VS-LLM: Visual-Semantic Depression Assessment based on LLM for Drawing Projection Test

Meiqi Wu[1], Yaxuan Kang[2], Xuchen Li[2], Shiyu Hu[2,3], Xiaotang Chen[2], Yunfeng Kang[2], Weiqiang Wang[1], and Kaiqi Huang[2,3,4]

[1] School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[4] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China
wumeiqi18@mails.ucas.ac.cn, {yaxuan.kang, lixuchen2024}@ia.ac.cn,
hushiyu199510@gmail.com, xtchen@nlpr.ia.ac.cn, yunfeng.kang@ia.ac.cn
wqwang@ucas.ac.cn, kqhuang@nlpr.ia.ac.cn

Abstract. The Drawing Projection Test (DPT) is an essential tool in art therapy, allowing psychologists to assess participants' mental states through their sketches. Specifically, through sketches with the theme of "a person picking an apple from a tree (PPAT)", it can be revealed whether the participants are in mental states such as depression. Compared with scales, the DPT can enrich psychologists' understanding of an individual's mental state. However, the interpretation of the PPAT is laborious and depends on the experience of the psychologists. To address this issue, we propose an effective identification method to support psychologists in conducting a large-scale automatic DPT. Unlike traditional sketch recognition, DPT more focus on the overall evaluation of the sketches, such as color usage and space utilization. Moreover, PPAT imposes a time limit and prohibits verbal reminders, resulting in low drawing accuracy and a lack of detailed depiction. To address these challenges, we propose the following efforts: (1) Providing an experimental environment for automated analysis of PPAT sketches for depression assessment; (2) Offering a **V**isual-**S**emantic depression assessment based on **LLM** (**VS-LLM**) method; (3) Experimental results demonstrate that our method improves by 17.6% compared to the psychologist assessment method. We anticipate that this work will contribute to the research in mental state assessment based on PPAT sketches' elements recognition. Our datasets and codes are available at: https://github.com/wmeiqi/VS-LLM.

Keywords: Drawing projection test · Art therapy · LLM · Multimodal depression assessment.
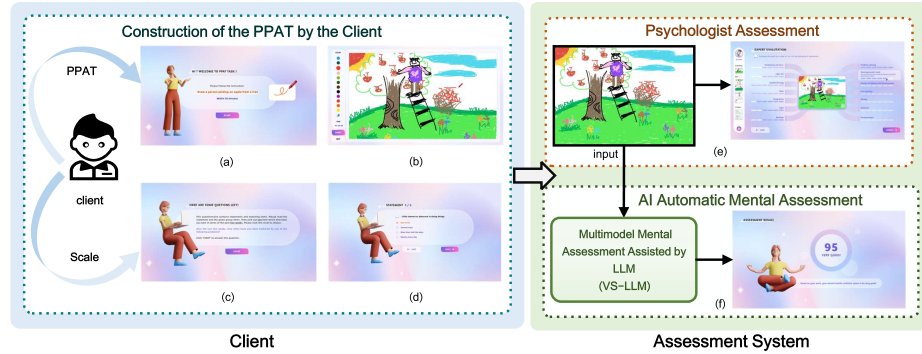
1   Introduction

Fig. 1. System Flowchart. Our mental state assessment on the electronic DPT system consists of two parts. The first part involves the participant, who is asked to participate in the PPAT task and finish the scale. **(a) (b)** represents the user interface of the drawing flow of the PPAT task, while **(c) (d)** depicts the assessment flow of the participant's depression scale. The second part is the mental state assessment system. **(e)** illustrates an expert-oriented evaluation system to collect psychologists' reviews of PPAT images from 14 dimensions. **(f)** demonstrates the automated assessment flow of VS-LLM which serves as the multi-modal psychological assessment algorithm proposed by us.

In computer vision, the sketch can be regarded as an expression of the human brain's internal representation of the visual world [1]. Moreover, in the field of psychology, sketches of the Drawing Projection Test (DPT) play a crucial role in evaluating mental disorders [2]. Compared with self-reported scale [45], it has been widely proven that the DPT can enrich psychologists' understanding of individuals' mental states, such as depression, anxiety, aggressiveness, and their normal states. However, despite its potential benefits, DPT has not seen widespread adoption in clinical practice. This raises the question: *Why hasn't DPT been more widely utilized in psychological assessment and treatment?*

In Fig. 1, psychologists evaluate participants' mental states based on the "a person picking an apple from a tree (PPAT)" task [4], which entails grading along 14 dimensions. This process demands significant time, effort, and expertise. It is hard to deploy evaluation tasks in large-scale settings like schools or companies. Moreover, unlike traditional sketch recognition methods in computer vision [8,9,10,11,12], PPAT sketches present unique challenges due to imposed time limits and the prohibition of verbal reminders, resulting in lower drawing accuracy and less detailed depictions. To address these issues, some AI-based automatic assessment methods [5,6,7] are proposed, they primarily use CNNs to extract visual features from the final sketch. However, these approaches often overlook detailed analysis of the painting process and fail to provide compre-

hensive evaluations, including aspects like color usage and space utilization. In addition, there is still no unified testing environment available for study.

