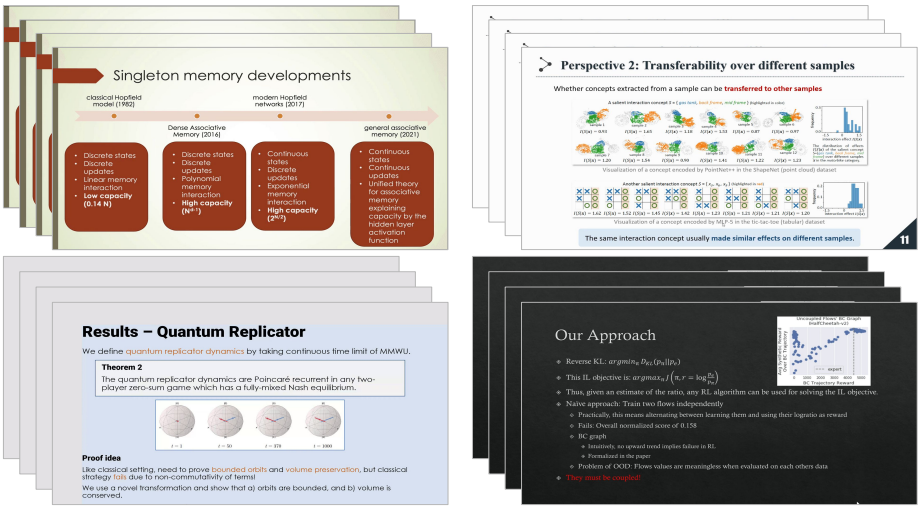


Fig. 3. This figure showcases a set of presentation slides from our dataset. The dataset consists of slides with varying layouts, color schemes, and content structures, reflecting the diversity of real-world academic and professional presentations.

	Min	Max	Avg	Median
Duration (in sec)	2	56	21.10	18
Word Count in OCR	1	799	99.93	83
Word Count in ASR	5	307	72.71	59

Fig. 4. Statistics for Audio Segment Durations, Word Count in ASR and OCR

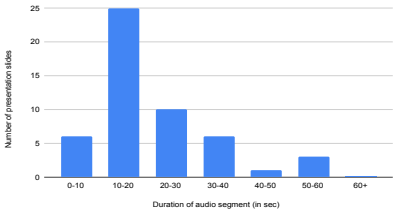


Fig. 5. The distribution of presentation slides based on the duration of their corresponding audio segments.

3.3 Recognizing text in slides and speech

Presentation slides often contain structured layouts with tables and dense text, making accurate segmentation critical for aligning visual and spoken content. We use Amazon Textract[1] for its OCR and layout analysis capabilities, enabling precise segmentation into meaningful regions. For speech, we employ WhisperX [3], which achieves low error rates (CER = 0.0129, WER = 0.0297; see Table 1), though it struggles with technical terms, speaker names, and accents (Figure 6). While Textract handles structured content well, open-source OCR tools like Tesseract often fail on multi-line text, complex fonts, or embedded figures (Figure 7), underscoring the need for advanced OCR like Textract.

	CER	WER	Edit Distance	Precision	Recall	F1-score
WhisperX	0.0129	0.0297	1.0327	0.9745	0.9718	0.9731
Corrected	0.0366	0.0832	3.3224	0.9287	0.9266	0.9271

Table 1. Comparison of ASR performance before and after post-correction using OCR. Metrics include Character Error Rate (CER), Word Error Rate (WER), Edit Distance, Precision, Recall, and F1-SCORE. WhisperX represents the original ASR output, while "Corrected" refers to the ASR output refined using OCR-based post-correction.

3.4 Alignment of speech and text

Our approach matches OCR-extracted text from slides with ASR-generated transcripts to identify relevant slide regions corresponding to spoken content, as illustrated in Fig. 8. There are two broader methods for matching: (1) String Matching: which identifies exact or near-exact textual correspondences, and (2) Semantic Matching: which leverages contextual similarity to establish meaningful associations between spoken content and visual elements. We will discuss these methods in more detail below.

String Matching One straightforward method for matching OCR and ASR outputs is fuzzy matching. This approach relies on identifying textual similarities, even in the presence of minor differences such as spelling errors, word order variations, or formatting inconsistencies between the spoken content and the slide text. While effective for simple matches, the accuracy of this method may decrease in more complex scenarios where the alignment is less direct.

