Generate, Align and Predict (GAP): Detecting Neurocognitive Disorders via Cross-modal Consistency in Narratives

Junan Li*
Jinchao Li*
jli@se.cuhk.edu.hk
jcli@se.cuhk.edu.hk
The Chinese University of Hong Kong
Hong Kong SAR, China

Ka Ho Wong
khwong@se.cuhk.edu.hk
The Chinese University of Hong Kong
Hong Kong SAR, China

Xixin Wu
wuxx@se.cuhk.edu.hk
The Chinese University of Hong Kong
Hong Kong SAR, China

Helen Meng
hmmeng@se.cuhk.edu.hk
The Chinese University of Hong Kong
Hong Kong SAR, China

Abstract

The early detection of neurocognitive disorders (NCDs), such as Alzheimer's disease, remains a critical global health challenge due to the limitations of conventional diagnostic tools with high costs and limited accessibility. Visual-stimulated narrative (VSN)-based approach offers a promising alternative by capturing narrative patterns related to holistic cognitive domains. While prior work has focused on unimodal microstructural features (e.g., pauses, lexical diversity), macrostructural impairments—such as disrupted coherence and cross-modal inconsistency between narratives and visual stimuli—remain understudied despite their clinical importance. To address this, we propose GAP (Generate, Align, and Predict), a novel multimodal framework leveraging advances in Multimodal Large Language Models (MLLMs), Dynamic Programming (DP), and Vision-Language Model (VLM) to evaluate dynamic semantic consistency in VSNs. Evaluated on the Cantonese CU-MARVEL-RABBIT dataset, GAP achieved state-of-the-art performance (F1=0.65, AUC=0.75), outperforming traditional acoustic, linguistic, and pattern-matching baselines. In addition, we conduct an in-depth analysis that reveals key factors that provide insights into cognitive assessment using VSNs.

CCS Concepts

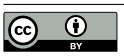
• **Information systems** → *Multimedia and multimodal retrieval*; • **Applied computing** → **Health informatics**.

Keywords

Neurocognitive disorder detection, Alzheimer's Disease, visual-stimulated narrative, text-image processing, multi-modality.

ACM Reference Format:
Junan Li, Jinchao Li, Ka Ho Wong, Xixin Wu, and Helen Meng. 2025. Generate, Align and Predict (GAP): Detecting Neurocognitive Disorders via Cross-modal Consistency in Narratives. In *Proceedings of the 2nd International Workshop on Multimedia Computing for Health and Medicine (MCHM '25), October 27–28, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3728424.3760767

*These authors contributed equally to this research.



MCHM '25, Dublin, Ireland

ACM ISBN 979-8-4007-1836-6/2025/10
https://doi.org/10.1145/3728424.3760767

1 Introduction

Neurocognitive disorders (NCDs) [1, 2], particularly Alzheimer's disease (AD), represent a growing global health challenge demanding sensitive and objective detection methods [3]. The NCD detection task is critical to enable timely interventions that may slow disease progression, yet conventional tools – such as neuroimaging [4, 5], cerebrospinal fluid (CSF) analysis [6] and neuropsychological assessments [7] – are often constrained by high resource demands and limited accessibility [8, 9]. These gaps have spurred interest in alternative paradigms that balance diagnostic precision with practicality through effective and automatic analysis. Among these, visual-stimulated narrative (VSN)-based assessment, where participants describe isolated images or tell a story based on sequential images, emerges as a promising solution.

VSN serves as a window of holistic cognitive domains through both its production and reflective nature. On the one hand, narrative generation represents an integrative cognitive process requiring coordinated engagement of attentional control, visual-spatial comprehension, linguistic fluency, executive functioning, memory retrieval, and theory-of-mind capabilities [10, 11, 12, 13, 14, 15]. On the other hand, the resulting narrative outputs contain observable biomarkers that mirror underlying cognitive states, such as dysfluencies [16, 17], reduced lexical diversity [18, 19], incomplete or irrelevant information [20], and inconsistencies with visual stimuli [21, 22, 23]. These quantifiable features form a multidimensional signature that can reveal subtle disruptions of neurocognitive systems.

Despite the promise of VSN-based NCD detection, translating narrative outputs into reliable diagnostic features remains a significant challenge. Existing research primarily focus on exploring unimodal patterns linked to the language production abilities, such as pause-related statistics, part-of-speech ratios, and acoustic or textual embeddings [24, 25, 26, 27, 28]. While these microstructural analyses provide valuable insights into specific language impairments that are related to bottom-up production of linguistic units,

they systematically overlook higher-order macrostructural features that represent top-down integration of basic units. This discrepancy is particularly crucial when considering macrostructural features reflect higher-order cognitive abilities like attention, memory, and executive function. Clinical research indicates that individuals with NCDs exhibit more pronounced impairments in macrostructuring than microstructuring, manifesting as disrupted topic maintenance, incoherent sequencing, and inconsistency with visual stimuli – features particularly relevant to temporal and cross-modal consistency between visual scenes and verbal descriptions [29, 30].

However, these macrostructural features, especially cross-modal consistency, are difficult to model because it demands aligning abstract semantic concepts in visual scenes with flexible linguistic expressions – not merely matching keywords. Many studies have employed automated metrics to evaluate the consistency between visual stimuli and narratives, such as calculating hit-rates of domain-specific information units or pattern-matching scores with fixed references [22, 31]. However, this automated approach, while efficient, fails to capture semantic alignment between linguistic expressions and visual concepts, reducing contextual nuance to binary co-occurrence metrics. Therefore, a significant gap persists in the development of automated, clinically viable frameworks capable of modeling semantic consistency in VSNs.

To address these critical challenges, we propose a novel framework termed GAP (Generate, Align, and Predict), to explicitly model the temporal and cross-modal consistency within VSNs for NCD detection. This framework leverages recent advances in Multimodal Large Language Models (MLLMs) and introduces a novel dynamic programming-based alignment paradigm to capture dynamic semantic alignments between narrative text and image sequences. Specifically, it operates through three core stages: (1) MLLM-based text generation for frame-level stories in sequence; (2) dynamic programming (DP)-based inter-image alignment for narrative segmentation, and intra-image similarity using finetuned contrastive vision-language model (VLM); (3) NCD detection using a classifier based on derived cross-modal consistency scores.

Established experiments on Cantonese CU-MARVEL corpus validate the efficiency of our proposed framework. The GAP model achieved the highest performance (F1=0.65, AUC=0.75) in NCD detection task, surpassing those of handcrafted acoustic, linguistic and pattern-matching feature sets.

In addition to achieving state-of-the-art performance on NCD detection, we conduct a in-depth analysis of the narrative responses collected from the Rabbit Story task. Our results demonstrate that specific images are particularly informative for evaluating crucial cognitive functions, such as object naming and story comprehension. By identifying these key moments and analyzing the narrative patterns of both healthy controls and NCD subjects, our study provides insights into the cognitive mechanisms underlying visual-stimulated narrative tasks and offers guidance for future assessment design and research.

