Petroleum Science 22 (2025) 5020–5033

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

ThinGPT: describing sedimentary rock thin section images with a multimodal large language model

Xin Luo [a,b,*], Jian-Meng Sun [a,b], Peng Chi [a,b,c], Ran Zhang [c], Rui-Kang Cui [a,b,c], Xing-Hua Ci [d], Wei Liu [d]

[a] National Key Laboratory of Deep Oil and Gas, China University of Petroleum (East China), Qingdao, 266580, Shandong, China
[b] School of Geosciences, China University of Petroleum (East China), Qingdao, 266580, Shandong, China
[c] Exploration and Development Research Institute of Shengli Oilfield Company, Dongying, 257015, Shandong, China
[d] Geosteering and Logging Research Institute, Sinopec Matrix Corporation, Qingdao, 266003, Shandong, China

ARTICLE INFO

Article history:
Received 2 January 2025
Received in revised form
26 June 2025
Accepted 4 September 2025
Available online 10 September 2025

Edited by Meng-Jiao Zhou

Keywords:
Rock thin section description
Large language model
Contrastive language-image pre-training
Generative Pre-trained

ABSTRACT

Rock thin section description is an essential method for examining lithology, structure, diagenesis, and sedimentary environment, playing a pivotal role in fields such as geology, geophysics, and petroleum exploration. To overcome the challenges of subjectivity, low efficiency, and high expertise requirements in describing rock thin sections, we design a multimodal mapping network, ThinGPT, which aligns the feature spaces of the contrastive language-image pre-training (CLIP) and Generative Pre-trained (GPT-2) through network training. Given the high frequency of keywords and the structured sentence patterns in thin-section descriptions, we introduce a tokenization method tailored for rock thin sections. This approach enhances GPT-2's ability to effectively encode text and produce text feature vectors. We conducted comparative experiments using ThinGPT and other models on common sedimentary rocks. The results demonstrate that ThinGPT exhibits excellent potential in generating thin-section feature descriptions of rocks. Based on the geological expert evaluation criteria proposed in this study, ThinGPT achieved a score of 1.62 on the test set. For model complexity, ThinGPT avoids heavy initial training of large language models (LLMs). This training strategy makes the model lighter and improves the efficiency of rock thin section descriptions. As an innovative application of a LLMs within a lightweight architecture for rock thin section description, ThinGPT has significant implications for intelligent geology, geophysics, and petroleum exploration.

1. Introduction

Rock thin section description is a widely used technique in geology, geophysics and petroleum exploration (Chen et al., 2019; Sotelo et al., 2023). The core principle of this technology involves professionals examining rock thin sections under polarized light to assess their optical properties and crystallographic features (Guo et al., 2023; Chi et al., 2024a,b; Zhang et al., 2024). This examination reveals the mineral composition, structural framework, generation sequence, and the rock's type and genetic traits, culminating in the production of description outcomes for the rock thin sections (Luo et al., 2024). Historically, manual methods were predominant in the description of rock thin sections, yet these approaches were plagued with issues such as high labor costs, low efficiency, and the potential for divergent outcomes due to variations in descriptive expertise (Qi et al., 2021; Li, Li, Liu, et al., 2023; Chi et al., 2024). To make rock thin section description work free from knowledge barriers, there is an urgent need for a high-precision, high-generalization intelligent rock thin section description generation method. Now, the huge progress in computer vision, natural language processing (NLP), and large language model (LLM) fields has made it possible to intelligently describe the rock thin sections.

In recent years, high-parallel computing has further driven the benign development of deep learning in various vertical fields, leading to a series of works that combine rock thin section images

* Corresponding author.
E-mail address: 17853243001@163.com (X. Luo).
Peer review under the responsibility of China University of Petroleum (Beijing).

https://doi.org/10.1016/j.petsci.2025.09.009

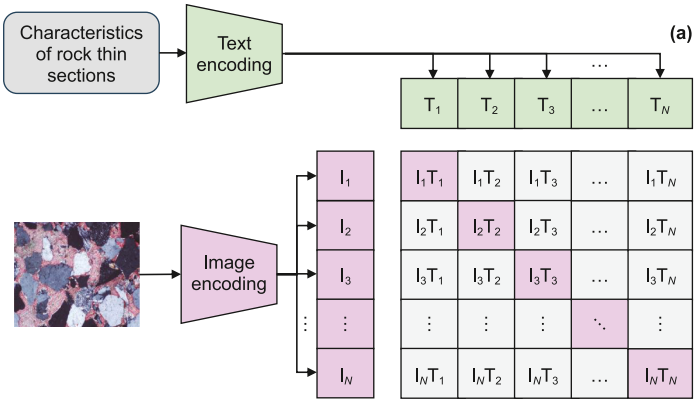
and pattern classification with super-resolution reconstruction, semantic and instance segmentation. Concretely, On rock thin section data, convolutional neural network is used to achieve multi-mineral segmentation with satisfactory results (Saxena et al., 2021). In the identification of thin sections of sedimentary rocks, metamorphic rocks, and igneous rocks, the MaSE-ResNeXt model, which adopts enhancement strategies such as hierarchical filter groups and bottleneck stacking, has effectively improved the identification ability for multiple rock types (Ma et al., 2021). The generative adversarial network model is also used to realize the super-resolution reconstruction of rock thin section, which significantly improves the traditional segmentation and classification results (Liu et al., 2022). Additionally, In order to realize the comprehensive intelligent recognition of rock thin sections, the models of particle segmentation, mineral identification and pore type classification are constructed based on PANet network (Liu et al., 2024). Despite the efforts of practitioners to enhance the accuracy of component recognition in deep learning models on rock thin section data, they have overlooked the exploration of information related to feature description in rock thin section data. It is essential to conduct research on feature description generation in rock thin section. However, traditional supervised models adapted for generating feature descriptions for rock thin section data encounter several issues (Liu and Mukerji, 2022). First, the occurrence of numerous different categories and a scarcity of instances within the same category as labels for rock thin section data feature descriptions, leading to the model's inability to learn semantic information corresponding to each category. Second, low generalization, resulting in new rock thin section data being unable to find corresponding category labels. Finally, a substantial sample size is required to train a model with effective feature description generation capabilities. In addition, retraining many LLMs from scratch for downstream tasks is difficult and unnecessary. LLMs have demonstrated formidable text generation capabilities, achieving notable progress in the fields of medical image diagnosis assistance, video subtitle generation, and product quality inspection (Bhayana et al., 2023; Wang, Wang, et al., 2024; Wu et al., 2024; Yan et al., 2025). Consequently, LLMs may prove effective for describing rock thin sections.

The task of rock thin section description can be classified as image text description for LLMs, which falls under the category of image captioning. Currently, image description generation methods often employ an encoder-decoder approach, where the input image is passed to a visual encoder and a self-regressive language decoder generates the textual description (Radford et al., 2018; Raffel et al., 2020). Image-text description involves both visual and natural language domains, and the effective fusion of multimodal information requires more training costs. One of the main ways to integrate multimodal information is contrastive learning, whose principle is to match the features of images and text for learning. Among them, Contrastive Language-Image Pre-training (CLIP), the most classic contrastive learning model, has become a successful foundation for a general multimodal pre-training LLMs, driving the progress of image description generation research (Radford et al., 2021). However, in specific professional application scenarios within certain fields, the CLIP trained on natural general data will no longer be applicable. While downstream tasks such as image feature description can be addressed through fine-tuning based on the characteristics of the images, a substantial amount of data is still required to effectively train and deploy LLMs. Recently, efforts have been made to reduce the cost of model training, such as ClipCap and I-Tuning (Chen et al., 2020; Mokady et al., 2021). These models utilize pre-trained visual encoders and language decoders. The parameters of the pre-trained networks with effective weights are frozen, and training is focused solely on mapping the image-text latent space. This approach significantly reduces the number of trainable parameters and shortens the training time. However, it may result in a lack of adaptability to professional data in the model's output. Therefore, this paper takes into account the fact that the learning cost required for image information in multimodal fields is much greater than that for text information, and also considers the superiority of generative pre-trained models in the field of NLP (Radford et al., 2019). We propose a multimodal fusion approach for LLMs, which employs a training strategy where the image domain is frozen and the text domain is fine-tuned. This ensures that the LLM generates results in line with rock thin section description style, while also lowering the threshold for model training. Our approach provides a more efficient and convenient technical solution for the laborious and highly experience-dependent manual rock thin section description work.