Semantic Alignment

Semantic Embeddings: A more advanced approach is to use semantic similarity for aligning transcript lines to slide regions. This method compares the semantic embeddings of the transcript lines with the OCR-generated text of slide

[1] https://aws.amazon.com/textract/

GT: To bridge the gap, we released METS-CoV data set.
Original: To bridge the gap, we released MassCov data set.
Corrected: To bridge the gap, we released METS-CoV data set.

GT: Hey, everyone, I'm Peilin Zhou.
Original: Hey, everyone, I'm Pei Leng Zhou.
Corrected: Hey, everyone, I'm Peilin Zhou.

GT: Tulane incorporates balanced and domain randomized images from simulation as the source domain and the well-known TuSimple dataset as the target domain.
Original: Tulane incorporates balanced and domain randomized images from simulation as the source domain and the well-known to simple data set as the target domain.
Corrected: Tulane incorporates balanced and domain randomized images from simulation as the source domain and the well-known TuSimple dataset as the target domain.

GT: In total, there are 45 unique blocks leading to a search base size of more than 91,000 models.
Original: In total, there are 45 release blocks leading to a search base size of more than 91,000 models.
Corrected: In total, there are 45 unique blocks leading to a search base size of more than 91,125 models.

GT: This can be attributed to the lack of publicly available data
Original: This can be attributed to the lack of publicly available data
Corrected: This can be attributed to the lack of publicly available datasets.

Fig. 6. Comparison of ASR-generated text (GT) with the original and OCR-guided corrected versions.

regions. Unlike fuzzy matching, which relies on surface-level text similarity, semantic similarity captures the underlying meaning of the content, allowing for more robust alignment even in the presence of paraphrasing or significant lexical variation. To achieve this, we employ a variety of pre-trained models that generate semantic embeddings.
MINILM[24]: A lightweight, efficient model known for producing high-quality sentence embeddings that capture the semantic meaning of the text.
SCI-BERT [4]: A specialized BERT-based model fine-tuned on scientific literature, making it well-suited for technical and domain-specific content often found in academic and professional presentations.
SPECTER [8]: A model trained on scholarly documents to generate embeddings that are particularly effective in capturing the relationships and semantic structures within academic and scientific text.

LLM: LLMs such as Flan-T5 and Qwen-Instruct 2.5 are utilized for further enhancing the alignment process, in addition to above mentioned semantic embedding models.
Flan-T5 [6]: An instruction-tuned version of the T5 model (Text-to-Text Transfer

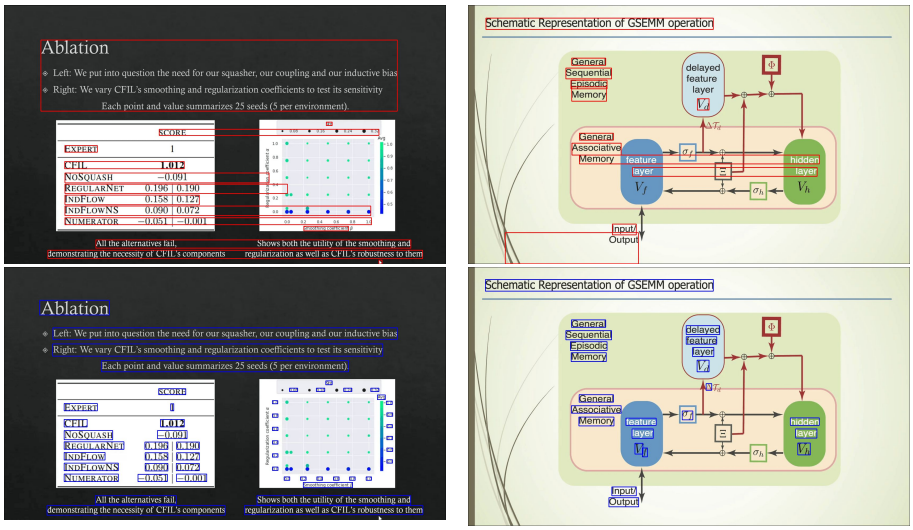


Fig. 7. Comparison of OCR results from Tesseract (red) and AWS Textract (blue) on slides with complex layouts.