To address those limitations, we have developed an Artificial Intelligence (AI) automatic depression assessment method, Visual-Semantic Depression Assessment Based on Large Language Model (VS-LLM), to support psychologists in conducting a large-scale automatic DPT. In particular, the Visual Perception Module utilizes stroke information to model the painting process by decomposing the original sketch into sequences, enabling a more detailed analysis. The Mental Semantic Caption Generation Module generates mental captions by designing tailored prompts (such as color usage and space utilization) for LLM (specify what LLM stands for if possible). Subsequently, the Semantic Perception Module captures overall information under the guidance of expert knowledge.

To verify the effectiveness of our method, we first constructed a PPAT dataset containing 690 PPAT sketches aligned with the PPAT task in psychology. Finally, the experimental results demonstrate the superior performance of our method, which improves by 17.6% compared to the psychologist assessment method. In summary, our contributions are:

- We developed a visual-semantic depression assessment system based on LLM (**VS-LLM**) method, where the visual perception module and the mental semantic caption generation module are respectively used to obtain more detailed and overall information from the painting, enabling more effective analysis of PPAT.
- We first provided an experimental environment for automated analysis of PPAT sketches for depression assessment.
- Our experiments demonstrate the superior performance of our method, which improves by 17.6% compared to the psychologist assessment method.

2   Related Work

2.1   Drawing Projection Test and PPAT

DPT aims to depict the authentic psychological state of participants by guiding them to draw specific topics that can reflect subconscious or unconscious information that cannot be expressed in language [13]. So far, it has been widely proven the effectiveness of identifying mental state in various clinical settings (e.g. depression [14], stress [15], and child sexual abuse [16]). Psychologists evaluate sketches based on experiences and professional standards, such as the use of color, the detail of elements in the sketch, or the coherence and logic between elements. Silver's work reviews a body of research on the Silver Drawing Test (SDT) and Draw A Story (DAS) art-based assessments, which span 40 years of development [17].

As a kind of DPT method, the PPAT test originated in 1990 and is currently widely used in clinical and non-clinical settings, with revisions made for different cultural backgrounds [18]. For example, in India, participants were asked to draw "a person picking a mango from a tree (PPMT)" and scored by the Formal

Elements Art Therapy Scale (FEATS). It was proven that PPMT can effectively distinguish depression patients [18]. In the "person" and "details and environment" dimensions of the PPAT, Scott et al. [19] discovered that patients diagnosed with depression according to DSM-5 performed significantly differently from the control group. Numerous studies have shown that PPAT can effectively differentiate between subjects with subclinical psychological conditions such as depression, anxiety, and aggression, and those in a normal state [4,20].

While the DPT is effective in assessing mental states, it relies on the expertise of psychologists and can be time-consuming and labor-intensive. Therefore, there is a growing trend towards developing automated detection methods.

2.2   AI Technology for Art Therapy

In recent years, with the rapid development of AI, AI-based art therapy has attracted the attention of many scholars. For instance, Chen et al. [5] proposed a scoring system ranging from 1 to 6 based on clock drawings. By comparing the performance of VGG16 [44], ResNet-152 [42], and DenseNet-121 [43], they illustrated how neural networks can effectively screen individuals for dementia and quantitatively estimate its severity. Similarly, Kim et al. [21] utilized ResNet for classifying mental states and numerical evaluation tasks.

Beyond classification techniques, other researchers have explored classic detection methods. For example, Lee et al. [22] applied deep learning to analyze House-Tree-Person sketches by extensive feature extraction and object detection, which got high accuracy in mental state analysis. Moreover, Kim et al. [23] employed YOLOv5 [24] to effectively evaluate Draw-A-Person-in-the-Rain (DAPR) [25].

Compared with traditional expert assessment methods, AI-based art therapy methods will save more costs and the analysis results will be more authentic. However, they typically applied CNN to extract features from the final sketch, regardless of the details analysis of the painting process and the overall features of sketches. In addition, the sketches they analyzed were different and lacked a unified experimental environment [26] for everyone to study.