The paper is organized as follows: Section 2 reviews related works, Section 4 introduces the proposed approaches, Section 5 presents our experimental settings and results, Section 6 analyzes and discusses experimental results, and Section 7 concludes this work and prospects future works.

2 Related Work

2.1 Unimodal NCD Detection

Early approaches to neurocognitive disorder (NCD) detection primarily relied on either acoustic or linguistic patterns independently, including prosody (e.g., increased dysfluencies and longer, more frequent pauses), phonetics (e.g., abnormal Mel Frequency Cepstral Coefficients patterns), phonology (e.g., higher rates of mispronunciation of words), lexicon (e.g., reduced type-token-ratio and elevated stopword ratio), syntax (e.g., reduced parse-tree depth), and semantics (e.g., abnormal patterns in word and sentence embeddings) [25, 27, 32, 33, 34, 35, 36, 26, 37, 38, 39, 40, 19]. The advent of deep learning shifted this paradigm from computerized handcrafted feature engineering toward embedding-based pretraining and fine-tuning approaches, yet remained constrained by single-modality limitations, failing to capture the relationship between narratives and corresponding visual stimuli [41, 8, 28, 42].

2.2 Reference-based Alignment for NCD Detection

To overcome unimodal constraints, extensive work attempted cross-modal alignment between visual stimuli and verbal narratives. These methods mainly depended on manually predefined keywords or rigid reference texts for calculating statistical correlations, such as ratios of information units from the images [20, 43, 22, 31]. However, these pattern-matching metrics are restricted to fixed-reference tasks (such as machine translation), making it lack adaptability to diverse narrative expressions in creative tasks like storytelling.

2.3 Application of Large Language Models for NCD Detection

Recent advances in multimodal large language models (MLLMs) have enabled more sophisticated cross-modal analysis, image quality assessment [44], image complexity evaluation [45], and – crucially for our work – textual quality metrics like relevance [46] and coherence scoring [47]. However, these works mainly focus on the static picture description task, which lacks the assessment of temporal organization skills. Moreover, the work directly using LLM as evaluator lacks interpretability due to their untraceable decision pathways, undermining reliability in clinical applications. These limitations prevent comprehensive modeling of cross-modal consistency between narratives and sequential images – a critical gap our framework addresses by integrating multimodal LLMs with dynamic alignment techniques.

3 Dataset

This study is conducted based on the CU-MARVEL-RABBIT (Cognitive Assessment Using Machine LeArning EmpoweRed VoicE AnaLysis – RABBIT Story Part) Corpus [38], which is designed to assess the cognitive capacities of individuals with NCDs in integrating sequential visual information and constructing coherent narratives.

The corpus records Cantonese-speaking participants narrating a story about a girl finding a rabbit through 15 cartoon images, as illustrated in Fig. 1. These images generally depict 5 sequential scenes (or topics): (1) the discovery and subsequent loss of the rabbit

Figure 1: Visual stimuli used in CU-MARVEL-RABBIT, which shows a girl and her dog searching for a rabbit and encountering various animals and adventures.

(images 1-4), (2) an encounter with a squirrel and some ants (images 5-7), (3) a meeting with a goose (images 7-9), (4) a run-in with a cow (9-11), and (5) a reunion with the rabbit followed by a farewell wave to its family (images 12-15).

These images are presented in a booklet, with each image displayed on a separate page. To eliminate the interference from non-narrative cognitive processes, participants will first review all 15 images to familiarize themselves with the content. Then, they will examine each image individually, providing a description of its content before proceeding to the next image. This process will continue sequentially until the stories for all 15 images have been completed.

Fig. 2 illustrates representative narrative examples (with coarse English translations for reader comprehension), highlighting the typical differences between HC and NCD narratives. Specifically, HC tends to capture most visual details, such as characters (girl, dog, rabbit, squirrel, goose, cow), objects (tree hollow, rock), and settings (house, garden, forest). In contrast, NCD often misses or misidentifies key details, such as overlooks squirrel, misidentifies goose as duck.

The CU-MARVEL-RABBIT Corpus (version "2024-06-13") consists of 381 manually-labeled samples, including 230 HCs and 151 individuals with NCDs. The data distribution and train-test split are listed in Table 1, where labels are assigned as follows: 0 (Healthy Control), 1 (Mild NCD), and 2 (Major NCD).

Table 1: Distribution of CU-MARVEL-RABBIT Corpus.

Label	Train	Test	*Total*
0	192	38	*230*
1	105	22	*127*
2	20	4	*24*
Total	*317*	*64*	*381*

4 Approach

To systematically assess cross-modal consistency in narrative descriptions provided by elderly participants, we propose a comprehensive pipeline that aligns narrative segments to visual story elements and quantifies their correspondence. Fig 3 shows the overall framework of our Generate, Align and Predict (GAP) system. Step 1 is generating a rich set of reference stories for each image using MLLM-generated narratives. Next, we introduce a dynamic programming-based segmentation algorithm to partition each subject's narrative transcription into segments, with each segment aligning to a specific image. After obtaining the image-text pairs, we are able to quantitatively evaluate the alignment degree between the segmented narratives and the corresponding images. To this end, we employ the Bootstrapped Language-Image Pretraining (BLIP) [48] multimodal model, leveraging its Image-Text Contrastive (ITC) and Image-Text Matching (ITM) scores as quantitative measures of cross-modal consistency. Recognizing the domain shift between our dataset and BLIP's pretraining data, we further finetune BLIP on our task-specific references and images to enhance its evaluative capacity. Finally, the extracted ITC and ITM scores for each image-text pair are aggregated as features and input into an XGBoost classifier for neurocognitive disorder detection.

4.1 Generate: MLLM-based Reference Story Generation

VSNs produced by participants are usually long and unstructured, often lacking explicit temporal markers that link narrative segments to their corresponding visual elements. This structural ambiguity complicates the quantification of temporal and cross-modal consistency, as unstructured narratives may conflate disparate visual concepts. To address this, we adopt MLLMs (e.g., GPT-4o [49]) to generate structured stories that explicitly align to each image, which serve as foundational references for subsequent segmentation step.

To systematically refine the quality of generated reference stories, our framework employs an iterative optimization loop between a generator and an evaluator, as illustrated in the Step 1 of Figure 3. The process begins with the generator model – GPT-4o [49], which is capable of processing interleaved text and image inputs. We carefully design the prompt that formatted in the template: "`<system instruction> + <interleaved text-image instruction> + <integrate instruction> + <examples>`".

The `<system instruction>` component outlines the overall goal and constraints for generation, such as: "*You are engaged in a storytelling task based on a sequence of images. Your story should be in Cantonese and keep a concise, coherent and colloquial style.*"

The `<interleaved text-image instruction>` guides the model to process each image sequentially, generating contextual descriptions that incorporate visual elements and narrative continuity. For example, the instruction of the first image is "*Let's start by describing the first image (e.g., a girl finds a rabbit aside a rock), including the settings (e.g., forest, garden, house, living room, daytime, morning, afternoon) and the main characters (e.g., girl, dog, rabbit).* `<IMAGE 1>`"[1]. Subsequent instructions not only include similar descriptive prompts but also emphasize narrative continuity by adding a clause "*Based on the events in previous images,*" at the beginning.