Transformer), Flan-T5 excels at following specific instructions, making it highly suitable for post-processing tasks. We go through each ocr region and ask Flan-T5 whether the ocr region is relevant to the transcript line as a yes/no question for a given transcript line.
Qwen-Instruct 2.5 [11]: A powerful language model that excels in instruction-following tasks. In this approach, the model is presented with a transcript line along with the OCR outputs from all slide regions, and its task is to select the most relevant slide regions for the given transcript line.

For our dataset, the best thresholds were: Fuzzy (textual: 0.45), S-BERT (textual: 0.55), SPECTER (textual: 0.85), and Sci-BERT (textual: 0.75), each achieving the highest F1-score for its respective method. Since optimal thresholds may not be known beforehand, we used three general-purpose threshold settings—T-1 (textual: 0.8, visual: 0.6), T-2 (textual: 0.7, visual: 0.6), and T-3 (textual: 0.6, visual: 0.6)

3.5 Highlighting Regions

With the correspondence between the speech and slide regions established, the next step is to visualize the relevant slide regions in alignment with the spoken narrative. Several approaches can be used, such as drawing a box around the region, shading it with a distinct color, magnifying or enlarging the area of interest, or completely removing non-essential slide content. If the goal is to allow users to follow the speech while focusing on the highlighted regions but

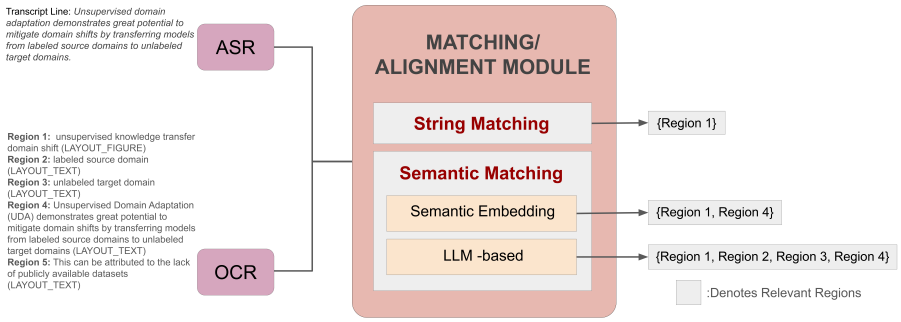


Fig. 8. The figure illustrates the different categories in the matching module and how they function for an given example.

still explore other parts of the slide, then hiding the background content may not be a suitable choice. However, if the slide contains multiple distinct sections covering different topics that are not related to each other, removing irrelevant regions may be a more effective approach to maintaining clarity and focus.

3.6 Using OCR to improve ASR accuracy

ASR in presentations often misinterprets technical or domain-specific terms, affecting alignment accuracy. To address this, we employ a prompt-based LLM approach that leverages slide context to refine ASR output. We use Qwen/Qwen2.5-1.5B-Instruct2 [**?**] with both transcription and slide text as input to correct misspellings and improve recognition of key terms. As shown in Fig. 6, this enhances the accuracy of domain-specific content identification.

4 Evaluation

4.1 Metric

Correctness Score: The Correctness Score (S_c) evaluates how accurately the identified slide regions match the corresponding transcript content. A high score indicates that most detected regions align correctly with the transcript, while a low score suggests errors, such as incorrect regions being identified or excessive regions being detected for a given transcript line. If no predictions are made for a transcript line, the score is automatically set to 1. The Incorrectness Score, defined as $1 - S_c$, represents the proportion of incorrect alignments.

Missing Score: The Missing Score (S_m) measures how many relevant slide regions were overlooked during alignment. A high missing score suggests that relevant regions were not detected, leading to incomplete alignments. It is the ratio of missing alignments in the prediction to the total number of expected alignments for a transcript line. If no alignments are needed for a transcript line, the score is set to 0.