2.3   Large Language Model

The rise of Large Language Models (LLMs) signifies the dawn of a transformative period in Artificial Intelligence, fundamentally restructuring the entire domain. For example, GPT-4V [27] marks a significant advancement, allowing users to instruct GPT-4 to analyze image inputs. LLaVA [40] integrates a vision encoder from CLIP [41] with the LLM, enabling it to process visual information alongside language. QWen-VL [31] is a multi-modal large language model that demonstrates remarkable capability in understanding and generating content that seamlessly integrates visual and linguistic elements.

Specifically, the integration of Large Language Models (LLMs) in PPAT presents innovative opportunities to enhance psychological understanding and

analysis. Leveraging extensive text data for training, LLMs demonstrate superior performance in natural language processing tasks, making them as valuable tools for interpreting symbolic content in PPAT sketches. Furthermore, LLMs have the potential to automate certain aspects of PPAT image analysis, streamline workflow, and enable psychologists to focus on complex cases.

In summary, integrating LLMs into PPAT promises to enrich psychological comprehension, offering new pathways for delving into the human psyche.

3   Methods

This section demonstrates how the proposed VS-LLM effectively assesses depression. As depicted in Fig. 2, VS-LLM comprises three modules. The Visual Perception Module extracts detailed visual information by decomposing the sketch into sketch sequences; The Mental Semantic Caption Generation Module generates an overall caption for the Text Encoder to obtain global information on the sketch; The Mental Classification Module integrates visual and semantic features for classification.
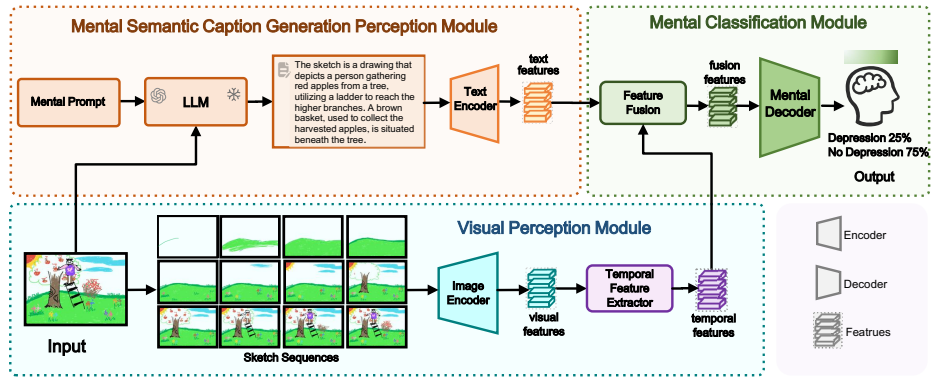

Fig. 2. The architecture of VS-LLM. VS-LLM comprises the Visual Perception Module, Mental Semantic Caption Generation Perception Module, and Mental Classification Module.

3.1   Visual Perception Module

Decomposition of Sketch. When participants draw, they typically adhere to a specific sequence of strokes or components, providing insights into their thought processes. We propose decomposing the sketch into several sequences to capture more detailed information.

According to equations (1) and (2), the resulting sketch drawing is depicted in Fig. 3. Each original sketch is decomposed into 12 consecutive sub-sketches, each containing all the stroke information of the preceding sub-sketches. For the i-th sketch, the division rules for sub-sketches can be formalized as:

$$step_i = \frac{sn_i}{12},$$ (1)

$$sn_{ij} = \begin{cases} j \times 1 & \lfloor step_i \rfloor < 1, j < sn_i \\ sn_i & \lfloor step_i \rfloor < 1, j \geq sn_i \ or \ \lfloor step_i \rfloor \geq 1, j = 12 \ , \\ j \times \lfloor step_i \rfloor & \lfloor step_i \rfloor \geq 1, j < 12 \end{cases}$$ (2)