[1]The details are adopted from the Narrative Scoring Scheme of Rabbit Story.

嗷呢有個小朋友，諗女仔嚟嘅，佢帶住隻狗，就去花園嗰度，見到隻兔仔喺度喎。嗷佢啊帶咗隻兔仔呢，就返咗入屋企玩囉喎。嗷呢，跟住佢啊同隻狗狗去沖涼，隻兔仔啊出咗嚟。嗷啊突然之間佢返到入去，唔見咗隻兔仔喎。**(Rough meaning:** A girl, while walking her dog in the garden, found a rabbit. She took it home, but it escaped while she was bathing her dog.)佢就周圍搵啦。出翻個樹嗰度，個窿嗰度叫啦。嗷啊狗狗呢，就睇住個蟻蟻蟻竇啦嗷樣樣。嗷啊原來個樹窿嗰度嗰個松鼠嚟個喎，佢話見唔到佢。嗷嗰啲蟻呢就躡咗落去隻狗狗度啦。啊，躡到隻狗狗度呢就、就隻狗狗都唔舒服。(She then searched for it. She went out to the forest and called out to a tree hole. The dog watched an ant nest. A squirrel went out the hole and said it hadn't seen the rabbit. The ants crawled onto the dog, making it uncomfortable.)嗷佢又見到隻鵝囉喎，有隻佢問隻鵝見唔見到隻兔仔啊。嗷呢。嗷隻鵝呢就話帶佢去囉喎，嗷隻狗狗呢，嗰啲。諗諗蟻呢就開始跌、跳走啦。(The girl then saw a goose and asked if it had seen the rabbit. The goose offered to help, and the ants began to fall off the dog.)嗷佢哋，諗行行行，行到一嚿大石嗰度企上去嗌囉喎，嗷樣樣。喺度嗌嗰隻，兔仔喎。點知呢，原來嗰隻呢，係隻牛嚟嘅，唔係嚿石，啊嚇到佢兩個呢。就、跌咗出落嚟喎。原來跌咗落水池度喎。(They walked until reached what they thought was a large rock and stood on it to call for the rabbit. It turned out to be a cow, not a rock, startling and causing them to fall into a pond.)嗷樣呢。嗷佢哋要爬翻上去啦嗷。爬翻上去呢，咦，爬翻出嚟嗰陣時，見到草叢入便有兔仔喺度喎。原來呢，兔仔嘅家族住晒喺入便嘅，嗷佢哋啊好開心啦見到佢。嗷啊最尾就同佢哋拜拜啦，返屋企啦。(They climbed back up and then noticed the rabbit in the grass. It turned out the rabbit's family lived there. They were happy to see it and finally said goodbye and headed back home.)

(a) HC

個女仔，就同隻狗一齊。嗷跟住佢就同，隻兔仔講嘢嗷啊。跟住個女仔呢。就同隻狗。同隻兔仔一齊玩。嗯啊。係咪嗷啊。嗯。跟住個女仔又同隻狗仔沖涼。洗面。隻兔仔下便。哦嗷樣啊 。嗰隻狗喺上便，個女仔喺下便，諗住，唔知講乜嘢。**(Rough meaning:** A girl, with a dog, talked to a rabbit. Then the girl and the dog played with the rabbit. Is it like that? Then the girl took a bath with the dog, washing its face. A rabbit under… Oh that's it. The dog was on the top and the girl was on the bottom… thinking… I don't know what to say.)嗰個女仔，就喺度，叫唔知，同佢狗講嘢，佢話好大聲嗷樣講。嗯。嗯跟住有隻唔知乜嘢走出嚟啊我唔識講啊。吓。跟住個女仔又喺度嗌。(The girl, right there, said don't know, talking to her dog, she said very loudly. Then something I don't know came out. I can't say. Huh. Then the girl shouted again.)嗷啊有一隻狗仔喺度。跟住隻女仔，個女仔就同隻狗仔呢。去追隻企鵝。跟住個女仔呢。就喺度同隻狗仔望喺個樹上便。嗯個女仔坐喺度。嗯啊依隻唔知係咪狗仔我睇唔到啊。(Then a dog, then a girl... A girl with a dog was chasing after a penguin. Then the girl here was looking at the tree with the dog. The girl was sitting there. Um, I don't know if it's a dog. I can't see it.)又有隻牛喺度，喺個水嗰度玩。嗰隻，嗰個女仔呢，就喺度喺個、草地嗰度玩，同嗰隻狗喺度游水。(There was another cow playing by the water. That, that girl was playing on the grass and swimming with the dog.)哎睇唔到。跟住隻狗仔就同，個女仔一齊，同啲兔仔玩。哎。諗仲有呢個。唔好意思吓。喱個女仔呢就同啲兔仔喺度，揮手。嗰隻狗仔喺後便，嗷企住囉。(I can't see it. Then the dog played with the rabbits with the girl. Uh, there's... I'm sorry. The girl was there with the rabbits, waving. The dog was standing in the back.)

(b) NCD

Figure 2: Illustrative Cantonese narratives with rough English meaning in CU-MARVEL-RABBIT.

Then, the model is instructed to weave all previous elements into a coherent and unified story using the `<integrate instruction>`, along with randomly selected `<examples>` to guide the storytelling style and structure. The `<integrate instruction>` emphasizes continuity, logical flow, and emotional engagement, such as: "*Now, review all fifteen images and their descriptions again. Then, tell me a continuous, cohesive, and engaging story that weaves all these elements together. Your story should be in Cantonese and keep a concise, coherent and colloquial style. Structure your story by numbering each paragraph to correspond with each image (e.g., 1., 2., ..., 15.). Think about how each scene contributes to the development of characters, settings, and plot. Develop a storyline that evolves with each image, ensuring that the transitions between paragraphs are fluid and meaningful. For each transition, think deeply about the characters' motivations and actions that logically and emotionally bridge the each image to the previous or the next images.*" While the `<examples>` are selected from seed references curated by six research team members who are familiar with the clinical scoring scheme[2].

To evaluate the quality of generated stories, we calculate the text similarity metrics, including BLEU-n (n=1,2,3,4), METEOR, and ROUGE-L, between each generated text and our six seed reference stories. These metrics can guide the improvement of the designed prompt and the selection of top-scored stories, resulting in 80 high-quality references retained for subsequent processing.

4.2 Align: Dynamic Programming for Narrative Segmentation

Since there are no manual annotations indicating which part of the transcription corresponds to each image, we design an automatic segmentation algorithm that aligns each narrative segment to its corresponding image in the sequence. Specifically, we aim to identify 14 segmentation points located at punctuation marks, such that the resulting 15 text segments align with the 15 images in the correct order.

For the purpose of this work, we define a *text unit* as the substring of the transcription that lies between two consecutive punctuation marks (e.g., periods, commas, question marks, etc.). Thus, the entire transcription can be represented as a sequence of text units, which serve as the minimal granularity for possible segmentation boundaries.