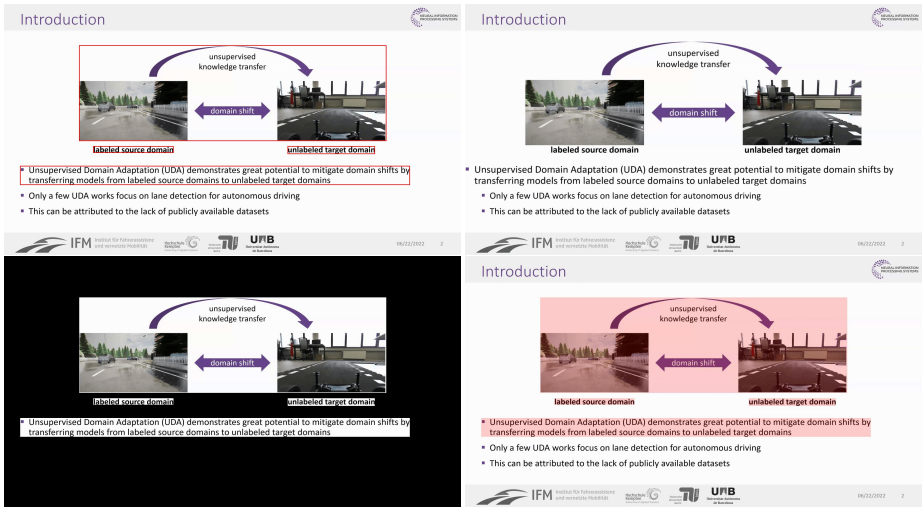


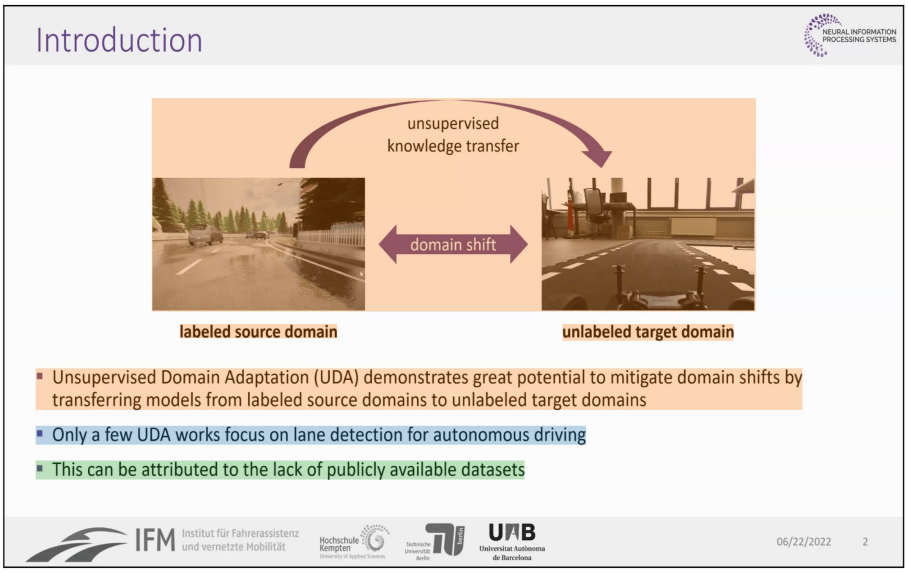
Fig. 9. Illustration of different techniques for highlighting relevant regions in presentation slides, including transparent bounding boxes, color overlays, content removal, and magnification of key areas.

F1-Score: To compute the F1-Score, both *precision* and *recall* must be defined. Recall is given by $1-S_m$. Precision is defined as the proportion of correctly predicted regions among all predicted regions. It is important to note that precision is **not** equivalent to the Correctness Score. If no predictions are made for a given transcript line, the precision is defined as zero.

4.2 Results

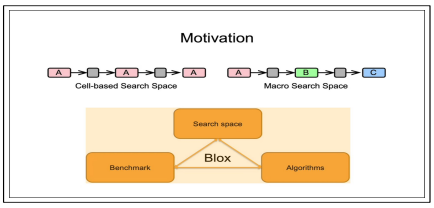
OCR and ASR Matching Results

Qualitative Results: Semantic alignment demonstrates good performance in mapping spoken content to visual slide regions. As shown in Figure 10, T5 effectively links transcript lines to their corresponding areas on the slide. To better understand the behavior of alignment models, Figure 11 presents an analysis of Flan-T5 across four representative cases: (1) perfect alignment with $S_c = 1$, $S_m = 0$; (2) perfect correctness with additional mismatches ($S_c = 1$, $S_m = 0.6$); (3) perfect missing score with partial correctness ($S_c = 0.66$, $S_m = 0$); and (4) a failure case where predictions are made despite no true match ($S_c = 1$, $S_m = 0$). Interestingly, S-BERT T-3 achieves the best F1 score overall. This performance arises from the fact that while LLMs like T5 and Qwen are context-aware and can handle complex semantic matches, they sometimes misinterpret generic phrases, which leads to incorrect region predictions. For example, Qwen associates phrases like “That’s all” or “Thank you for your attention” with conclusion sections, while T5 aligns them with the “Summary” label, as illustrated