Where sn_i is the total number of strokes in the i-th original image. $i = 1, 2, 3, ..., n$, n is the number of all sketches. $step_i$ is the stroke step in the i-th original image. $\lfloor step_i \rfloor$ is $step_i$ round down. sn_{ij} is the number of strokes representing the j-th sub-sketch of the i-th sketch, $j = 1, 2, ..., 12$.
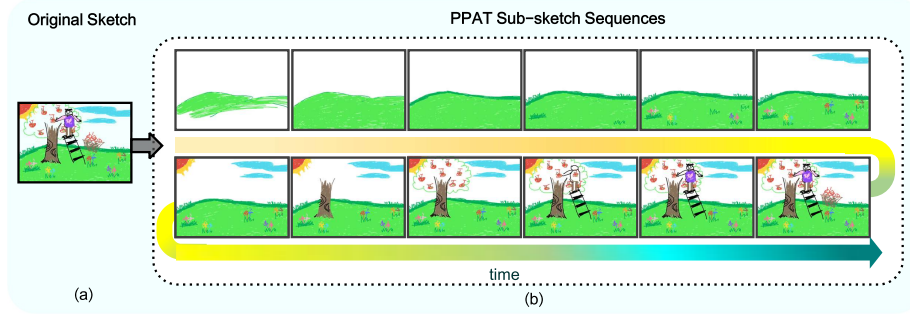

Fig. 3. PPAT Sub-sketch Sequence. (a) represents the complete original sketch, and (b) represents the generated sub-sketch sequence sketch. The specific order of the sub-sketch is arranged from left to right and from top to bottom, forming a sequence of sub-pictures with timing information.

Image Encoder. In the Visual Perception Module, the convolution uses a part of ResNet18 [28], the input is $x \in \mathbb{R}^{W \times H \times C}$, where H,W are the size of x and C is channels, and the output after layer-by-layer convolution pooling (removing full connections) is $(512, 7, 7)$. Finally add a layer of max pooling to become $(512, 3, 3)$ [29]. For sketch x we have:

$$F_v = MaxPooling\left(\mathcal{R}\left(x\right)\right) \in \mathbb{R}^{512 \times 3 \times 3},$$ (3)

where $\mathcal{R}(x)$ represents the process of ResNet processing the input sketch x. F_v represents visual features.

Temporal Feature Extractor. Temporal Feature Extractor consists of two-layer LSTM [30], which captures temporal dependencies and learns semantic knowledge from high-level feature sequences. First, flatten the visual features F_v

into a one-dimensional tensor $(3 \times 3 \times 512 = 4,608)$. Then, the flattened visual features are taken as the input of LSTM. Finally, we add a linear mapping layer.

$$F_t = LSTM\left(Flatten\left(F_v\right)\right) \in \mathbb{R}^{100 \times 12}, \tag{4}$$

where F_t represents temporal stroke features.

3.2   Mental Semantic Caption Generation Perception Module

Mental Semantic Caption Generation. As shown in Fig. 4, to obtain advanced mental semantic captions from PPAT sketches, we designed mental prompts to utilize Qwen-VL [31]. These mental prompts emphasize psychological elements, such as color usage and space utilization, ensuring the quality of text generated by LLM. For PPAT sketches,

$$Caption_{psy} = LLM(P_{psy}, I_{ppat}), \tag{5}$$

where $Caption_{psy}$, P_{psy}, I_{ppat} represent psychological captions, prompts with psychological elements, and PPAT sketches, respectively.
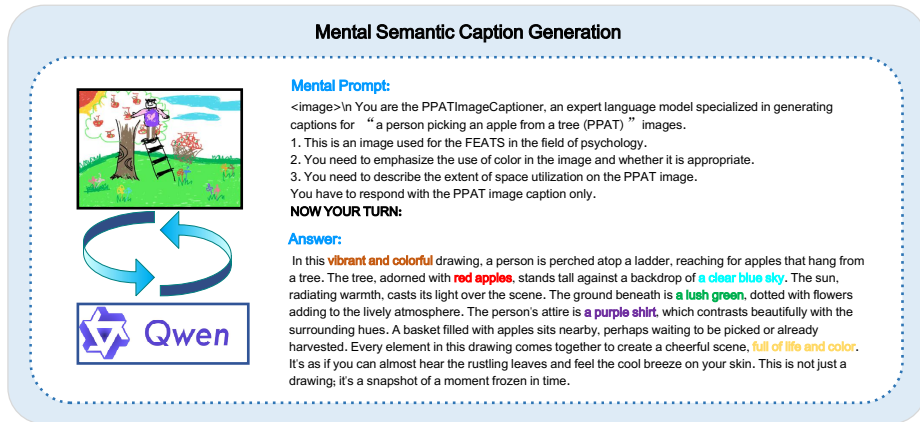

Fig. 4. Mental Semantic Caption Generation Framework. We designed mental prompts and input them into Qwen-VL to obtain the psychological semantic description of the sketch.