We formulate this task as an optimization problem, where the objective is to maximize the overall similarity score between segmented texts and their corresponding images. Considering the sequential nature and the existence of optimal substructure in this problem, we employ a dynamic programming (DP) approach to decompose it into overlapping subproblems.

State Definition. Let $dp[j][k]$ denote the maximum cumulative similarity score achievable by dividing the first j text units into k segments, where $j \geq 1$, $1 \leq k \leq K$, and K is the number of visual stimuli.

Recurrence Relation. The DP transition can be formulated as:

$$dp[j][k] = \max_{1 \leq i < j} \left\{ dp[i][k-1] + \text{sim}\left(T_{i:j}, I_k\right) \right\} \tag{1}$$

where $T_{i:j}$ denotes the concatenated text segment formed by joining text units from position i to j, and I_k is the k-th image in the sequence. The boundary conditions are (1) $dp[1][1] = 0$, as no actual split occurs at the anchor point; and (2) $dp[j][k] = -\infty, \forall k > j$, since we assume each image must correspond to at least one text unit and thus we cannot have more segments than the number of text units.

Similarity Calculation. The similarity score $\text{sim}\left(T_{i:j}, I_k\right)$ between a candidate text segment and image I_k is calculated by leveraging the 80 previously generated reference captions for each image. We encode both the candidate segment and each reference using a

[2]The scoring rubric provides detailed criteria, elaborations and examples for evaluating narratives across seven dimensions: introduction, character development, mental and emotional states, referencing, conflict/resolution (or event/reaction), cohesion, and conclusion. The seed references are curated to align with the proficient level.

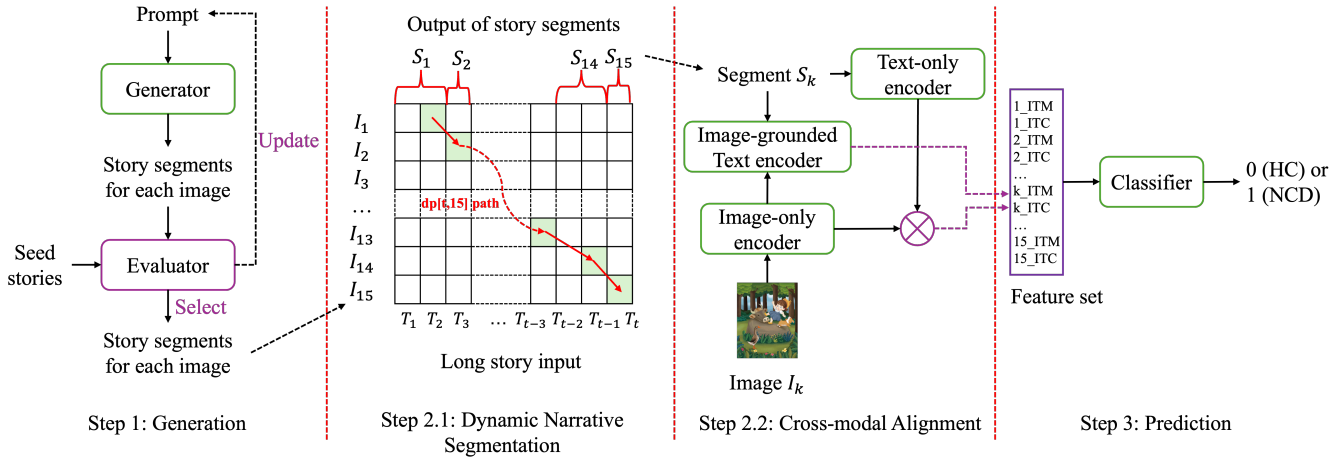


Figure 3: The overall framework of GAP, 0.95 three core steps: (1) MLLM-based text generation for frame-level stories; (2) dynamic programming (DP)-based text segmentation for image-text alignment, and BLIP-based intra-image cross-modal consistency evaluation with alignment scores; (3) classification.

Cantonese SentenceBERT encoder [50], then compute the mean cosine similarity:

$$\text{sim}\left(T_{i:j}, I_k\right) = \frac{1}{N}\sum_{n=1}^{N} \text{CosSim}\left(\text{SBERT}(T_{i:j}), \text{SBERT}(R_k^{(n)})\right) \quad (2)$$

where $N = 80$, $R_k^{(n)}$ is the n-th reference for image I_k, and $\text{SBERT}(\cdot)$ denotes the SentenceBERT encoding function. $\text{CosSim}(\cdot,\cdot)$ computes the cosine production between two vectors. This similarity score is only used in segmentation algorithm.

Fig. 3 Step2.1 visualize this optimization process with DP matrix. By maximizing the total similarity score via this dynamic programming formulation, we obtain an optimal segmentation of the narrative, in which each text segment is aligned with its corresponding image in sequence.

To verify the accuracy of the proposed DP-based automatic segmentation algorithm, we manually annotate the ground-truth segmentation boundaries for all narratives in the test set. The test set comprises a total of 3,358 text units. Among these, only 437 text units were assigned to incorrect images by the automatic algorithm, resulting in an overall segmentation accuracy of 87%. Notably, of the misclassified cases, only 32 text units were assigned to images that are not adjacent to their ground-truth images, indicating that the majority of segmentation errors involve only minor boundary shifts rather than substantial misalignments.

4.3 Align: Task-specific Fine-tuning of the BLIP Model & Alignment Scores Extraction

To quantitatively evaluate the alignment between each image and its corresponding narrative segment, we propose to leverage VLM originally designed for image retrieval and image captioning tasks etc. Such models provide an alignment score that reflects the semantic consistency between visual and textual modalities.

However, the availability of pre-trained VLMs in Cantonese is extremely limited due to the low-resource nature of the language. To address this challenge, we employ the Bing Translation API to translate all reference narratives and subject transcriptions from Cantonese to English. This translation enables us to exploit the robust capabilities of established English-language VLMs, specifically the BLIP model [48].

To adapt BLIP to the domain of the Rabbit Story, we perform task-specific fine-tuning using our generated references as ground-truth narratives paired with the 15 corresponding Rabbit story images. Each image is paired with 80 MLLM-generated reference captions, resulting in a total of 1,200 positive image-text pairs. Negative pairs are constructed by shuffling images and captions such that the text does not correspond to the associated image.

Given that our objective is to assess the degree of alignment between images and textual descriptions, we focus exclusively on optimizing the ITC and ITM components of BLIP during fine-tuning. Both ITC and ITM scores are designed to capture fine-grained and global cross-modal correspondence, respectively. The ITC objective encourages representations of corresponding image-text pairs to be close in the joint embedding space, while the ITM objective promotes accurate binary matching decisions between images and texts. The details of the ITC and ITM fine-tuning procedures are consistent with those described in the original BLIP paper [48].