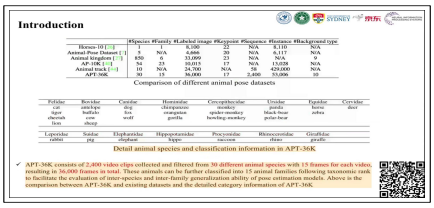
"Unsupervised domain adaptation demonstrates great potential to mitigate domain shifts by transferring models from labeled source domains to unlabeled target domains. However, only a few user works focus on lane detection for autonomous driving. This can be attributed to the lack of publicly available data"

Fig. 10. The figure illustrates the method's output, with each transcript line and its corresponding slide region highlighted in the same color.



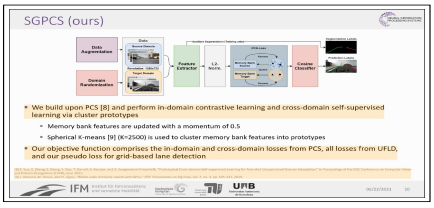
Transcript Line: To provide a systematic study of the performance of North's algorithm on a macro surface, we release blocks for a first large-scale benchmark and study of a macro surface.

$S_c = 1, S_m = 0$



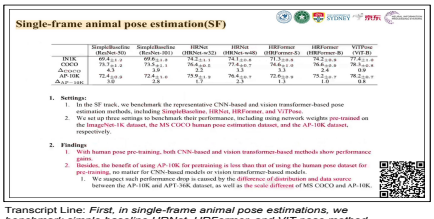
Transcript Line: *Above is a comparison between existing datasets and detailed category information of it.*

$S_c = 1, S_m = 0.6$



Transcript Line: *Similar to PCS, we perform in-domain contrastive learning and cross-domain self-supervised learning between the features of the feature extractor and the prototypes.*

$S_c = 0.66, S_m = 0$



Transcript Line: *First, in single-frame animal pose estimations, we benchmark simple baseline HRNet, HRFormer, and VIT pose method.*

$S_c = 0, S_m = 0$

: Denotes the relevant regions

Fig. 11. t5-based alignment results across three examples: the first shows perfect correctness and missing scores ($S_c = 1, S_m = 0$); the second has perfect correctness ($S_c = 1$); and the third has perfect missing score ($S_m = 0$).

in Figure 12. On the other hand, simpler baselines like Fuzzy rely on near-exact string matches and tend to make fewer predictions. This conservative behavior results in inflated correctness scores (e.g., $S_c = 1$) when no match is predicted, even if relevant matches are missed. The missing score component helps account for such omissions, providing a more balanced evaluation of alignment quality.

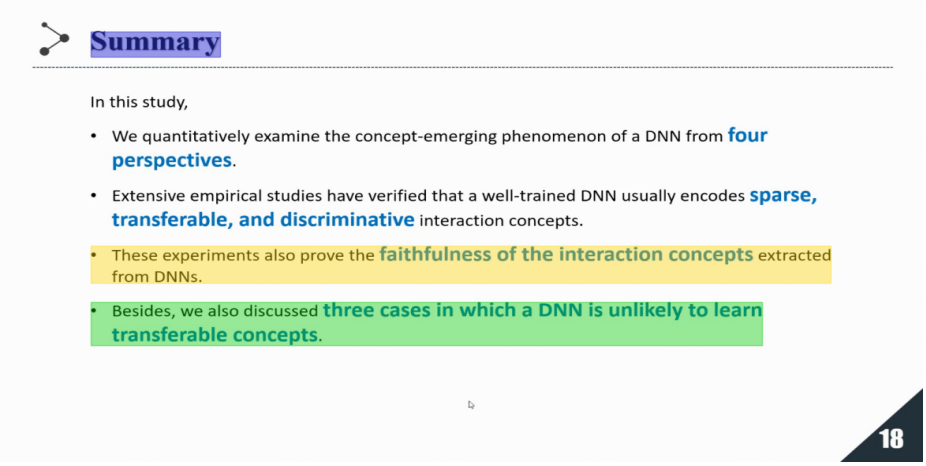


Fig. 12. The figure shows that for "That's all," T5 highlights blue and Qwen highlights yellow; for "Thank you for your attention," T5 highlights blue and Qwen highlights green.