2. Methodology

2.1. Thin section image encoding of CLIP

CLIP is a leading foundational multimodal pre-training model widely applied in various visual domains. The model encompasses both visual and textual modalities, as illustrated in Fig. 1(a). Guided by unsupervised contrastive loss, the CLIP has been trained on over 400 million image-text pairs, thereby establishing a rich semantic latent space for shared visual and textual data (Gupta et al., 2023; Ramos et al., 2023). The details of the CLIP framework are shown in Fig. 1(a), which consists of an image encoder



Generation effect of the CLIP	Human identification results
A rock wall with a variety of colors	The result of identification is that the mud is thin film, fibrous, etc., and some of it has been sericite, and the distribution is relatively uniform. The silica is distributed around the particles. Occasionally, quartz wavy extinction intensity and apatite-containing inclusions, feldspar with clayificationor sericite. Rare zircon, garnet, etc. The long axis of the detrital is approximately directional.
A blue rock surrounded by black and blue stones and earth	

Fig. 1. The CLIP model and its generated effects.

and a text encoder. The image encoder uses the visual Transformer (VIT) encoding, which divides the image into blocks and forms a sequence, inputs the sequence into the Transformer structure, and obtains the image feature vector (Liu, Lin, et al., 2021; Saad et al., 2022). The text encoder, based on the Transformer architecture, vectorizes the text and encodes it to generate a text feature vector. By computing the cosine similarity between the image feature vector and the text feature vector, it achieves alignment between images and texts. The abundance of high-quality data has enabled the CLIP to demonstrate superior performance in computer vision tasks that require understanding auxiliary text. However, the CLIP trained on natural universal image-text pairs cannot be directly applied to thin section image description tasks. To address the problem of adapting CLIP to the task of identifying rock slabs, we tried three options: retraining the pre-trained model from scratch, freezing some weights, and completely freezing all weights. However, the collection of high-quality training samples was extremely difficult, and the limited amount of rock thin section data was far from meeting the data requirement for a fully trained contrastive learning model. Freezing part of the pre-trained model's weights and then fine-tuning them is a commonly used and effective approach to reduce data requirements. However, there are significant differences between the image and text domains of the CLIP and the rock thin section images as well as their corresponding description results, particularly in the text domain. The text domain of the CLIP typically consists of simple sentences, usually 1–2 in length, and utilizes “A” as a prompt word for text description. In contrast, rock thin section description texts are generally composed of short sentences describing rock properties, structures, oil content sedimentary environment, and diagenesis without a fixed sentence structure. When fine-tuning the CLIP using rock thin sections image and description text data, it is often challenging to effectively update the weights within the CLIP's text domain. As shown in Fig. 1(b), the generated text has significant differences in content and format from the original text samples. The words “rock”, “black”, “blue”, and “stones” in the generated text indicate that the CLIP has learned the information in the rock thin section image, but it cannot correctly decode the text features of the image features according to the format of the rock thin section description text, mainly because the words in the rock thin section description results are rarely, if ever, present in the training dataset. In contrast to medical imaging, descriptions of rock thin sections contain longer sentences and exhibit a higher frequency of key terms. Moreover, the sequence of key term occurrences is relatively consistent, following a pattern based on lithology, structural type, and minerals. Consequently, it is essential to develop an encoding methodology tailored to rock thin section descriptions to facilitate the creation of a pre-trained text generation model.

2.2. Thin section text encoding of GPT-2

This study enhances the textual expression of rock thin section descriptions by employing a Generative Pre-trained Transformer 2 (GPT-2). The model's architecture comprises 12 decoder layers from the Transformer network. The training regimen includes tokenization, word embedding, encoding, and text preprocessing for rock thin section descriptions before text encoding (Devlin et al., 2019). While the BERT model built into CLIP is effective for text tokenization, it sometimes fails to accurately segment domain-specific phrases. Consequently, this paper adopts a single-character format, using the rock thin section description samples as a corpus to create a single-character list that includes only characters, numbers, and symbols present in the corpus. This approach further minimizes the training parameters and enhances training efficiency. Whether it is natural image or medical image data, the division method based on word groups is reliable, and both CLIP and GPT-2 can encode the text correctly. However, there are too many professional terms in rock thin section descriptions, and word group division errors often occur. Therefore, as shown in Fig. 2(a), this paper adopts a single-character word segmentation scheme. Of course, this scheme will increase the reasoning burden of the model. The solution is to use the rock thin section description samples as a corpus to create an individual character list that includes only the characters, numbers, and symbols present in the corpus. This method further reduces the training parameters and improves the training efficiency.

As a classic autoregressive language model in NLP in recent years, the GPT-2 has rich text generation capabilities and its accuracy has reached the same level as the BERT while its

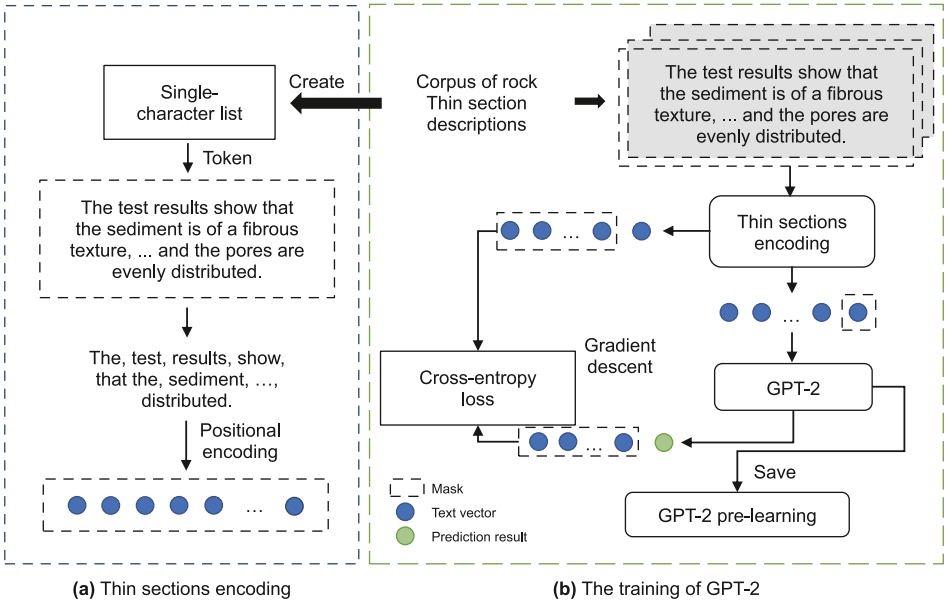


Fig. 2. The thin sections encoding and training of GPT-2.

generalization ability is higher and has excellent zero-shot learning ability (Schneider et al., 2021). The principle of the GPT-2 involves semi-supervised training on a large corpus of natural language text to obtain a universal pre-trained text generation model, followed by supervised fine-tuning for specific subtasks. The main framework of the GPT-2 is based on the decoder part of the Transformer architecture, which enables the model to predict unknown text information, but also increases the demand for model training samples. In the training phase, the GPT-2 employs sequence masking to enhance contextual relationships. The overall training process of GPT-2 is as follows: A segment of text from the training set undergoes tokenization and vectorization. By maximum likelihood function, the model is optimized to produce results that are most likely to be similar to the original text (Shrestha et al., 2024).

Prior studies have demonstrated that the official pre-trained GPT-2 model can achieve remarkable performance in general natural language tasks without additional training or fine-tuning (Bird et al., 2021; Lee et al., 2024). However, the description of rock thin-section features deviates significantly from typical natural language datasets. To enhance the model's generation capabilities for this specific domain, we conducted initial training on GPT-2. As illustrated in Fig. 2(b), text vectors processed through a thin-section encoding layer are masked and subsequently fed into the GPT-2 model, which generates the prediction outputs. The cross-entropy loss function is then computed based on the prediction results and the true labels (the next character in the input sequence). This loss function guides the gradient updates of the model. The optimizer adjusts the model parameters iteratively until the loss converges, ultimately yielding the fine-tuned GPT-2 model.

2.3. Cross-modal fusion between CLIP and GPT-2

The CLIP possesses powerful image-text semantic matching capabilities. However, in the task of rock thin section description, issues such as inconsistent image-text correspondence, monotonous text structure, and weak contextual relationships have been observed in the model's outputs. Considering the GPT-2's rich text generation abilities and drawing inspiration from the ClipCap (Barraco et al., 2022; Li et al., 2024; Yan et al., 2025), this paper proposes a cross-modal fusion network to achieve a unified semantic representation of rock thin section images and description text (Fig. 3). This aims to enhance the textual expression effects of CLIP. As depicted in Fig. 3(d), the cross-modal fusion network effectively bridges the independent latent spaces of CLIP and GPT-2, enabling ThinGPT to encode thin section images using CLIP and to derive image coding vectors. At this point, the image encoding vectors are still not understood by GPT-2. The ThinGPT utilizes a mapping network to transform image encoding vectors into prompt vector space, which is then integrated with the text encoding space in a prefix manner through ordered vector concatenation. It is important to note that the prompt vector space must precede the text encoding space, and this is related to the approach of masking the end of sequences used by GPT-2.