Text Encoder. We use RoBERTa [32], which is a transformer encoder model pre-trained on a large corpus of text data. We tokenize input sentences into a series of text tokens and then feed the token sequence into the RoBERTa language model to extract text embedding vectors. In this semantic perception module, the input text $Caption_{psy} \in \mathbb{R}^{W \times C}$, where W is the length of the text sequence and C is the dimension of the features, while the output is the feature representation transformed through various layers.

$$F_t = \text{Pooling}\left(\mathcal{RO}(Caption_{psy})\right) \in \mathbb{R}^{\text{D}}, \tag{6}$$

Where $\mathcal{RO}(Caption_{psy})$ represents the process of RoBERTa model processing the input text $Caption_{psy}$, $Pooling$ represents global pooling operation, and F_t is the final text feature representation with dimension D.

3.3   Mental Classification Module

After extracting semantic and visual information, we concatenate them. Then, we design a mental decoder. Specifically, the mental decoder consists of three consecutive linear layers to further learn multi-modal mental features. The decoder outputs psychological categories.

3.4   Training with Focal Loss

To address the imbalance between positive and negative samples, we utilized Focal Loss during the training process [33]. The abundance of negative samples, constituting the majority of the total loss, often consists of easily classifiable instances. Consequently, the model's optimization direction may not align with our expectations. This function aims to redirect the model's focus towards hard-to-classify samples by attenuating the weight of easy-to-classify ones. Focal Loss introduces a modulating factor $(1 - p_t)^{\gamma}$ to the cross-entropy loss, where $\gamma \geq 0$ is a tunable focusing parameter. We define the focal loss as:

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t). \tag{7}$$

The focal loss is visualized for several values of $\gamma \in [0, 5]$. The focusing parameter γ smoothly adjusts the rate at which easy examples are down-weighted. When $\gamma = 0$, Focal Loss is equivalent to Cross-Entropy Loss (CE), and as γ increases, the effect of the modulating factor becomes more pronounced.

4   Dataset Contruction

To establish an experimental environment for the automated analysis of PPAT sketches, we constructed a PPAT dataset. This section begins by discussing the collection and annotation process of the dataset, followed by the presentation of statistics and division of the dataset.

4.1   Dataset Collection and Annotation

Dataset Collection Process. As depicted in Fig. 1, participants were invited to participate in the research project. The software operation process involved providing basic personal information and completing sketch tasks, which consisted of two parts: sketch adaptation exercises and formal sketch tasks. Additionally, participants were required to complete the Patient Health Questionnaire (PHQ-9) scale [45], which is an established diagnostic instrument designed to screen for the presence and measure the severity of depressive symptoms in individuals.

Table 1. PPAT Dataset Statics.

Metric	Type	Value
Age	0-20	123/690
	20-40	531/690
	40-60	32/690
	>60	4/690
Gender	Male	302/690
	Female	388/690
	Neutral	0/690
Mental State	Depression	117/690
	No Depression	573/690

Dataset Annotation. The sketch was evaluated by three psychologists according to FEATS, and the average score was calculated. Subsequently, each sketch obtained 14 scores, corresponding to the 14 dimensions in FEATS, thus forming a 14-dimensional vector of real numbers. The ground truth is the results of the PHQ-9 scale.

Checkout Flow. We have implemented a stringent data review process to uphold the dataset's quality. Trained professional participants understand the nuances of the PPAT task. Subsequently, the psychologist conducted a second-round review of the collected data. Finally, authors and psychologists make the ultimate judgment of whether to accept or reject the data in the third-round confirmation. Any rejection during self-check, verification, or data acceptance necessitates recollection. We believe this verification mechanism ensures the generation of a high-quality dataset.

4.2 Dataset Statistics and Division

We have used a threshold of 10 points for PHQ-9 scores and the statistics of depression and non-depression are provided in Table 1. In addition, we also collected the gender and age information for the dataset. As shown in Table 2, we split the train dataset and test dataset by using 5-fold cross-validation for conducting experiments.

5 Experiment

In this section, we first introduce the experimental implementation details. Secondly, we conduct experiments to compare the performance of the psychologist assessment and the AI automatic assessment. Thirdly, we demonstrate the effectiveness of different modules and Focal Loss through ablation studies.

Table 2. The division of PPAT Dataset.

	Depression	No Depression	Total
Train	88	465	553
Test	29	108	137
Total	117	573	690

5.1   Implementation Details

The VS-LLM is constructed by the PyTorch[34], and the Focal Loss is implemented according to [33]. The network is optimized with the Adam [35] by setting the initial learning rate to 0.001 and the batch size to 8. All the experiments are performed on a workstation with an NVIDIA GTX Titan X 12G GPU.