After fine-tuning the BLIP model, we utilize it to extract alignment scores for downstream analysis. As illustrated in Fig. 3 step 2.2, when a subject's narrative transcription is provided, we first apply the proposed dynamic programming-based segmentation algorithm to divide the text into 15 segments, each corresponding to a specific image. Each of these segments is then translated into English, resulting in 15 English image-text pairs for each subject. These pairs are subsequently input into the fine-tuned BLIP model, from which we extract the ITC and ITM scores for every pair. The ITC score represents the cosine similarity between two embedding vectors, ranging from -1 to 1, while the ITM score reflects a probability output, ranging from 0 to 1. Consequently, each subject is represented by a feature vector of dimension $15 \times 2 = 30$, encapsulating the cross-modal consistency between the narrative and the visual story.

4.4 Predict: NCD Prediction via XGBoost

Following feature extraction, the task of NCD prediction is formulated as a binary classification problem, where each subject is represented by a 30-dimensional feature vector composed of ITC and ITM scores for the 15 image-text pairs. In this work, we employ the XGBoost classifier due to its strong predictive performance, high interpretability, and the ability to facilitate subsequent analysis of feature importance. XGBoost is a gradient-boosted decision tree framework that efficiently handles structured data and provides built-in mechanisms for evaluating the contribution of individual features to the classification outcome. This property is particularly valuable for interpreting the diagnostic relevance of each image-text consistency measure in relation to neurocognitive status.

5 Experiments & Results

5.1 Experimental settings

For the multimodal alignment model, we utilize the *blip-itm-large-coco* checkpoint available on HuggingFace[3], which is pretrained for image-text matching on the COCO dataset using a Vision Transformer (ViT-Large) as its backbone. During fine-tuning, we employ the AdamW optimizer with an initial learning rate of 5×10^{-5}, and fine-tune the model for 5 epochs using four NVIDIA A800 GPUs.

For the classifier, we adopt the XGBoost framework and conduct an exhaustive grid search to optimize its hyper-parameters. The grid search is performed using 5-fold cross-validation on the training set, and the combination of parameters yielding the highest average F1 score are selected for the final model. In this work, we merge the Mild NCD and Major NCD groups into a single positive class (label 1), while the Healthy Control group is designated as the negative class (label 0), thereby formulating the task as a binary classification problem.

To ensure experimental reproducibility, all random seeds and random states are fixed to 42 across the entire pipeline, including data shuffling, model initialization, and classifier training.

5.2 Baseline

For comparison, we adopt both micro-based and reference-based systems as established in [23]. The micro-based system computes statistics from basic narrative units, encompassing 10 acoustic and 13 linguistic features. In contrast, the reference-based system evaluates narratives at a higher level by comparing them to curated references, yielding 10 coverage ratios and 6 pattern-matching metrics. Together, these systems provide a comprehensive set of baselines—including 23 acoustic and linguistic features and 16 similarity and coverage metrics—against which we evaluate the effectiveness of our proposed method using the same classifier. These baselines serve as a benchmark for evaluating the effectiveness of our proposed method using the same classifier.

5.3 NCD Detection Results

Table 2 presents the results of NCD detection on the CU-MARVEL-RABBIT corpus, comparing various feature sets using the same classifier, with AUC as the primary evaluation metric. Values in bold indicate the highest performance achieved for each respective metric. As shown in the table, the combination of ITM and ITC features (ITM+ITC) yields the best overall performance across all evaluation metrics, achieving an AUC of 0.7500, accuracy of 0.7500, and precision of 0.7500. Notably, removing either ITM or ITC features results in a drop in performance, indicating that these two feature sets complement each other and together contribute to the improved effectiveness of the model.

[3]https://huggingface.co/Salesforce/blip-itm-large-coco

Table 2: Performance Metrics for Different Features on CU-MARVEL-RABBIT.

Feature	AUC	F1	Acc	Recall	Precision
Micro [23]	0.6377	0.5106	0.6406	0.4615	0.5714
Reference [23]	0.6690	0.5556	0.6250	**0.5769**	0.5357
ITC	0.7115	0.5417	0.6562	0.5000	0.5909
ITM	0.7065	0.5532	0.6719	0.5000	0.6190
ITM+ITC	**0.7500**	**0.6522**	**0.7500**	**0.5769**	**0.7500**

6 Discussion & Analysis

6.1 Feature Importance Analysis

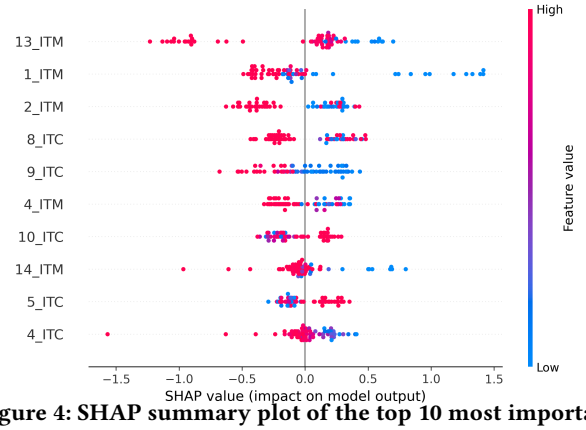
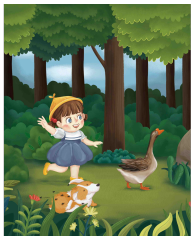


Figure 4: SHAP summary plot of the top 10 most important features for the XGBoost classifier. Feature values are color-coded from low (blue) to high (red); SHAP values reflect each feature's impact on model prediction

To further interpret the feature importance of our NCD detection framework for the Rabbit Story task, we investigate which images or patterns contribute most to distinguishing NCD subjects from HC. Figure 4 presents the SHAP summary plot of the top 10 most important features, where SHAP values quantify each feature's impact on the XGBoost classifier's predictions.

As shown in Figure 4, features associated with images 1, 2, 9, and 13 demonstrate particularly high importance for NCD classification. Notably, these images all correspond to moments of "scene transition," where the narrative context shifts to a new scene. Accurate understanding and narration of these transitions require subjects to comprehend the evolving storyline and maintain logical coherence. Such cognitive abilities are strongly associated with higher cognitive functioning; thus, subjects with neurocognitive impairment are



Low scores example:
ID: TBS0072, 8_ITC: -0.1267, 8_ITM: 0.0017, Label: 1
Text segment: The puppy, little girl and the duck are going to find the bunny together.

High score example:
ID: TBS0122, 8_ITC: 0.9874, 8_ITM: 0.9126, Label: 0
Text segment : So the little goose has come out along with the girl and the puppy.

(a)

Low score example:
ID: TBS0852, 9_ITC: -0.1237, 9_ITM: 0.0303, Label: 1
Text segment : I saw a little rabbit coming, and I have no idea what it was doing. It was here playing or something, but I really don't understand. Oh, and those two are standing there admiring something at a tree. Really.

High scores example:
ID: TBS0429, 9_ITC: 0.9764, 9_ITM: 0.9923, Label: 0
Text segment : So, as they walk, she stands there thinking that there might be a rock, and she's standing on top, looking into the distance to see if she can find the rabbit.'