The key to multimodal fusion lies in training a cross-modal fusion network, which converts the image semantic information extracted by CLIP into a text prompt vector and embeds it into the latent space of the GPT-2 network. When the GPT-2 is involved in fine-tuning during the training process, the difficulty of mapping in the cross-modal fusion network is reduced, thereby enabling ThinGPT to easily control the mapping network and non-frozen parts of the GPT-2. ThinGPT adopts a multi-layer perceptron

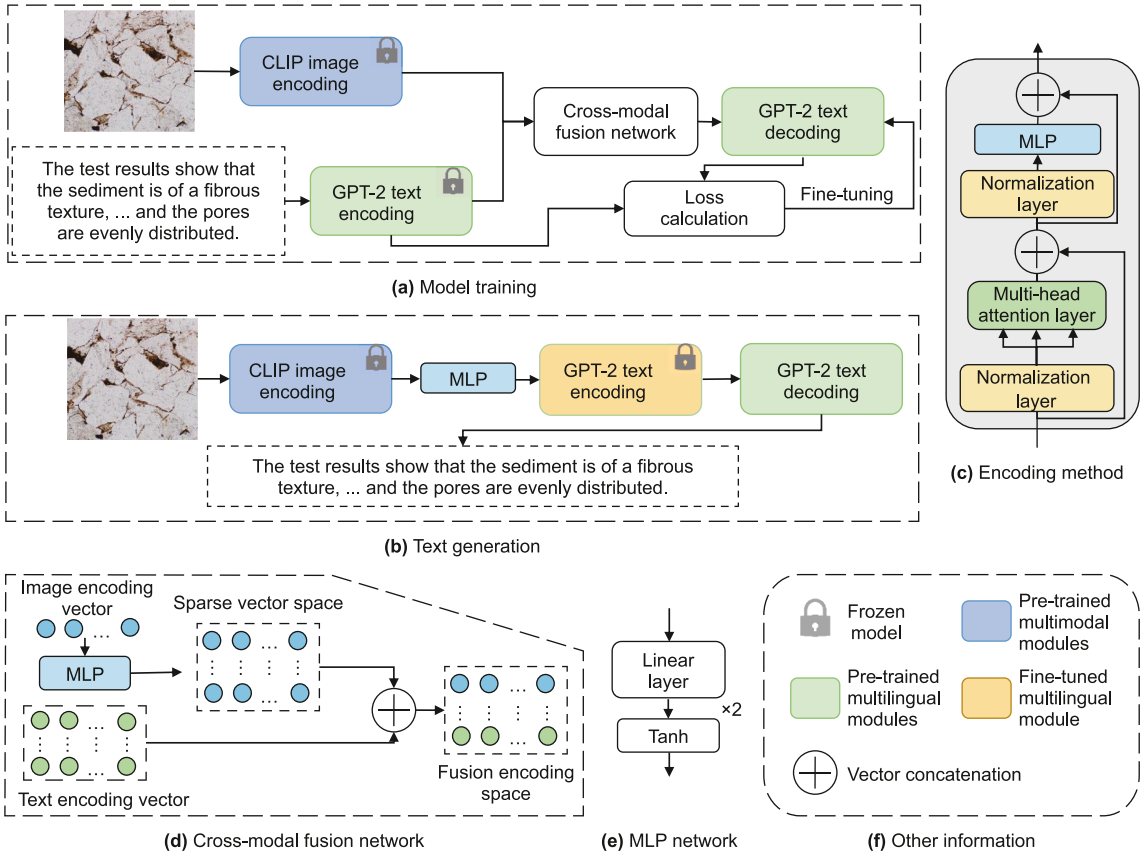


Fig. 3. The ThinGPT network architecture.

(MLP) as a cross-modal fusion network (Tolstikhin et al., 2021; Wang, Liu, et al., 2024), given that the language model is not entirely frozen. During the training process of the ThinGPT model (Fig. 3(a)), the CLIP image encoder remains frozen, the cross-modal fusion network and GPT-2 jointly participate in weight propagation and gradient updates. This enables even a relatively simple design of the mapping network to align text space with image space, thereby allowing ThinGPT to achieve realistic and contextually relevant rock thin section description results.

The ThinGPT architecture demonstrates that the CLIP network's layers are frozen, precluding participation in weight transfer or gradient updates. This approach is justified by observations indicating that fine-tuning CLIP neither enhances result quality nor decreases training time and complexity. Moreover, achieving loss convergence equilibrium while simultaneously fine-tuning two LLMs is inherently challenging. For text generation, ThinGPT utilizes CLIP's image encoding to process thin section images, extracting an image encoding vector (Wu et al., 2023; Gao et al., 2024; Li et al., 2025). This vector undergoes processing through a MLP network (Fig. 3(e)) to match the dimensionality of the fused encoding space. The aligned vector is then fed into the fine-tuned GPT-2 for text decoding, yielding description text results that correspond to the rock thin section images.

2.4. Model evaluation metrics

In the domain of the NLP, the n-gram model is a prevalent general-purpose algorithm for text analysis and processing (Hu et al., 2022). This model operates on the principle of statistically determining the frequency of consecutive N-word sequences within a given text (Aziz et al., 2023; Feng et al., 2024). It is frequently applied in tasks including text generation, machine translation, speech recognition, and image captioning (Brown, 1992). Numerous image captioning evaluation metrics, as well as established machine translation evaluation metrics such as the BLEU series, are predicated on the n-gram model. The definition of n-gram for these metrics is as follows:

$$P(w_i|w_{i-1},\ldots,w_{i-n+1}) = \frac{C(w_{i-n+1},\ldots,w_{i-1},w_i)}{C(w_{i-1},\ldots,w_{i-n+1})} \tag{1}$$

P is the probability of the i-th character when the first $n-1$ characters of the i-th character appear, $C(w_{i-n+1}, \ldots, w_{i-1}, w_i)$ is the number of co-occurrences of the first $n-1$ characters of the i-th character in the corpus in a fixed order, and $C(w_{i-1}, \ldots, w_{i-n+1})$ denotes the number of co-occurrences of the first $n-1$ words in random order.

2.4.1. Bilingual evaluation understudy

The bilingual evaluation understudy (BLEU) uses n-gram matching rules to evaluate translation quality, with the principle of comparing the degree of matching between the generated text and the original text based on the N consecutive words (Papineni et al., 2002). The higher the order of the n-gram, the more it reflects the fluency and contextual coherence of the production effect, and BLEU is defined as follows:

$$\mathrm{BLEU} = \mathrm{BP} \times \exp\left(\sum_{n=1}^{N} W_n \log\left(P_n\right)\right) \tag{2}$$

BP is the penalty factor, and P_n represents the conditional probability of n-grams. W_n is used to assign weights to different orders of n-grams. In this study, BLEU@4 is employed to quantitatively evaluate the generation performance of the ThinGPT model. Specifically, when N equals 4, a higher weight is assigned to W_n in order to emphasize the fluency of generated text. BLEU@4 yields a percentage value; scores between 30 and 40 indicate good generation quality, while scores between 40 and 50 signify high-quality results. Scores exceeding 50 suggest that the generation performance closely approximates that of the original text.

2.4.2. Consensus-based image description evaluation (CIDEr)

CIDEr is an effective evaluation metric for image descriptions that, in contrast to the BLEU series, aligns more closely with human understanding when assessing the similarity of generated sentences (Vedantam et al., 2015). The principle value of CIDEr involves segmenting a complete sentence into multiple word groups based on length n, converting these word groups into vector form using term frequency-inverse document frequency (TF-IDF), and subsequently calculating the cosine similarity between the TF-IDF vectors of the original sample and the generated sample. The definition of TF-IDF is as follows:

$$g_k(c_i) = \mathrm{TF}(k) \times \mathrm{IDF}(k) \tag{3}$$

Term frequency (TF) is the number of times a specific phrase appears in a sentence divided by the length of the sentence, while inverse document frequency (TF-IDF) is the total number of sentences in the corpus divided by the number of sentences that contain the specified phrase. The cosine similarity is calculated as follows:

$$\mathrm{CIDEr}_n\left(c_i, s_j\right) = \sum \frac{g^n(c_i) \times g^n\left(s_j\right)}{\|g^n(c_i)\| \times \|g^n\left(s_j\right)\|} \tag{4}$$

2.4.3. Geological expert evaluation

BLEU and CIDEr are widely recognized evaluation metrics in the domains of machine translation and image caption generation. Given that descriptions of rock thin sections often contain numerous keywords and adhere to specific sentence structures, we introduce a geological expert evaluation (GEE). This method specifically pertains to the confidence (C) in geological keywords and their accurate position (P) within the text. Initially, the confidence of keywords is defined as follows:

$$C = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\frac{Z(i,j) \cap z(i,j)}{Z(i,j)}\right) \tag{5}$$

Z represents the keywords of the original sample, while z represents the keywords of the generated sample. M is the number of keywords in the i-th sentence, and $\cap$ denotes the overlap between the keywords of the generated sample and the original sample.

$$P = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\frac{\mathrm{Dis}(Z(i,j) \cap z(i,j) - a)}{\mathrm{Dis}(Z(i,j) - b)}\right) \tag{6}$$

P represents the probability that the keyword z appears in the correct position. a is the first word in the generated sample, and b is the first word in the original sample. Dis(.) is the distance between words. The GEE can be defined as

$$\mathrm{GEE} = C + P \tag{7}$$

3. Application and analysis

3.1. Data collection and model training

The application context of ThinGPT is focused on the intelligent description of common sedimentary rock thin sections. This study

compiled thin section images and corresponding description text data for lithological types, including clastic rocks, shale, and carbonate rocks. As shown in Fig. 4(a), the clastic rock thin section data were sourced from the northern block of the Xihu depression in the Xishan Sag, encompassing 14 wells and comprising 2623 pairs of rock thin section description text.

The shale thin section data were collected from Block 8 of the Weixinan Sag, involving 8 wells and totaling 1,386 pairs of rock thin section description text. The carbonate rock thin section data were derived from 13 wells in the Maokou Formation of Zigong City, Sichuan Province, summing to 1649 pairs of rock thin section description text. To ensure consistency between image and text features, we preprocessed the feature descriptions to retain characteristics specific to plane-polarized light images, given that the majority of the collected rock thin-section images were of this type. In the process of language translation, we have verified and corrected all the key terms in the text descriptions of rock thin sections (Ulmer-Scholle et al., 2015).