Table 3. Experimental Results of Psychologist Assessment and AI-Automated Assessment on PPAT Dataset. The input for the psychologist assessment method consists of scores from the 14 dimensions of FEATS. The input for the AI automated assessment method is the PPAT sketch. Note that we only calculate the number of trainable parameters of the model.

Method		Acc(%)	Params(M)	FLOPs
Psychologist Assessment	Random Forest [36]	**70.2**	-	-
	SVM [37]	57.3	-	-
	Logistic Regression [38]	56.4	-	-
	MLP [39]	64.3	-	-
AI Automatic Assessment	Resnet18 [42]	83.3	11.18	4.11
	Sketch-a-net [9]	85.7	8.41	1.51
	VS-LLM (Ours)	**87.8**	8.87	1.51

5.2   The Performance of Psychologist Assessment

When psychologists psychoanalyze participants based on sketches, they quantitatively analyze the sketches' scores on 14 global variables (color prominence, appropriate color, effort, space use, integration, authenticity, logic, problem-solving, development level, detail and environment, line quality, people, rotation, continuous repetition). These variables are scored on a scale of 0 to 5.

Based on psychologist scores, we compared the use of Random Forest [36], SVM [37], Logistic Regression [38], and MLP [39] to analyze the depression of the visitors. The experimental results are given in the Psychologist Assessment Method of Table 3. Among those methods, the random forest approach shows the best results for psychologist ratings, with only 70.2% prediction accuracy on the test.

5.3   The Performance of AI Automatic Assessment

Regarding AI automated assessment, we conducted experiments, as shown in Table 3, titled "AI Automatic Assessment". Compared to the psychologist assessment, AI automated assessment outperform those of psychologists. We attribute this result to two main factors: firstly, psychologist ratings are heavily influenced by the skills and cognitive levels, and each individual may have biases in their interpretation of sketches. On the other hand, the psychologist assessment relies on scores across 14 dimensions as input, which is relatively simplistic, whereas AI automated assessment utilizes sketches as input, allowing for the extraction of richer visual features. This greatly aids in the assessment of depression.

In AI automated assessment methods, our VS-LLM method have performed the best on the PPAT dataset, achieving an accuracy of 87.8%. This success stems from the integration of psychological semantic descriptions, assisted by LLM, into the VS-LLM framework. The visualization of recognition results is shown in Fig. 5, revealing that in examples of depression, the colors appear dull and there is low utilization of space.

Compared with the best Psychologist Assessment method, our method improves by 17.6%, indicating that our method surpasses the level determined by psychologist ratings. This suggests the potential for automated depression assessment.
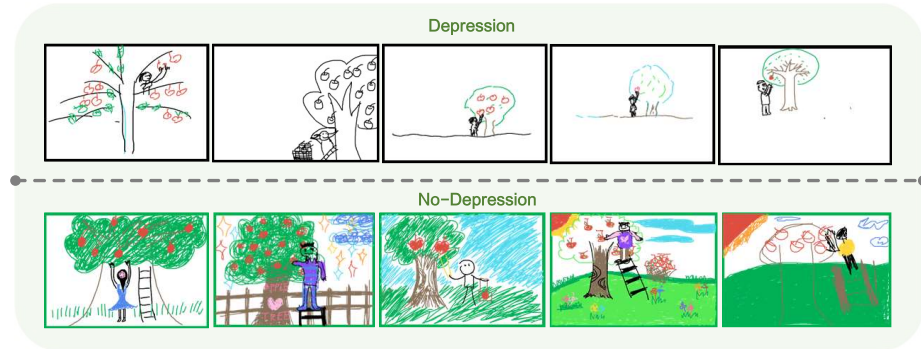

Fig. 5. Visualization of Recognition Results. **Top:** the examples of depression; **Bottom:** the examples of no-depression.

5.4   Ablation Study

Table 4 presents the results of ablation studies. Specifically, the experimental results in the first and 5th rows demonstrate the effectiveness of the Mental Semantic Caption Generation Module, with a performance improvement of 3.7% achieved by incorporating psychological semantic descriptions. Comparing the results between the second and 5th rows highlights the effectiveness of the Temporal Feature Extractor, which achieved an improvement of 4.5%. The third

and 5th rows demonstrate the performance of different image encoders, with our method achieving an improvement of 0.6%. Finally, contrasting the 4th and 5th rows, we compared the impact of different loss functions on performance, with focal loss (FL) outperforming cross-entropy loss (CEL) by 1.4%.

Table 4. Ablation Studies on PPAT Dataset. FL is the abbreviation of focal loss and CEL is the abbreviation of cross-entropy loss.