(b)

Figure 5: Example text segments with low and high ITC & ITM scores for (a) image 8 and (b) image 9. Red highlights indicate abnormal or unclear expressions in low score examples.

more likely to struggle with these transitions, leading to inaccurate narrative with low alignment scores.

Additionally, features corresponding to image 8 are prominent among the top contributors. Image 8 includes multiple characters and objects (the girl, dog, goose, and ant), requiring subjects to correctly identify and name a diverse set of visual elements. This aspect of the narrative engages object and character naming abilities, which are often impaired in NCD patients. The elevated importance of this feature supports its relevance for differential diagnosis in the Rabbit Story paradigm.

6.2 Case Studies and Linguistic Pattern Analysis

To gain deeper insight into the cognitive mechanisms underlying NCD detection with our approach, we conduct in-depth case studies focusing on image 8 and image 9, two pictures that capture distinct cognitive evaluation aspects highlighted in Section 6.1. Our objective is to elucidate which visual objects are frequently misrecognized by NCD subjects in image 8, and whether subjects with NCD misinterpret or lose track of the storyline in image 9.

We first employ a qualitative comparison by selecting representative examples with high and low ITC/ITM scores for each image. As shown in Fig. 5(a), for image 8, the low-score narrative (from an NCD subject) erroneously identifies the "goose" as a "duck," while the high-score narrative (from a healthy control) correctly recognizes all key characters and constructs a coherent description of the visual scene. To quantitatively validate this observation, we conduct a TF-IDF analysis over all narrative segments corresponding to image 8. Specifically, we construct a TF-IDF matrix for these segments, compute the average TF-IDF vector for the HC group and for the NCD group, and then subtract the NCD group average vector from the HC group average vector to obtain the group-level differences. Fig. 6(a) displays the top 20 terms with the largest TF-IDF score differences between groups. Red terms are more frequent in NCD, blue in HC. Consistent with the qualitative case, NCD subjects mention "duck," "ducks," and "duckling" more often, while "goose" is much more salient in HC narratives. Furthermore, HC subjects are more likely to mention secondary objects, such as "ants," suggesting better attention to visual detail. These results indicate that precise object recognition—especially for "goose" and "ant"—is a key discriminator in this context.

For image 9, as depicted in Fig. 5(b), the low-score narrative from the NCD group is marked by confusion and lack of narrative coherence, with explicit expressions of uncertainty (e.g., "I have no idea what it was doing" or "I really don't understand"). In contrast, the high-score example is well-structured and closely aligned with the depicted story context. The TF-IDF analysis for image 9, presented in Fig. 6(b), yields further insights. Surprisingly, the term "rabbit" appears much more frequently in HC narratives, despite the rabbit not being directly shown in image 9. This can be explained by understanding the narrative logic: at this point in the story, the girl is attempting to locate the rabbit, so she climbs onto a rock and calls for the rabbit from a higher vantage point. This prompt HC subjects to refer to the rabbit even in its absence. This ability to maintain narrative continuity and refer to relevant but non-present entities is a hallmark of healthy cognition. Additionally, the HC group frequently uses terms like "calling," indicating a coherent grasp of the plot. By contrast, NCD narratives often include irrelevant actions or objects (e.g., "singing," "playing"), reflecting confusion or failure to follow the story. The term "don" (from "don't") is also prominent in NCD narratives, which may indicate negative emotions or cognitive difficulty. Overall, these findings suggest that the ability to accurately interpret and logically narrate the storyline is a crucial factor in distinguishing between NCD and HC in image 9.

Our analytic framework is readily extensible to other images and yields similar observations. For example, in image 6, NCD subjects often misidentify a squirrel as a mouse, while in image 11, NCD subjects may fail to grasp the storyline that the girl falls from the back of a buffalo into a pond, instead describing her as swimming. These results collectively demonstrate that the proposed multimodal and linguistic analysis, together with the GAP system, provides interpretable insights into the operation of visual-stimulated narrative-based approaches for NCD detection. This methodology not only identifies key points for distinguishing NCD from HC but also informs future research on cognitive evaluation using narrative-based tasks.

6.3 Representative Case Comparison by Cognitive Rank

In addition to the analysis of our model-derived features, we compare the ITC and ITM scores with subjects' performance on well-established cognitive assessments. The original CU-MARVEL dataset includes standardized test scores, such as MoCA and AD8, which were normalized for each subject by age and education level to generate cognitive ability rankings. Based on these normalized ranks, we select representative samples from the top, middle, and bottom positions within each diagnostic subgroup – HC, Mild NCD, and Major NCD – and compute the mean and standard deviation of their ITC and ITM scores.

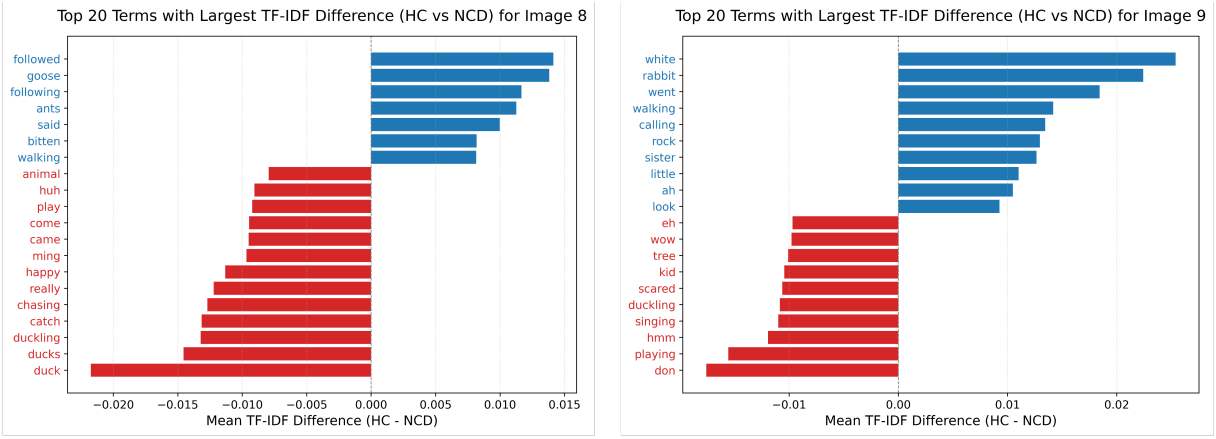


Figure 6: Top 20 terms with the largest TF-IDF differences between the HC and NCD groups for (a) image 8 and (b) image 9. Red terms are more frequent in the NCD group, while blue terms are more frequent in the HC group.

Table 3: Summary statistics of ITC and ITM scores across groups and cognitive ranks. Ranking (top, mid, bottom) within each group is determined based on age- and education-normalized scores from established cognitive assessments (e.g., MoCA, AD8).