The feature descriptions for these three rock types primarily encompass composition, structure, minerals, pore characteristics (distribution and filler), clasts (orientation, grain size, and bioclasts), and organic matter (Fig. 4(b)). Respectively, the average lengths of the description texts for clastic rocks, shale, and carbonate rocks are 69.3, 29.6, and 41.2 (Fig. 4(c)). For the thin section description dataset of different rock types, we conducted a complexity analysis of thin-section descriptions under various rock types. Specifically, as shown in Fig. 4(d), we calculated the average number of keywords and the standard deviation of text data length in the descriptions across different rock types.

We performed a statistical analysis on three lithology thin section texts to develop a GPT-2 training corpus, ensuring that ThinGPT's output conforms to the stylistic norms of rock thin section description texts. In the training of corpora, GPT-2 employs a comprehensive training strategy to ensure that all feature descriptions are fully learned. Consequently, the GPT-2 dataset is not partitioned; instead, all text data are allocated to the training set. For accuracy validation, the convergence of the loss function and quantitative metrics are used to evaluate the generated results. The specific hyperparameters for GPT-2 training are detailed in Table 1, and the model's loss value quickly converges during fine-tuning, as depicted in Fig. 5(a). The GPT-2 training loss reaches its minimum at 3450 iterations. Concurrently, the training process takes merely 20 min, indicating GPT-2's high portability and low computational resource requirements. Following training, a GPT-2 model optimized with rock thin section petrographic text is acquired. To assess the GPT-2 pre-trained model's text generation capabilities, we employed prompt phrases like “the identification result is” and “the primary components of the rock are” to

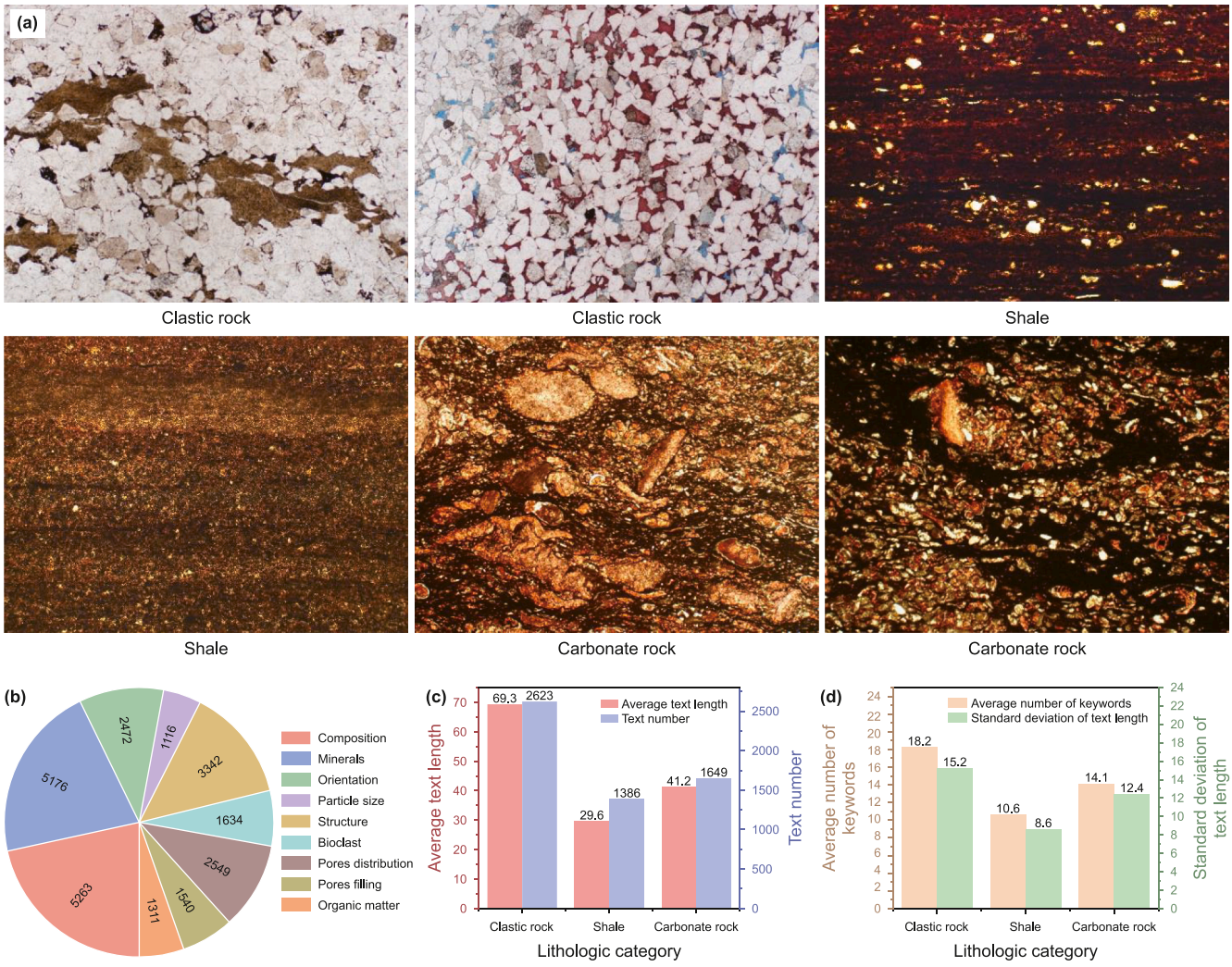


Fig. 4. Rock thin section image and text description dataset.

Table 1
Hyperparameter settings for training of GPT-2.

Hyperparameters	learning rate	Batch samples	Context window size	Layers numbers	Position numbers	Label number	Hidden states	Dimensions	Attention mechanism capabilities
Value	$1.5 \times e^{-4}$	2	1024	10	1024	1	Yes	768	12

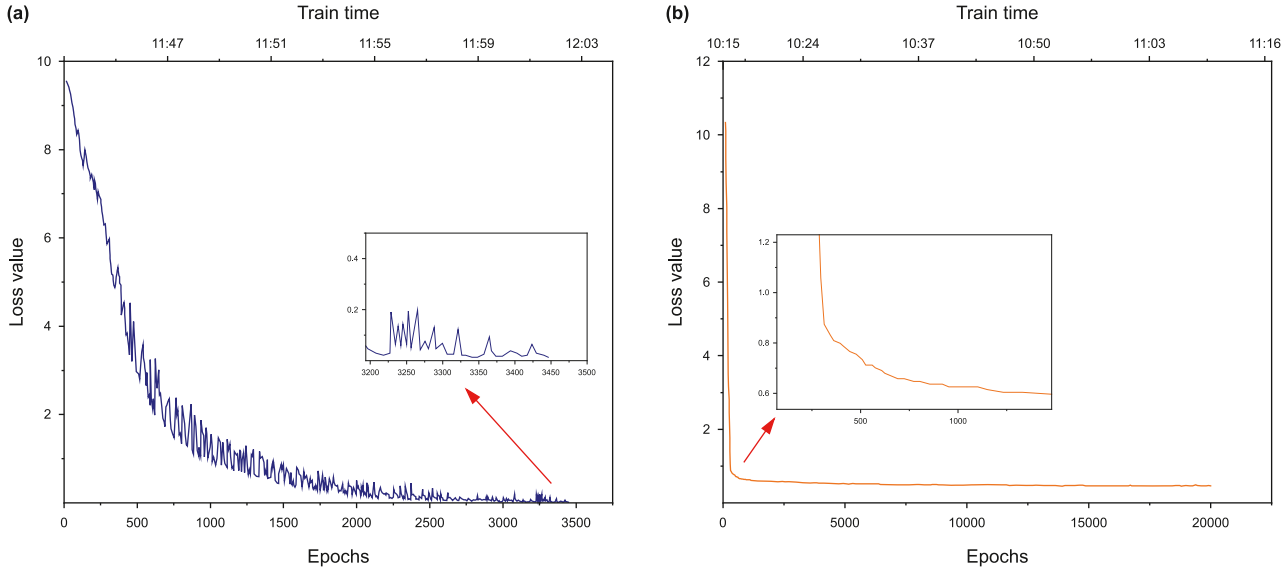


Fig. 5. The training loss and training time of GPT-2 and ThinGPT models.

direct the model's text production. Validation revealed that the GPT-2-generated text closely resembled the style of original rock thin section description texts.

For the implementation of the ThinGPT, we utilized the general CLIP pre-trained model and fine-tuned GPT-2 pre-trained model as inputs. A total of 5658 pairs of rock thin section description text were collected to serve as the dataset, the training set (3395), validation set (1697) and test set (566) are randomly divided in the ratio of 6:3:1. The hyperparameters required for running ThinGPT are listed in Table 2, and the corresponding descriptions are provided in Appendix A. During the training process (Fig. 5(b)), it was observed that the loss value of the ThinGPT rapidly decreased to around 0.7 within 0–1000 iterations. Subsequently, while the loss value continued to decrease, its progress became abnormally slow, indicating that the ThinGPT possesses an advantage in short-term rapid convergence. In terms of training time, our proposed model completed training in a total duration of 57 min. This demonstrates high efficiency compared to other models in the domain of LLMs.