Number	Mental Sematic Caption Generation	Visual Perception Module		Loss	Acc(%)
		Image Encoder	Temporal Feature Extractor		
1	✗	Resnet18	✓	FL	84.1
2	✓	Resnet18	✗	FL	83.3
3	✓	Sketch-a-Net	✓	FL	87.1
4	✓	Resnet18	✓	CEL	86.4
5(Ours)	✓	Resnet18	✓	FL	87.8

6   Conclusion, Discussion, and Future Work

In this paper, We developed a visual-semantic depression assessment based on LLM (**VS-LLM**), where the visual perception module and the mental semantic caption generation module are respectively used to obtain more detailed and overall information from the sketch, enabling more effective analysis of PPAT. We provided an experimental environment for automated analysis of PPAT sketches for depression assessment. Our experiments demonstrate the superior performance of our method and confirm the importance of incorporating mental descriptions assisted by LLM.

We have outlined the main challenges and provided insightful perspectives on automated psychological analysis. Moving forward, we will further explore anxiety assessment and other related aspects. Moreover, there are numerous promising opportunities for creating a platform that enhances research and assessment for both researchers and patients.

Acknowledgements

References

1. Peng Xu, Timothy M. Hospedales, Qiyue Yin, Yi-Zhe Song, Tao Xiang, and Liang Wang. Deep learning for free-hand sketch: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):285–312, 2023.

2. Pui-Kwan Au, Calvin Kai-Ching Yu, and Siu-Sing Wong. The application of the formal elements art therapy scale in the person-in-the-rain drawing test. *Art Therapy*, 2024.

3. Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613, 2001.

4. Linda Gantt and Carmello Tabone. *The formal elements art therapy scale: The rating manual*. Gargoyle Press Morgantown, WV, 1998.

5. Shuqing Chen, Daniel Stromer, Harb Alnasser Alabdalrahim, Stefan Schwab, Markus Weih, and Andreas K. Maier. Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports*, 10, 2020.

6. Ochilbek Rakhmanov, Nwojo Nnanna Agwu, and Steve Adeshina. Experimentation on hand drawn sketches by children to classify draw-a-person test images in psychology. In Roman Barták and Eric Bell, editors, *Proceedings of the Thirty-Third International Florida Artificial Intelligence Research Society Conference, Originally to be held in North Miami Beach, Florida, USA, May 17-20, 2020*, pages 329–334. AAAI Press, 2020.

7. Woosuk Seo, Joonyoung Jun, Minki Chun, Hyeonhak Jeong, Sungmin Na, Woohyun Cho, Saeri Kim, and Hyunggu Jung. Toward an ai-assisted assessment tool to support online art therapy practices: A pilot study. In *Proceedings of 20th European Conference on Computer-Supported Cooperative Work*. European Society for Socially Embedded Technologies (EUSSET), 2022.

8. Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Transactions on Graphics (TOG)*, 31:1 – 10, 2012.

9. Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net: A deep neural network that beats humans. *International journal of computer vision*, 122(3):411–425, 2017.

10. Jun-Yan He, Xiao Wu, Yu-Gang Jiang, Bo Zhao, and Qiang Peng. Sketch recognition with deep visual-sequential fusion model. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 448–456, 2017.

11. Lei Li, Changqing Zou, Youyi Zheng, Qingkun Su, Hongbo Fu, and Chiew-Lan Tai. Sketch-r2cnn: An attentive network for vector sketch recognition. *arXiv preprint arXiv:1811.08170*, 2018.

12. Raul Fernandez-Fernandez, Juan G. Victores, David Estévez, and Carlos Balaguer. Quick, stat!: A statistical analysis of the quick, draw! dataset. *ArXiv*, abs/1907.06417, 2019.

13. Emanuel F Hammer. The clinical application of projective drawings. 1958.

14. Simeng Gu, Yige Liu, Fei Liang, Rou Feng, Yawen Li, Guorui Liu, Mengdan Gao, Wei Liu, Fushun Wang, and Jason H Huang. Screening depressive disorders with tree-drawing test. *Frontiers in psychology*, 11:1446, 2020.

15. Kate Kravits, Randi McAllister-Black, Marcia Grant, and Christina Kirk. Self-care strategies for nurses: A psycho-educational intervention for stress reduction and the prevention of burnout. *Applied Nursing Research*, 23(3):130–138, 2010.

16. Galit Amir and Rachel Lev-Wiesel. Dissociation as depicted in the traumatic event drawings of child sexual abuse survivors: A preliminary study. *The arts in psychotherapy*, 34(2):114–123, 2007.