Quantitative Results The effectiveness of the proposed alignment method is evaluated using the Correctness Score (S_c) and Missing Score (S_m), as presented in Table 2. Fuzzy matching consistently achieves the highest correctness ($S_c \approx 0.99$); however, it suffers from high missing scores, particularly at T-1 ($S_m = 0.663$), due to its inability to detect more difficult matches. In contrast, embedding-based methods such as Sci-BERT and SPECTER perform poorly overall, with correctness scores decreasing sharply across thresholds. Among the evaluated methods, MiniLM T-1 yields the best F1 score, followed closely by T5. LLM-based models offer a better balance between correctness and missing: Flan-T5 achieves $S_c = 0.687$ and $S_m = 0.272$, while Qwen-2.5 shows higher correctness ($S_c = 0.843$) at the cost of a higher missing score ($S_m = 0.490$). When comparing alignments on original (ASR) versus corrected transcripts, LLMs like Qwen benefit in correctness (e.g., a gain of $+0.084$ in S_c), and most methods exhibit improved missing scores. Overall, while fuzzy matching excels at exact match scenarios due to its high correctness, LLMs demonstrate more robust semantic alignment capabilities, particularly when provided with corrected input.

Method	**Avg** S_c	**Avg** S_m	**Avg** F_1	**Avg** S_c (W)	**Avg** S_m (W)	**Avg** F_1 (W)
Fuzzy [T-1]	0.998	0.663	0.254	1.000	0.673	0.246
Fuzzy [T-2]	0.998	0.603	0.307	1.000	0.617	0.295
Fuzzy [T-3]	0.989	0.525	0.364	0.989	0.526	0.360
S-BERT [T-1]	0.983	0.578	0.335	0.993	0.624	0.299
S-BERT [T-2]	0.943	0.460	0.403	0.966	0.507	0.373
S-BERT [T-3]	0.868	0.372	0.429	0.902	0.391	0.430
Sci-BERT [T-1]	0.524	0.356	0.280	0.636	0.474	0.279
Sci-BERT [T-2]	0.313	0.169	0.254	0.307	0.169	0.252
Sci-BERT [T-3]	0.222	0.056	0.003	0.219	0.053	0.220
SPECTER [T-1]	0.318	0.139	0.277	0.322	0.159	0.276
SPECTER [T-2]	0.211	0.035	0.219	0.211	0.0345	0.220
SPECTER [T-3]	0.174	0.010	0.191	0.175	0.011	0.191
T5	0.687	0.272	0.420	0.680	0.314	0.400
Qwen2.5	0.843	0.490	0.342	0.845	0.499	0.323

Table 2. Comparison of the average correctness score S_c and the average missing score S_m for different alignment methods under varying threshold settings. W denotes results obtained using the original transcript before correction. **T-1**, **T-2**, and **T-3** represent different threshold values for textual and visual regions: **T-1:** Textual region threshold = 0.8, Visual region threshold = 0.6, **T-2:** Textual region threshold = 0.7, Visual region threshold = 0.6 **T-3:** Textual region threshold = 0.6, Visual region threshold = 0.6

Results of ASR Post-Correction with OCR Guidance: Table 1 presents a comparison of the ASR performance before and after correction. The evaluation was conducted on a subset of the dataset, consisting of 96 samples. Notably, the corrected ASR exhibits an increase in Character Error Rate (CER) and Word Error Rate (WER) compared to the original version. This increase is attributed to the OCR-guided correction process, which enforces alignment with the extracted text from slide images. A closer examination of Fig. 6, particularly the last two examples, illustrates the nature of these corrections. In the third example, the corrected ASR transcribes the number as 91,125 instead of 91,000. This discrepancy arises because the speaker provided an approximate figure during the presentation, whereas the OCR correction enforces the exact value present in the slide. Similarly, in the fourth example, the word "data" is replaced with "datasets" due to OCR guidance, aligning the ASR output with the textual content extracted from the slides. Despite such variations, these corrections are beneficial for our task. The adjustments ensure better alignment with the visual content of the slides, aiding in the accurate localization of relevant information. Furthermore, as seen in the first three examples of Fig. 6, OCR-guided correction enhances the accuracy of technical terms and key concepts, thereby improving the overall quality of the transcription in a technical presentation setting.