During the generation process, an input image slice is encoded into a text prompt vector and passed through a trained ThinGPT to obtain a series of text feature spaces and corresponding probability distributions. Subsequently, employing TOP-P sampling with a setting of 0.8, the text features with a probability distribution exceeding 0.8 are retained for decoding (Holtzman et al., 2020), resulting in the output of description results corresponding to the rock thin section image. It is important to note that the probability distribution associated with the text requires maximization processing. Specifically, any characters generated by the ThinGPT that do not belong to the rock slice corpus are assigned negative infinity probability, effectively preventing unknown characters from appearing in the final generated result. Furthermore, regarding decoding specifics, the ThinGPT utilizes a BERT tokenizer loaded with pre-trained GPT-2. The decoding process returns textual information based on character feature positions.

The primary network parameters in ThinGPT are based on MLP and GPT-2. MLPs are relatively insensitive to hyperparameters due to their architectural design, whereas the fine-tuning efficacy of GPT-2 is highly sensitive to hyperparameter configurations, particularly the extent of fine-tuning and prefix length. As previously stated, the GPT-2 within ThinGPT is a model fine-tuned on a thin section and text corpus. Thus, during ThinGPT training, all GPT-2 network structures are locked, yet ThinGPT still demonstrates notable performance, with BLEU@4 and CIDEr scores of 15 and 61, respectively. To enhance accuracy, fine-tuning GPT-2 within ThinGPT yields superior text generation outcomes. We set the prefix length to 10 and assess the accuracy of the top linear layer and the even-numbered layers among the 12 multi-head attention layers in GPT-2, as illustrated in Fig. 6(a). Unfreezing the top linear layer significantly boosts ThinGPT's accuracy. Gradual unfreezing of the multi-head attention layers results in a modest improvement in model metrics. When the entire GPT-2 architecture is updated, that is, when GPT-2 is fully fine-tuned, the BLEU@4 and CIDEr scores for ThinGPT-generated outputs rise

Table 2
Hyperparameters for training and generation of ThinGPT.

Hyperparameters	Training processing					Generating processing	
	Learning rate	Batch samples	Prefix length	Maximum length	Dimension of image features	Temperature	Nucleus sampling
Value	$2 \times e^{-5}$	40	10	100	512	1	0.8

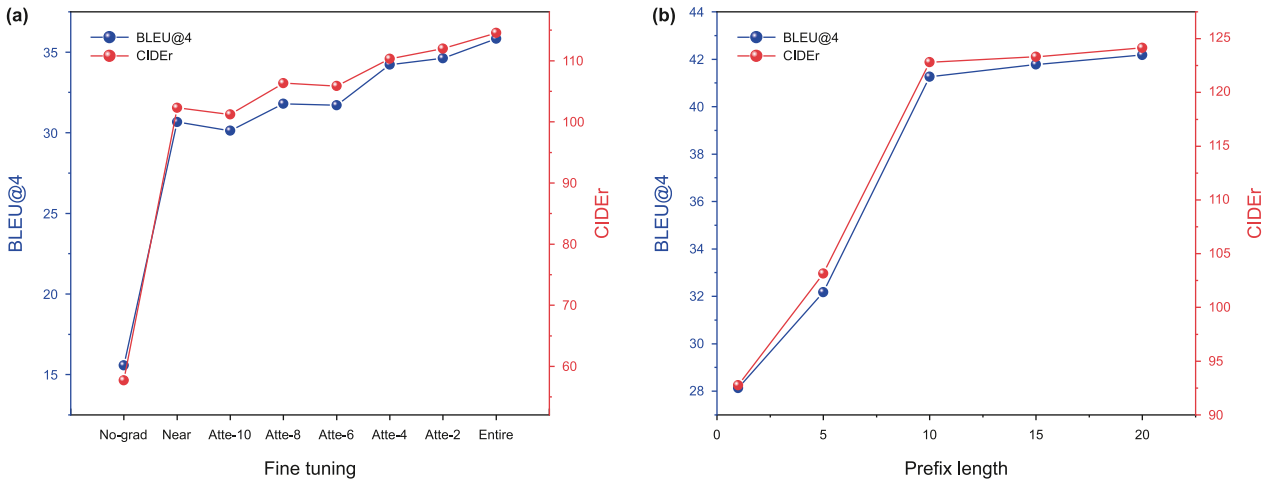


Fig. 6. The impact of fine-tuning degree and prefix length on BLEU@4 and CIDEr. No-grad: GPT-2 does not perform gradient updates; Near: Only updates the gradients of the final linear layer in GPT-2; Atte-X: Updates gradients for the *X*-th and subsequent multi-head attention layers (*X* are 10, 8, 6, 4, and 2, respectively); Entire: The entire network architecture of GPT-2 undergoes weight updates.

to 35.6 and 114.8, respectively. It is worth noting that the deeper the fine-tuning of the GPT-2, the higher the model accuracy, but the model's dependence on computing resources also increases significantly, especially when fine-tuning the multi-head attention layer.

The second important parameter that affects the generation effect of ThinGPT is the prefix length. In ThinGPT, GPT-2 serves as a component for text prediction. The key step is to mask the end of the sequence and leave the unmasked part as the prefix, which has the role of providing information hints for the entire sequence prediction. The longer the prefix length, the more effective information hints, and the better the text generation effect of the model. However, a too-long prefix will also lead to a deep feature dimension for parallel processing, wasting computing resources. As shown in Fig. 6(b), when the prefix length is 10, ThinGPT can show better generation effects.

3.2. Description of rock thin sections under different rock types

To visually demonstrate the rock thin section description effect of the ThinGPT, we used the description of geological expert as a benchmark to label the content in the model's generated output as "Right", "Not quite right", "unknown description", and "Mistake". In this study, ThinGPT generated feature descriptions of rock thin sections for the test set. We randomly selected three results for different rock types for visualization, and the visualization results are shown in Figs. 7–9. These images served as inputs for the pre-trained ThinGPT, which generated corresponding descriptions for

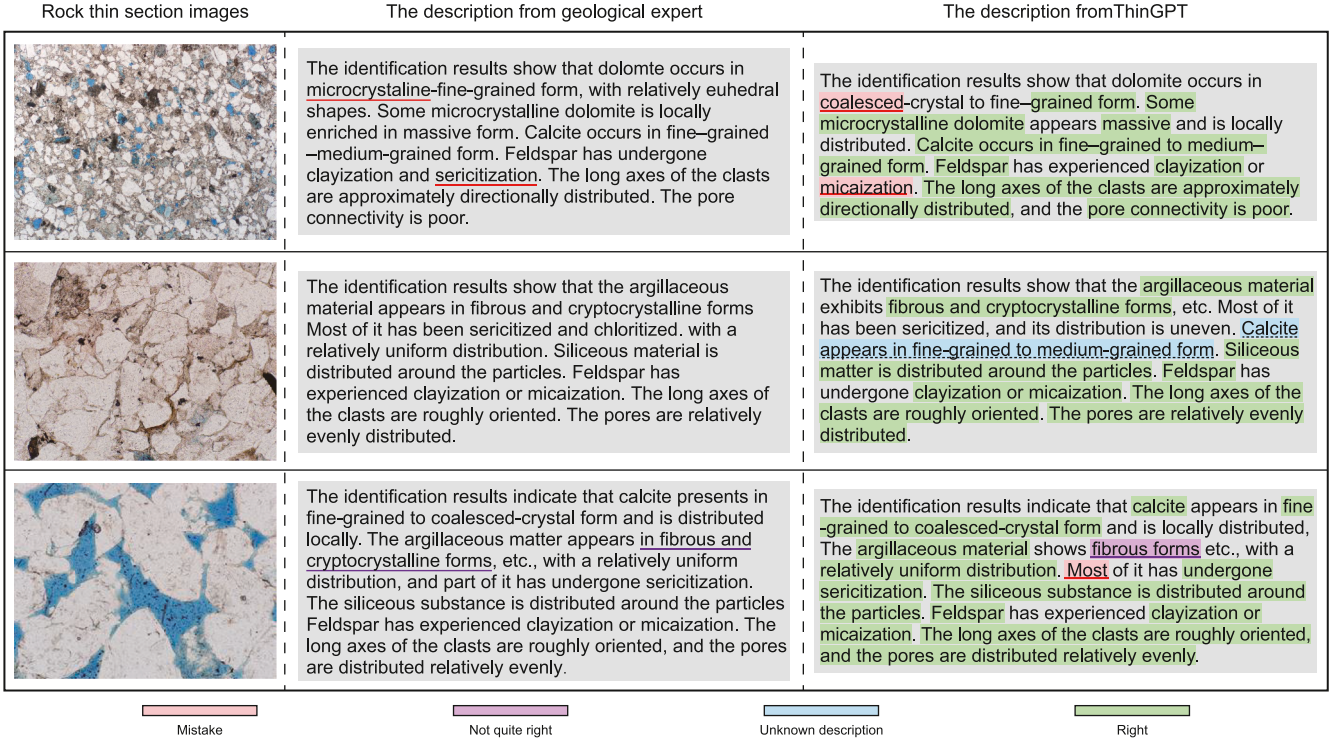


Fig. 7. Comparison of ThinGPT's description effect on clastic rock thin sections.