Group	Rank	ID	ITC mean	ITC std	ITM mean	ITM std
HC	Top	TBS0972	0.881	0.152	0.969	0.044
	Mid	TBS0932	0.675	0.393	0.661	0.435
	Bottom	TBS0223	0.676	0.427	0.658	0.438
Mild NCD	Top	TBS0899	0.613	0.452	0.598	0.411
	Mid	TBS0487	0.509	0.438	0.472	0.462
	Bottom	TBS0541	0.418	0.459	0.497	0.432
Mild NCD	Top	TBS0847	0.557	0.413	0.567	0.418
	Mid	TBS0898	0.376	0.509	0.344	0.387
	Bottom	TBS0042	0.359	0.465	0.355	0.404

The results are summarized in Table 3. A clear and consistent trend emerges: higher-ranked individuals (e.g., HC top) exhibit higher mean ITC and ITM scores, while those with lower cognitive rankings (e.g., Major NCD bottom) have markedly reduced scores. This concordance between multimodal feature scores and cognitive test rankings underscores the validity of our model's alignment metrics as indicators of cognitive health.

Despite this overall trend, it is important to note that the statistical gap between the HC bottom sample and the Mild NCD top sample is relatively narrow. These "boundary cases" lie at the diagnostic interface between healthy and impaired cognition, presenting unique challenges for NCD detection. Such cases highlight the complexity of distinguishing early-stage NCD from healthy aging, and point to the need for more granular, image-level feature analysis.

In future work, we plan to conduct an in-depth investigation of these boundary cases, examining ITC and ITM scores for individual images. Our goal is to identify specific visual stimuli that are most informative for resolving ambiguity in borderline diagnostic cases, thereby enhancing the sensitivity and specificity of VSN-based NCD detection.

7 Conclusion

In this work, we present the GAP framework for a Cantonese visual-stimulated narrative cognitive assessment. The GAP framework introduces several key innovations: (1) the use of multimodal LLM to generate high-quality reference narratives for each story image, (2) a dynamic programming-based segmentation algorithm to accurately segment and align long-form subject narratives to individual images, and (3) a task-specific fine-tuning approach for vision-language models on the Rabbit Story dataset, enabling robust extraction of ITC and ITM scores that quantify image-text alignment for NCD prediction. Our experiments demonstrate that this pipeline achieves state-of-the-art performance in automatic NCD detection, validating the effectiveness of the GAP framework.

Beyond the core methodology, we analyze cognitive aspects revealed by the Rabbit Story task, finding that certain images are key to assessing specific functions. Images with multiple characters or distinct objects aid in evaluating naming abilities, while those requiring storyline comprehension reveal narrative coherence and logical progression. Negative emotions or self-limiting expressions in narratives may also indicate cognitive impairment.

In summary, our study demonstrates the promise of multimodal narrative analysis for NCD detection and provides actionable guidance for future research on narrative-based cognitive assessment. The GAP framework offers a flexible and interpretable approach that can be further adapted to other languages, tasks, and clinical settings.

Acknowledgments

This work is supported by the HKSARG Research Grants Council's Theme-based Research Grant Scheme (Project No. T45-407/19N).

References

[1] APA et al. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Association (APA).

[2] Perminder S Sachdev, Deborah Blacker, Dan G Blazer, Mary Ganguli, Dilip V Jeste, Jane S Paulsen, and Ronald C Petersen. 2014. Classifying neurocognitive disorders: the DSM-5 approach. *Nature Reviews Neurology*.

[3] Alzheimer's Disease International. 2024. *World Alzheimer's Report 2024: Global changes in attitudes to dementia*. Alzheimer's Disease International.

[4] Yuejiao Wang, Xianmin Gong, Xixin Wu, Patrick Wong, Hoi-lam Helene Fung, Man Wai, and Helen Meng Mak. 2024. Naturalistic language-related movie-watching fmri task for detecting neurocognitive decline and disorder. *ISCA ISCSLP*.

[5] Yuejiao Wang, Xianmin Gong, Lingwei Meng, Xixin Wu, and Helen Meng. 2024. Large language model-based fmri encoding of language functions for subjects with neurocognitive disorder. *INTERSPEECH*.

[6] Yu Guo et al. 2024. Multiplex cerebrospinal fluid proteomics identifies biomarkers for diagnosis and prediction of alzheimer's disease. *Nature Human Behaviour*.

[7] Isabella Veneziani, Angela Marra, Caterina Formica, Alessandro Grimaldi, Silvia Marino, Angelo Quartarone, and Giuseppa Maresca. 2024. Applications of artificial intelligence in the neuropsychological assessment of dementia: a systematic review. *Journal of Personalized Medicine*.

[8] Rahul Sharma, Tripti Goel, Muhammad Tanveer, Chin-Teng Lin, and R Murugan. 2023. Deep-learning-based diagnosis and prognosis of alzheimer's disease: a comprehensive review. *IEEE Transactions on Cognitive and Developmental Systems*.

[9] Aaron C Lim, Lisa L Barnes, Gali H Weissberger, Melissa Lamar, Annie L Nguyen, Laura Fenton, Jennifer Herrera, and S Duke Han. 2023. Quantification of race/ethnicity representation in alzheimer's disease neuroimaging research in the usa: a systematic review. *Communications medicine*.

[10] Walter Kintsch and Teun van Dijk. 1978. Cognitive psychology and discourse: recalling and summarizing stories. *Current trends in text linguistics*.

[11] Andrea Marini, Anke Boewe, Carlo Caltagirone, and Sergio Carlomagno. 2005. Age-related differences in the production of textual descriptions. *Journal of psycholinguistic research*.

[12] William Labov. 2010. Oral narratives of personal experience. *Cambridge encyclopedia of the language sciences*.

[13] Juliana Onofre De Lira, Karin Zazo Ortiz, Aline Carvalho Campanha, Paulo Henrique Ferreira Bertolucci, and Thaís Soares Cianciarullo Minett. 2011. Microlinguistic aspects of the oral narrative in patients with alzheimer's disease. *International Psychogeriatrics*.

[14] Cíntia Matsuda Toledo, Sandra Maria Aluísio, Leandro Borges Dos Santos, Sonia Maria Dozzi Brucki, Eduardo Sturzeneker Trés, Maira Okada de Oliveira, and Letícia Lessa Mansur. 2018. Analysis of macrolinguistic aspects of narratives from individuals with alzheimer's disease, mild cognitive impairment, and no cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*.

[15] Dingyi Yang and Qin Jin. 2024. What makes a good story and how can we measure it? a comprehensive survey of story evaluation. *arXiv preprint arXiv:2408.14622*.

[16] Florence Pasquier, Florence Lebert, Laurence Grymonprez, and Henri Petit. 1995. Verbal fluency in dementia of frontal lobe type and dementia of alzheimer type. *Journal of Neurology, Neurosurgery & Psychiatry*.

[17] Laura M Wright, Matteo De Marco, and Annalena Venneri. 2023. Current understanding of verbal fluency in alzheimer's disease: evidence to date. *Psychology Research and Behavior Management*.

[18] Morteza Rohanian, Julian Hough, and Matthew Purver. 2021. Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. *arXiv:2106.15684*.