17. Rawley A. Silver. Identifying children and adolescents with depression: Review of the stimulus drawing task and draw a story research. *Art Therapy*, 26:174 – 180, 2009.

18. LSS Manickam and V Sajani. Elements of drawing and depression among adolescents in kerala, south india: An explorative study using a person picking a mango from a tree. *SIS Journal of Projective Psychology & Mental Health*, 23(1):41, 2016.
19. David E. Gussak. Effects of art therapy with prison inmates: A follow-up study. *Arts in Psychotherapy*, 33:188–198, 2006.
20. M Kruthi and LSS Manickam. Person picking a mango from a tree: Formal elements, anxiety and aggression among adolescents in conflict with the law. *SIS Journal of Projective Psychology & Mental Health*, 30(2), 2023.
21. Seong-in Kim, Kee-Eung Kim, and Seunghwan Song. Exploring artificial intelligence approach to art therapy assessment: A case study on the classification and the estimation of psychological state based on a drawing. *New Ideas in Psychology*, 73:101074, 2024.
22. Moonyoung Lee, Youngho Kim, and Young-Kuk Kim. Generating psychological analysis tables for children's drawings using deep learning. *Data & Knowledge Engineering*, 149:102266, 2024.
23. Jiwon Kim, Jiwon Kang, Taeeun Kim, Hayeon Song, and Jinyoung Han. Alphadapr: An ai-based explainable expert support system for art therapy. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 19–31, 2023.
24. Glenn R. Jocher, Alex Stoken, Ayushi Chaurasia, Jiří Borovec, NanoCode, TaoXie, Yonghye Kwon, Kalen Michael, Changyu Liu, Jiacong Fang, Abhiram, Laughing, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Jeba Nadar, imyhxy, Lorenzo Mammana, AlexWang, Cristi Fati, Diego Aristizábal Montes, Jan Hájek, Laurentiu Diaconu, Mai Thanh Minh, Marc, albinxavi, Fatih, Oleg, and wanghaoyang. ultralytics/yolov5: v6.0 - yolov5n 'nano' models, roboflow integration, tensorflow export, opencv dnn support. 2021.
25. Leonard Handler and AD Thomas. *Drawings in assessment and psychotherapy*. New York: Routledge/Taylor & Francis Group, 2014.
26. Shiyu Hu, Xin Zhao, and Kaiqi Huang. Sotverse: A user-defined task space of single object tracking. *International Journal of Computer Vision*, 132(3):872–930, October 2023.
27. OpenAI and Josh Achiam. Gpt-4 technical report, 2024.
28. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
29. Jennifer Cardona, Michael Howland, and John Dabiri. Seeing the wind: Visual wind speed prediction with a coupled convolutional and recurrent neural network. *Advances in Neural Information Processing Systems*, 32, 2019.
30. Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2):654–669, 2018.
31. Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
32. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
33. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

34. Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
35. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
36. Steven J Rigatti. Random forest. *Journal of Insurance Medicine*, 47(1):31–39, 2017.
37. Vladimir Cherkassky and Yunqian Ma. Practical selection of svm parameters and noise estimation for svm regression. *Neural networks*, 17(1):113–126, 2004.
38. Raymond E Wright. Logistic regression. 1995.
39. Yi-Chung Hu. Pattern classification by multi-layer perceptron using fuzzy integral-based activation function. *Applied Soft Computing*, 10(3):813–819, 2010.
40. Haotian Liu and Chunyuan Li and Qingyang Wu and Yong Jae Lee Visual Instruction Tuning *arXiv preprint arXiv:2304.08485*, 2023.
41. Alec Radford and Jong Wook Kim and Chris Hallacy and Aditya Ramesh and Gabriel Goh and Sandhini Agarwal and Girish Sastry and Amanda Askell and Pamela Mishkin and Jack Clark and Gretchen Krueger and Ilya Sutskever Learning Transferable Visual Models From Natural Language Supervision *arXiv preprint arXiv:2103.00020*, 2021.
42. Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun Deep Residual Learning for Image Recognition *arXiv preprint arXiv:1512.03385*, 2015.
43. Gao Huang and Zhuang Liu and Laurens van der Maaten and Kilian Q. Weinberger Densely Connected Convolutional Networks *arXiv preprint arXiv:1608.06993*, 2018.
44. Simonyan, Karen and Zisserman, Andrew Very deep convolutional networks for large-scale image recognition *arXiv preprint arXiv:1409.1556*, 2014.
45. Kroenke, Kurt and Spitzer, Robert L and Williams, Janet BW The PHQ-9: validity of a brief depression severity measure *Wiley Online Library*, 16(9):606–613, 2001.