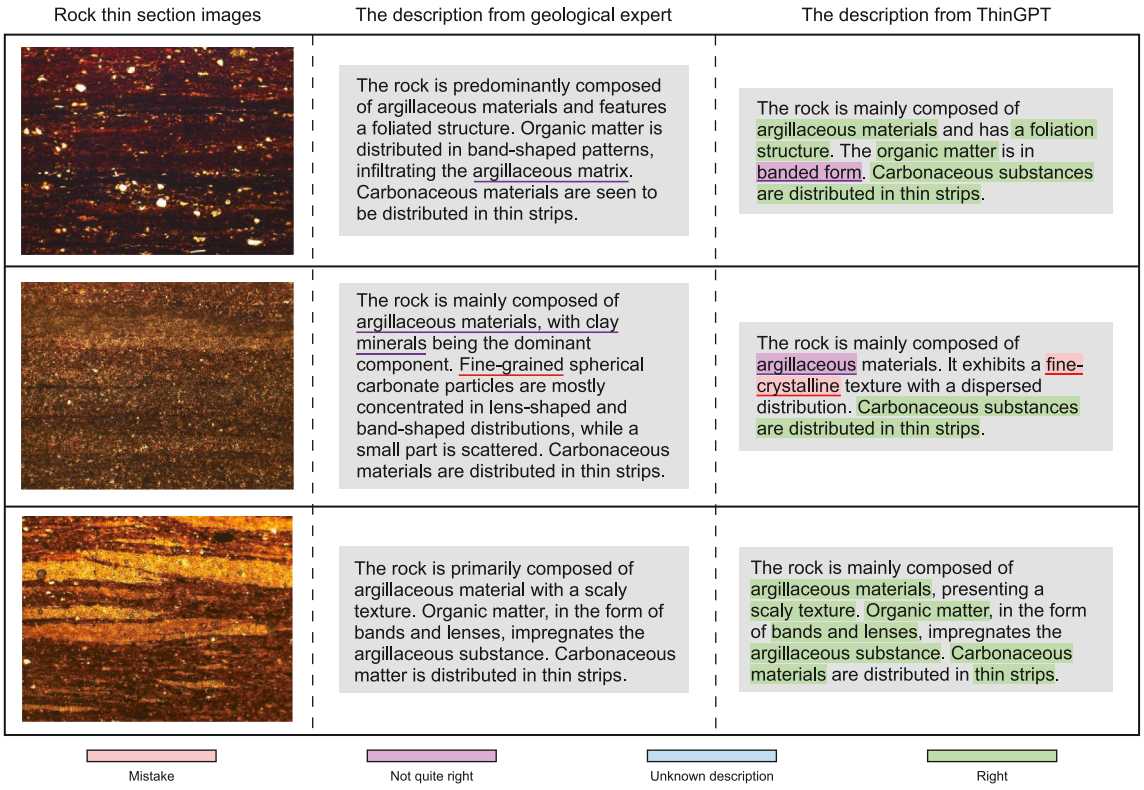


Fig. 8. Comparison of ThinGPT's description effect on shale thin sections.

the thin sections. As shown in Fig. 7, the labels “Right” and “Mistake” are displayed in green and red within the model's generated output, indicating whether they align with human assessment results. The label “Not quite right” is highlighted in purple text, signifying that the output generated by the model is part of the corresponding real description results. It typically occurs with the consecutive appearance of multiple geological keywords within a single sentence, such as mineral composition. In

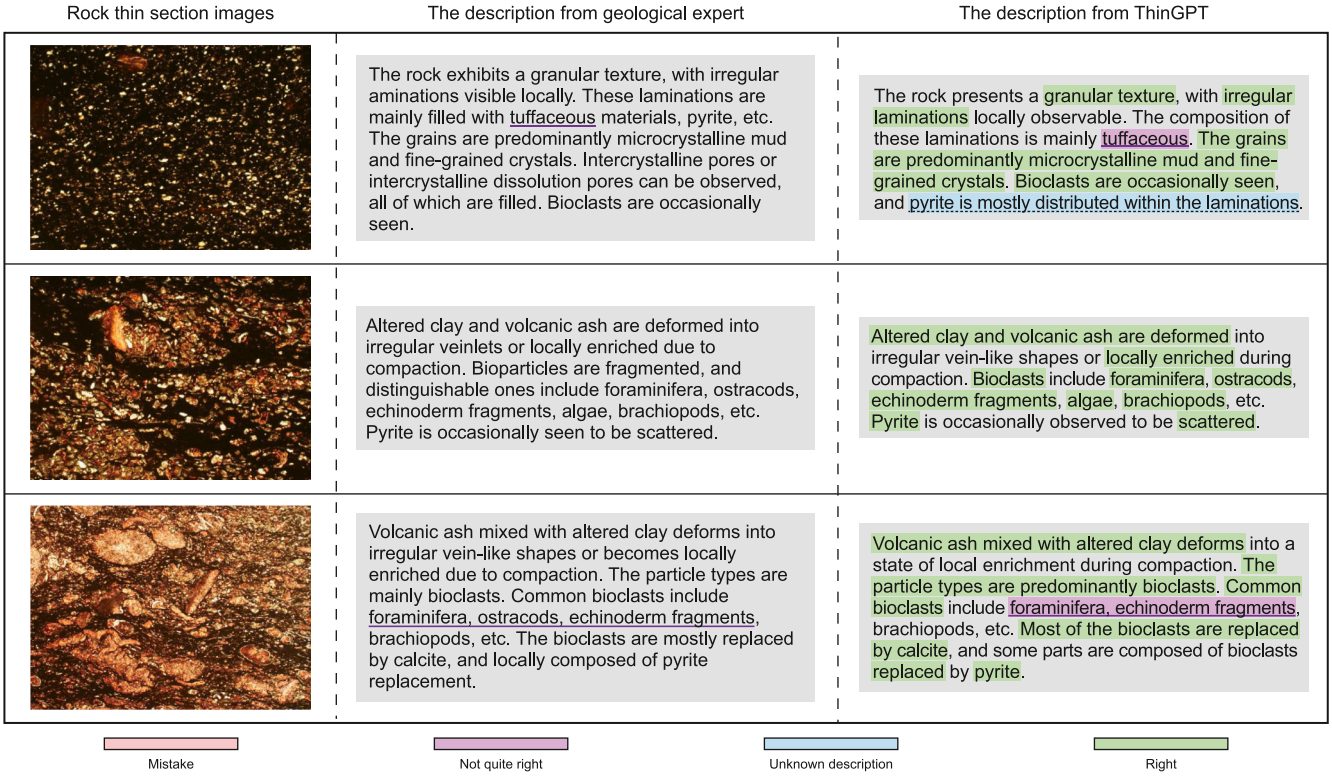


Fig. 9. Comparison of ThinGPT's description effect on carbonate thin sections.

the third row of Fig. 7, the model verification results indicate that the first sentence is “The result of description is that the mud is fibrous, cryptocrystalline”. In comparison to human-identified results, the model lacks descriptions such as “thin film”. Unknown description markers, displayed in light blue, indicate segments in the model-generated results that were not present in the human-identified test set but exist in the full rock thin section description corpus. These instances affect the correctness rate of the model-generated results, positioning them between incorrect and incomplete correctness. Ideally, the model's output should feature more green markers, indicating accuracy, while minimizing red markers. Occasional purple markers are acceptable, and the use of blue markers should be minimal. Regarding the accuracy of sentence phrases, after excluding non-keywords such as “The result of identification is” and “The rock composition is”. Figs. 7–9 show that the ThinGPT model-generated rock thin section description results predominantly feature green markings, with minimal purple, and very few cyan-blue and red markings. This indicates a high level of phrase conformity in the results generated by the ThinGPT model. In terms of sentence structure, the results generated by the ThinGPT are essentially consistent with human-identified results in terms of sentence style. They are generated in a sequence based on rock type, structure, mineral characteristics, grain size distribution, pore distribution and biological debris. Figs. 7–9 visually demonstrate that the sentence style of the ThinGPT's outputs exhibits high consistency with real sample results.

Although training the model on separate lithologies could potentially increase the accuracy of text generation, this experiment pooled rock thin section data from clastic rocks, shale, and carbonate rocks to train a multimodal LLM focused on common sedimentary rocks, due to constraints in sample availability and to optimize the use of parallel computing resources. In terms of model inference, we selected 150 thin-section descriptions for each rock type from the test set, totaling 450 descriptions, and calculated the frequency of “Right”, “Not quite right”, “Unknown description”, and “Mistake”. As shown in Fig. 10, the outputs generated by ThinGPT were generally satisfactory.

Regarding sample complexity, ThinGPT showed no significant performance variation across different lithologies of the test samples. However, the model's outputs for some shale thin sections were notably simpler (Fig. 8), in contrast to the highly complex and detailed results for other lithologies. Further examination of these simpler results for shale sections indicated a correlation with the distribution deviation of samples. Upon thorough examination of shale text data, substantial stylistic variations across different well sections were identified. Given that ThinGPT's corpus is solely derived from rock thin-section descriptions, disparities in sample types and descriptive styles could compromise the model's reliability. Conversely, it can be inferred that some generation errors in ThinGPT are attributable to dataset limitations in quantity and richness, rather than inherent model deficiencies.