[19] Eric Williams, Catherine Theys, and Megan McAuliffe. 2023. Lexical-semantic properties of verbs and nouns used in conversation by people with alzheimer's disease. *PLoS One*.

[20] Elaine Giles, Karalyn Patterson, and John R Hodges. 1996. Performance on the boston cookie theft picture description task in patients with early dementia of the alzheimer's type: missing information. *Aphasiology*.

[21] Katherine L Possin. 2010. Visual spatial cognition in neurodegenerative disease. *Neurocase*.

[22] Jinchao Li, Kaitao Song, Junan Li, Bo Zheng, Dongsheng Li, Xixin Wu, Xunying Liu, and Helen Meng. 2023. Leveraging pretrained representations with task-related keywords for alzheimer's disease detection. *IEEE ICASSP*.

[23] Jinchao Li* et al. 2025. Detecting neurocognitive disorders through analyses of topic evolution and cross-modal consistency in visual-stimulated narratives. In *ArXiv*. IEEE.

[24] Jiahong Yuan, Xingyu Cai, Yuchen Bian, Zheng Ye, and Kenneth Church. 2021. Pauses for detection of alzheimer's disease. *Frontiers in Computer Science*.

[25] Jinchao Li, Jianwei Yu, Zi Ye, Simon Wong, Manwai Mak, Brian Mak, Xunying Liu, and Helen Meng. 2021. A comparative study of acoustic and linguistic features classification for alzheimer's disease detection. *IEEE ICASSP*.

[26] Rini Pl and Gayathri Ks. 2024. Cognitive decline assessment using semantic linguistic content and transformer deep learning architecture. *International Journal of Language & Communication Disorders*.

[27] M Zakaria Kurdi. 2024. Dementia detection with phonetic and phonological features. In *FLAIRS*.

[28] S Saeedi, S Hetjens, MOW Grimm, and Ben Barsties v. Latoszek. 2024. Acoustic speech analysis in alzheimer's disease: a systematic review and meta-analysis. *The Journal of Prevention of Alzheimer's Disease*.

[29] Philippine Geelhand, Fanny Papastamou, Gaétane Deliens, and Mikhail Kissine. 2020. Narrative production in autistic adults: a systematic analysis of the microstructure, macrostructure and internal state language. *Journal of Pragmatics*.

[30] Anthony Pak-Hin Kong. 2023. *Spoken Discourse Impairments in the Neurogenic Populations*. Springer.

[31] Junan Li, Yunxiang Li, Yuren Wang, Xixin Wu, and Helen Meng. 2024. Devising a set of compact and explainable spoken language feature for screening alzheimer's disease. *ISCA ISCSLP*.

[32] Maysa Luchesi Cera, Karin Zazo Ortiz, Paulo Henrique Ferreira Bertolucci, Tamy Tsujimoto, and Thaís Minett. 2023. Speech and phonological impairment across alzheimer's disease severity. *Journal of Communication Disorders*.

[33] Jiawen Kang, Junan Li, Jinchao Li, Xixin Wu, and Helen Meng. 2024. Not all errors are equal: investigation of speech recognition errors in alzheimer's disease detection. *ISCA ISCSLP*.

[34] Thomas Melistas, Lefteris Kapelonis, Nikos Antoniou, Petros Mitseas, Dimitris Sgouropoulos, Theodoros Giannakopoulos, Athanasios Katsamanis, and Shrikanth Narayanan. 2023. Cross-lingual features for alzheimer's dementia detection from speech. In *INTERSPEECH*.

[35] Mahboobeh (Mah) Parsapoor (Parsa), Muhammad Raisul Alam, and Alex Mihailidis. 2023. Performance of machine learning algorithms for dementia assessment: impacts of language tasks, recording media, and modalities. en-US. *BMC Medical Informatics and Decision Making*.

[36] Ning Liu, Zhenming Yuan, Yan Chen, Chuan Liu, and Lingxing Wang. 2023. Learning implicit sentiments in alzheimer's disease recognition with contextual attention features. *Frontiers in Aging Neuroscience*.

[37] Spyridoula Varlokosta, Katerina Fragkopoulou, Dimitra Arfani, and Christina Manouilidou. 2024. Methodologies for assessing morphosyntactic ability in people with alzheimer's disease. *International journal of language & communication disorders*.

[38] Helen Meng et al. 2023. Integrated and enhanced pipeline system to support spoken language analytics for screening neurocognitive disorders. In *INTERSPEECH*.

[39] Olga Ivanova, Israel Martínez-Nicolás, Elena García-Piñuela, and Juan José G Meilán. 2023. Defying syntactic preservation in alzheimer's disease: what type of impairment predicts syntactic change in dementia (if it does) and why? *Frontiers in Language Sciences*.

[40] Maria Kaltsa, Anthoula Tsolaki, Ioulietta Lazarou, Ilias Mittas, Mairi Papageorgiou, Despina Papadopoulou, Ianthi Maria Tsimpli, and Magda Tsolaki. 2024. Language markers of dementia and their role in early diagnosis of alzheimer's disease: exploring grammatical and syntactic competence via sentence repetition. *Journal of Alzheimer's Disease Reports*.

[41] Israel Martínez-Nicolás, Thide E Llorente, Francisco Martínez-Sánchez, and Juan José G Meilán. 2021. Ten years of research on automatic voice and speech analysis of people with alzheimer's disease and mild cognitive impairment: a systematic review article. *Frontiers in Psychology*.

[42] Muath Alsuhaibani, Ali Pourramezan Fard, Jian Sun, Farida Far Poor, Peter S Pressman, and Mohammad H Mahoor. 2024. A review of deep learning approaches for non-invasive cognitive impairment detection. *arXiv:2410.19898*.

[43] Veronika Vincze, Gábor Gosztolya, László Tóth, Ildikó Hoffmann, and Gréta Szatlóczki. 2016. Detecting mild cognitive impairment by exploiting linguistic information from transcripts. In *ACL*.

[44] Jun Fu, Wei Zhou, Qiuping Jiang, Hantao Liu, and Guangtao Zhai. 2024. Vision-language consistency guided multi-modal prompt learning for blind ai generated image quality assessment. *IEEE Signal Processing Letters*, 31, 1820–1824.

[45] Yixiao Li, Xiaoyuan Yang, Yuqing Luo, Hadi Amirpour, Hantao Liu, and Wei Zhou. 2025. Unlocking implicit motion for evaluating image complexity. *Displays*, 103131.

[46] Youxiang Zhu, Nana Lin, Xiaohui Liang, John A Batsis, Robert M Roth, and Brian MacWhinney. 2023. Evaluating picture description speech for dementia detection using image-text alignment. *arXiv preprint arXiv:2308.07933*.

[47] Catarina Botelho, John Mendonça, Anna Pompili, Tanja Schultz, Alberto Abad, and Isabel Trancoso. 2024. Macro-descriptors for alzheimer's disease detection using large language models. In *INTERSPEECH*.

[48] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. (2022).

[49] OpenAI. 2024. GPT-4 technical report. *arXiv:2303.08774*.

[50] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.