Category	A	B	C
Magnifying/emphasizing	14.70	61.76	22.05
Bounding box	83.82	8.82	4.41
Shading	57.35	10.29	32.35
Hiding the background	38.23	14.07	47.05
Total	49.26	24.26	26.47

Fig. 13. User preference rating across different highlighting methods (all values in percentage).

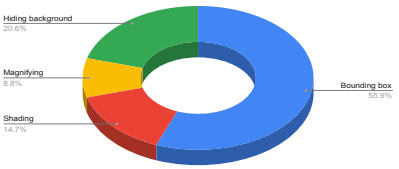


Fig. 14. Distribution of user preferences across four different highlighting methods.

5 User Study

We conducted a user study with 33 participants (Graduate, MS, PhD, and those pursuing MS/PhD) to evaluate four highlighting techniques. Each participant viewed six video clips (avg. 35.03s), sampled from the middle of presentations to simulate joining midway. After each clip, participants chose one of the following: A) The highlighting was useful — I preferred having the highlighting. B) I didn't mind whether the highlighting was present or not. C) I did not prefer the highlighting — I would rather watch the video without it. At the end, they indicated their most preferred technique and explained why. The results (Figure 14) showed bounding boxes were most preferred. Users found them effective for emphasizing key content while preserving full-slide visibility. Shading drew attention, but some noted it reduced readability by dimming the background. Magnification was least favored—while helpful for small text, it often obscured context and made focus areas unclear. Hiding was generally disliked, as users preferred being able to glance at surrounding content, though some appreciated its ability to reduce distractions. As summarized in Table 13, about 50% of participants preferred highlighting, especially when joining midway. Highlighting helped users catch up and stay focused, though its effectiveness depended on balancing emphasis with slide visibility.

6 Conclusion

We explore aligning speech with slide content using OCR and ASR matching, focusing on identifying where to direct attention during a presentation. We evaluated performance using the Correctness Score (S_c), Missing Score (S_m), and F1 score. We further explored string matching, semantic embeddings, and LLM-based approaches for deeper alignment. Future directions include leveraging VLMs for spatial and non-verbal references, and enabling real-time alignment in live scenarios.

7 Acknowledgments.

This work is supported by the MeitY Government of India, through the NLTM Bhashini https://bhashini.gov.in/ project.

References

1. Øistein Anmarkrud, A.A., Bråten, I.: Cognitive load and working memory in multimedia learning: Conceptual and measurement issues. Educational Psychologist **54**(2), 61–83 (2019). https://doi.org/10.1080/00461520.2018.1554484, https://doi.org/10.1080/00461520.2018.1554484 2
2. Antonacopoulos, A., Gatos, B., Bridson, D.: Page segmentation competition. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). vol. 2, pp. 1279–1283 (2007). https://doi.org/10.1109/ICDAR.2007.4377121 3
3. Bain, M., Huh, J., Han, T., Zisserman, A.: Whisperx: Time-accurate speech transcription of long-form audio (2023), https://arxiv.org/abs/2303.00747 3, 7
4. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text (2019), https://arxiv.org/abs/1903.10676 8
5. Bull, H., Afouras, T., Varol, G., Albanie, S., Momeni, L., Zisserman, A.: Aligning subtitles in sign language videos. CoRR **abs/2105.02877** (2021), https://arxiv.org/abs/2105.02877 2, 4
6. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling instruction-finetuned language models (2022), https://arxiv.org/abs/2210.11416 8
7. Clausner, C., Antonacopoulos, A., Pletschacher, S.: Icdar2017 competition on recognition of documents with complex layouts - rdcl2017. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 1404–1410 (2017). https://doi.org/10.1109/ICDAR.2017.229 3
8. Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D.S.: SPECTER: Document-level Representation Learning using Citation-informed Transformers. In: ACL (2020) 8
9. Daddaoua, N., Odobez, J., Vinciarelli, A.: Ocr based slide retrieval. In: Eighth International Conference on Document Analysis and Recognition (ICDAR'05). pp. 945–949 Vol. 2 (2005). https://doi.org/10.1109/ICDAR.2005.169 4
10. Haurilet, M., Roitberg, A., Martinez, M., Stiefelhagen, R.: Wise — slide segmentation in the wild. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 343–348 (2019). https://doi.org/10.1109/ICDAR.2019.00062 4
11. Hui, B., Yang, J., Cui, Z., Yang, J., Liu, D., Zhang, L., Liu, T., Zhang, J., Yu, B., Lu, K., Dang, K., Fan, Y., Zhang, Y., Yang, A., Men, R., Huang, F., Zheng, B., Miao, Y., Quan, S., Feng, Y., Ren, X., Ren, X., Zhou, J., Lin, J.: Qwen2.5-coder technical report (2024), https://arxiv.org/abs/2409.12186 9
12. Kumari, L., Singh, S., Rathore, V.V.S., Sharma, A.: A comprehensive handwritten paragraph text recognition system: Lexiconnet (2023), https://arxiv.org/abs/2205.11018 3