3.3. Quantitative index analysis

To evaluate the descriptive performance of ThinGPT quantitatively, we selected 450 generated samples from each of three distinct rock types and performed an assessment that included quantitative metric evaluation, analysis of keyword confidence, frequency statistics of keywords, and distribution statistics of keyword positions. The analysis revealed that ThinGPT achieved BLEU@4 and CIDEr scores of 41.6 and 127.4 for clastic rock thin sections, and scores of 45.2, 134.2 for shale, and 39.8, 132.8 for carbonate rock thin sections (Fig. 11(a)). Additionally, we conducted a quantitative analysis of the frequency, sequence position, and confidence levels of key nouns within ThinGPT's descriptions across all rock types. The analysis showed that larger diameters of green, yellow, and black areas corresponded to a higher occurrence of keywords such as minerals, pores, and mud. The confidence levels of these keywords predominantly ranged from 0.7 to 0.9 (Fig. 11(b)), suggesting that ThinGPT's descriptions are comprehensive. Regarding the position of keywords, descriptions of components like mud and tuff material were found in the initial keywords positions (0–12), while descriptions of

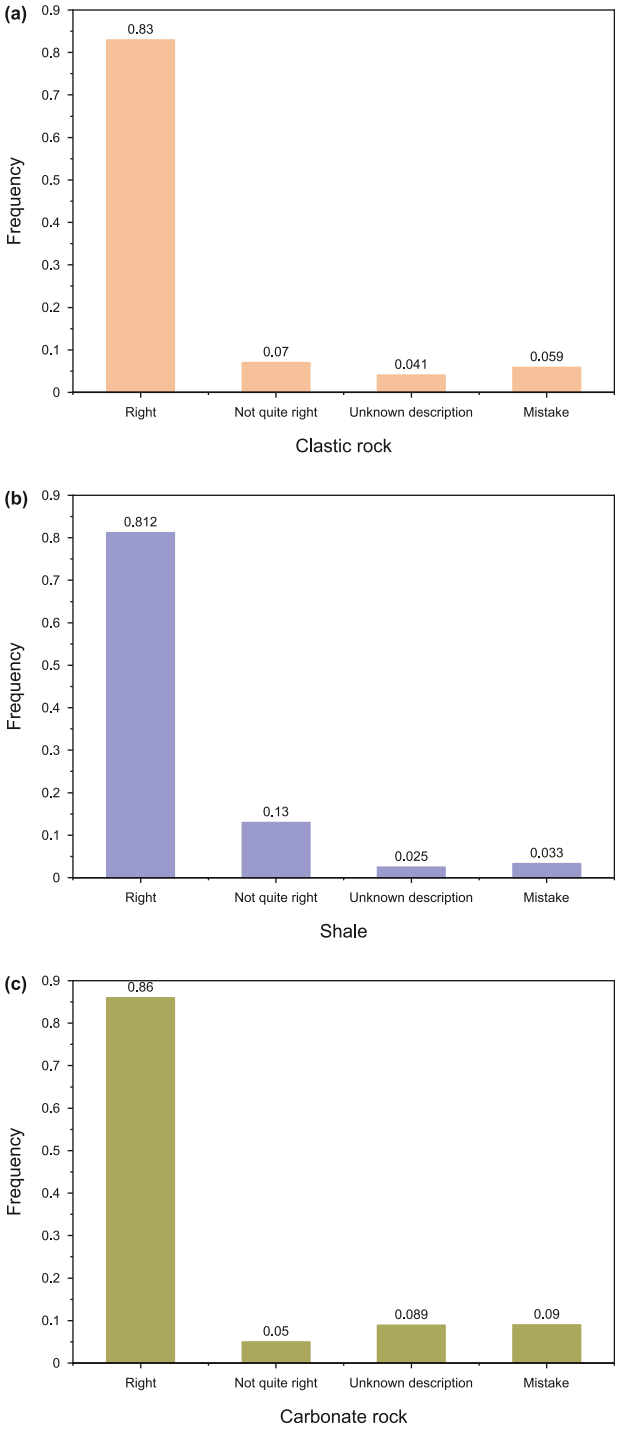


Fig. 10. The effectiveness of ThinGPT-generated samples under different rock types.

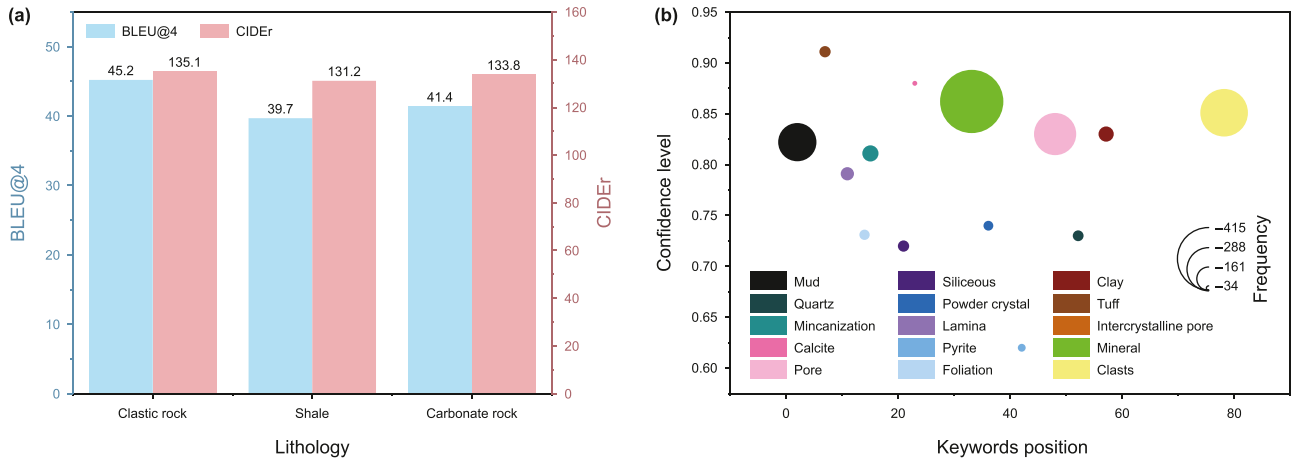


Fig. 11. Quantitative metric results for ThinGPT.

minerals such as quartz and calcite were located in the middle keywords positions (20–40). Pore descriptions were typically found at the end of the sentences, in keywords positions (40–60), and clastic-related descriptions were positioned at the very end, in keywords positions (70–80) (Fig. 11(b)). This distribution suggests that ThinGPT's generated descriptions exhibit logical contextual relationships.

To evaluate the quantitative performance of our model, we conducted comparative experiments with mainstream multi-modal image-to-text models. Regarding the design of comparative experiments, we selected several widely-used benchmark models in the image-to-text field: S&T, SA&T, Caption Transformer (CPTR), ClipCap and BLIP-2. Serving as early yet effective image captioning models, S&T and SA&T (with added attention mechanism) implemented automatic natural language description generation for images through a CNN-LSTM framework (Vinyals et al., 2015; Xu et al., 2016). CPTR became the first model to adopt Transformer for generating textual descriptions of images, possessing global context awareness capabilities (Liu et al., 2021a). Meanwhile, ClipCap and BLIP-2 achieved effective visual-text alignment by training modality mapping networks (Li et al., 2023a). In terms of quantitative metrics, we employed BLEU@4, CIDEr, and GEE to evaluate the quality of text descriptions generated by the models. As depicted in Table 3, our approach significantly outperformed S&T, SA&T, CPTR, ClipCap and BLIP-2. After multiple random divisions of the dataset, ThinGPT achieved average scores of BLEU@4, CIDEr and GEE on the test set of 42.2, 131.5 and 1.62. Consequently, the introduced ThinGPT model exhibits a competitive advantage in the domain of multimodal image-to-text models.

To further demonstrate the practical utility of ThinGPT, we conducted case analysis experiments to compare its geological keyword generation performance with that of other models. For this analysis, one thin-section image was randomly selected from each of three rock types: clastic rock, shale, and carbonate rock. These images were excluded from the training set to ensure unbiased evaluation. Different rock types have distinct geological focuses. For clastic rocks, we examined mineral types, porosity, grain size, and diagenesis. For shales and carbonate rocks, we emphasized stratification and bioclasts. As shown in Fig. 12, traditional image-to-text generation models (S&T and SA&T) struggled to generate accurate keywords related to grain size and diagenesis. The complexity of the keywords directly correlates with the demands placed on the model's text generation capabilities. Models incorporating pretrained text generation, such as ClipCap, BLIP-2, and ThinGPT, demonstrated superior performance. In particular, ThinGPT's unique thin-section encoding method enabled more precise descriptions of rock thin-section features.

4. Discussion

In the field of deep learning, thin section data has traditionally been utilized for lithology classification, porosity segmentation, and multi-mineral segmentation. The emphasis has primarily been on the compositional aspects of this data, with the descriptive features inherent to thin section data often being overlooked. This research introduces ThinGPT, capable of efficiently conducting intelligent description of thin section data and producing detailed, well-structured descriptions. The performance of ThinGPT in generation is significantly enhanced with an increased number of accurate outputs. However, results containing unknown descriptions also hold significant reference value. In the manual description of rock thin sections, analysts typically employ a fixed textual expression to describe prominent features in the data. This approach may lead to certain image features being overlooked. By learning the relationship between image features

Table 3
Comparative analysis of ThinGPT with other models regarding the description of rock thin sections.