13. Kumari, L., Singh, S., Rathore, V., Sharma, A.: Lexicon and attention based handwritten text recognition system (2022), https://arxiv.org/abs/2209.04817 3
14. Lakomkin, E., Wu, C., Fathullah, Y., Kalinli, O., Seltzer, M.L., Fuegen, C.: End-to-end speech recognition contextualization with large language models (2023), https://arxiv.org/abs/2309.10917 3
15. Min, X., Zhai, G., Zhou, J., Zhang, X.P., Yang, X., Guan, X.: A multimodal saliency model for videos with high audio-visual correspondence. IEEE Transactions on Image Processing **29**, 3805–3819 (2020). https://doi.org/10.1109/TIP.2020.2966082 3
16. Naphade, M., Huang, T.: Extracting semantics from audio-visual content: the final frontier in multimedia retrieval. IEEE Transactions on Neural Networks **13**(4), 793–810 (2002). https://doi.org/10.1109/TNN.2002.1021881 3
17. Plass, J.L., Homer, B.D.: Cognitive load in multimedia learning: The role of learner preferences and abilities. In: Proceedings of the International Conference on Computers in Education. p. 564. ICCE '02, IEEE Computer Society, USA (2002) 2
18. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision (2022), https://arxiv.org/abs/2212.04356 3
19. Rath, T., Manmatha, R.: Word image matching using dynamic time warping. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. vol. 2, pp. II–II (2003). https://doi.org/10.1109/CVPR.2003.1211511 4
20. Shahabaz, A., Sarkar, S.: Increasing importance of joint analysis of audio and video in computer vision: A survey. IEEE Access (2024) 3
21. Singh, D., Gupta, A., Jawahar, C., Tapaswi, M.: Unsupervised audio-visual lecture segmentation. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5221–5230. IEEE (2023) 3
22. Tanaka, R., Nishida, K., Nishida, K., Hasegawa, T., Saito, I., Saito, K.: Slidevqa: A dataset for document visual question answering on multiple images (2023), https://arxiv.org/abs/2301.04883 4
23. Tani, M., Manuguerra, M., Khan, S.: Can videos affect learning outcomes? evidence from an actual learning environment. Educational technology research and development **70** (08 2022). https://doi.org/10.1007/s11423-022-10147-3 3
24. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers (2020), https://arxiv.org/abs/2002.10957 8
25. Yang, H., Hsu, W.: Transformer-based approach for document layout understanding. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 4043–4047 (10 2022). https://doi.org/10.1109/ICIP46576.2022.9897491 4
26. Yang, H., Hsu, W.H.: Vision-based layout detection from scientific literature using recurrent convolutional neural networks. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 6455–6462 (2021). https://doi.org/10.1109/ICPR48806.2021.9412557 4
27. Yu, W., Zhang, C., Cao, H., Hua, W., Li, B., Chen, H., Liu, M., Chen, M., Kuang, J., Cheng, M., Du, Y., Feng, S., Hu, X., Lyu, P., Yao, K., Yu, Y., Liu, Y., Che, W., Ding, E., Liu, C.L., Luo, J., Yan, S., Zhang, M., Karatzas, D., Sun, X., Wang, J., Bai, X.: Icdar 2023 competition on structured text extraction from visually-rich document images (2023), https://arxiv.org/abs/2306.03287 3
28. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books (2015), https://arxiv.org/abs/1506.06724 2, 4