Model			
Metrics	BLEU@4	CIDEr	GEE
S&T	21.5	85.3	0.84
SA&T	22.4	88.3	0.91
CPTR	32.6	113.8	1.17
ClipCap	36.1	124.9	1.35
BLIP-2	40.7	121.8	1.46
ThinGPT	**42.2**	**131.5**	**1.62**

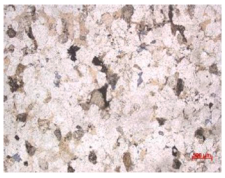
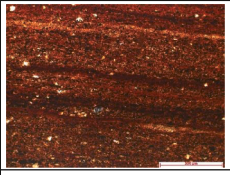
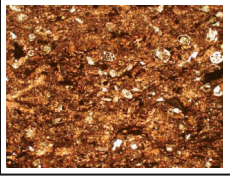
Rock types	Rock thin section images	Feature	Ture result	S&T	SA&T	CPTR	ClipCap	BLIP-2	ThinGPT
Clastic rock		Mineral types	Mainly feldspar	Feldspar	Feldspar	Feldspar and calcite	Feldspar	Feldspar	Mainly feldspar
		Pores	Poorly developed	Poorly developed	Poorly developed	Poorly developed	Poorly developed	Poorly developed	Poorly developed
		Grain size	Moderately sorted	\	\	Moderately sorted	Moderately sorted	Moderately sorted	Moderately sorted
		Diagenesis	Intergranular dissolved pore	\	\	Intergranular pores	Intergranular pores	Intergranular dissolved pore	Intergranular dissolved pore
Shale		Bedding structure	Lamellar	Massive	Lamellar	Lamellar	Lamellar	Lamellar	Lamellar
Carbonate rock		Bioclast	Foraminifera, ostracoda, echinodermata, and algae	Foraminifera and echinodermata	Foraminifera, echinodermata, and algae	Foraminifera, ostracoda, and echinodermata	Foraminifera, ostracoda, echinodermata, and algae	Foraminifera, echinodermata, and algae	Foraminifera, ostracoda, echinodermata, and algae

Fig. 12. Case analysis of ThinGPT with other models.

and textual information, ThinGPT can generate results with unknown descriptions. The emergence of unknown descriptions may provide valuable feedback for improving our initial thin section description work. The ThinGPT exhibits satisfactory performance in generating descriptions for three types of rock thin sections. However, its efficacy in describing shale lithology is comparatively lower. Analysis of the original data revealed a distribution deviation in the shale rock thin section descriptions, adversely impacting ThinGPT's generation capabilities. Future endeavors will focus on increasing the sample size and incorporating data augmentation techniques to enhance the accuracy of thinGPT's descriptions for rock thin sections.

Conversely, although the ThinGPT model delivers effective and reliable outcomes in textual description, its quantitative analysis capabilities fall short of expectations. The text labels of rock thin section descriptions have a small amount of quantitative analysis about the features, which include numerical and symbolic expressions. This situation was predictable. In the context of image quantitative segmentation or classification tasks, numerical symbols are commonly inputted as labels into the model. Typically, single-class labels map to multiple network structures within intricate underlying network spaces. However, in the original dataset of text descriptions, numerical values are merely treated as characters, leading to a substantial loss in their complex, nonlinear representational capacity. Consequently, the model prioritizes sentence structure and text matching, thereby diminishing its focus on the feature vectors associated with quantitative analysis. In subsequent research, enhancing the precision of quantitative descriptions will focus on two primary avenues. Firstly, we will augment the dataset, with a substantial volume of data, numerical types can be utilized as geological keywords. To further enhance the accuracy of ThinGPT, it is essential to focus on improving the quality of the dataset. Specifically, certain feature descriptions should be integrated with additional image types, such as crossed nicol images and high-magnification images. Moreover, ThinGPT should incorporate a multi-image fusion capability in future developments, which involves combining plane-polarized light images with crossed nicol images and high-magnification polarization images. This integration will ensure the completeness and accuracy of the results generated by ThinGPT. Secondly, we intend to integrate language models with image processing algorithms and semantic segmentation techniques, drawing on the methodologies employed in question-answering models.

From a network deployment standpoint, ThinGPT leverages existing models by implementing a straightforward mapping network for training. This strategy primarily entails calibrating the pre-existing semantic comprehension of the pre-trained model to align with the stylistic nuances of the target dataset, rather than acquiring new semantic knowledge. This training paradigm significantly conserves training resources and time. This paper describes the training and inference processes of ThinGPT, which were conducted on hardware with a computational capacity of 8.6 MH/s, eliminating the necessity for high parallel computing capabilities. Regarding the domain adaptability of ThinGPT, if datasets from other fields (such as seismic image-text pairs or medical image-text pairs) also exhibit multi-keyword and highly structured characteristics in textual descriptions (Chen et al., 2020; Si et al., 2024; Kanfar et al., 2025; Pham et al., 2025), then ThinGPT can still achieve the expected performance simply by replacing the dataset.

5. Conclusion

This study introduces a rock thin section intelligent description model (ThinGPT) based on multimodal pre-trained LLMs. It utilizes fine-tuned GPT-2 to replace the text encoding in the CLIP, ensuring consistency between generated results and the style of description text, thereby enriching the model's text expression capability. The ThinGPT is designed with frozen image domain, fine-tuned text domain, and a built-in text-image mapping network to effectively avoid repetitive training of LLMs, saving both training time and computing resources. Utilizing the pre-trained GPT-2 model, ThinGPT achieves full training for the intelligent description of rock thin sections in just a few hours. ThinGPT effectively generates descriptions for clastic rocks, shale, and carbonate rocks, exhibiting high keyword accuracy, suitable keyword distribution, and a coherent overall linguistic style in the text. Quantitative evaluation indicators such as BELU@4, CIDEr and GEE reach 42.6, 125.4 and 12.6. respectively. The introduction of the ThinGPT represents its first application in rock thin section intelligent

description and provides a feasible approach for constructing LLMs in the geology, geophysics, and petroleum exploration field.

CRediT authorship contribution statement

Xin Luo: Writing – original draft, Methodology, Conceptualization. **Jian-Meng Sun:** Funding acquisition. **Peng Chi:** Writing – review & editing, Validation. **Ran Zhang:** Resources. **Rui-Kang Cui:** Visualization. **Xing-Hua Ci:** Resources. **Wei Liu:** Resources.

Codes availability statement

The source codes and pre-training model are available for downloading at the link: https://github.com/luoxinggyyy/ThinGPT.git.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by a grant from the National Natural Science Foundation of China (Grant No. 42174156).

Appendix A

Appendix A describes in detail the hyperparameters that need to be set when ThinGPT and GPT-2 are fine-tuned and the meanings of corresponding hyperparameters (Table A1).

Table A1
Descriptions of hyperparameters.

Hyperparameters	Description
Learning rate	The speed of updating the model's weights
Batch samples	The sample input for each iteration
Context window size	The model takes into account a maximum number of characters when processing text
Layers numbers	The number of layers in the model
Position number	Used to store the position information of the phrases in the sentence
Label number	1 indicates that the model is used for the generation task
Hidden states	Output the hidden states from the past, which are used for the generation task
Dimensions	The dimension of the vector representation of a word group
Attention mechanism capabilities	Number of heads in a multi-head attention mechanism
Prefix length	The prefix length of a sequence
Maximum length	Maximum length limit for sequences
Dimension of image features	The number of channels in the image feature vector extracted by CLIP
Temperature	Used to control the randomness of text generation
Nucleus sampling	Use probability cumulative top-p sampling when generating text

References

Aziz, S., Sarfraz, M.S., Usman, M., et al., 2023. Geo-spatial mapping of hate speech prediction in roman Urdu. Mathematics 11 (4), 969. https://doi.org/10.3390/math11040969.

Barraco, M., Cornia, M., Cascianelli, S., et al., 2022. The unreasonable effectiveness of CLIP features for image captioning: An experimental analysis. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, pp. 4661–4669.. https://doi.org/10.1109/CVPRW56347.2022.00512.

Bhayana, R., Krishna, S., Bleakney, R.R., 2023. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 307 (5), e230582. https://doi.org/10.1148/radiol.230582.

Bird, J.J., Pritchard, M., Fratini, A., et al., 2021. Synthetic biological signals machine-generated by GPT-2 improve the classification of EEG and EMG through data augmentation. IEEE Rob. Autom. Lett. 6 (2), 3498–3504. https://doi.org/10.1109/LRA.2021.3056355.

Brown, P.F., 1992. Class-based n-gram models of natural language. Comput. Linguist. 18 (4), 467–480.

Chen, X., Yao, G., Luo, C., et al., 2019. Capillary pressure curve determination based on a 2-D cross-section analysis via fractal geometry: a bridge between 2-D and 3-D pore structure of porous media. JGR Solid Earth 124 (3), 2352–2367. https://doi.org/10.1029/2018JB016722.

Chen, Z., Song, Y., Chang, T.H., et al., 2020. Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1439–1449. https://doi.org/10.18653/v1/2020.emnlp-main.112.

Chi, P., Sun, J., Yan, W., et al., 2024a. From digital rock to digital wellbore: multi-scale reconstruction and simulation. Adv. Geo-energy Res. 13 (1), 1–6. https://doi.org/10.46690/ager.2024.07.01.

Chi, P., Sun, J., Yan, W., et al., 2024b. Investigating the causes of permeability anisotropy in heterogeneous conglomeratic sandstone using multiscale digital rock. Geophysics 89 (4), MR197–MR207. https://doi.org/10.1190/geo2023-0596.1.

Devlin, J., Chang, M.-W., Lee, K., et al., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805.

Feng, Z., Zhang, X., Zhou, B., et al., 2024. NGST-net: a n-gram based swing transformer network for improving multispectral and hyperspectral image fusion. Int. J. Dig. Earth 17 (1), 2359574. https://doi.org/10.1080/17538947.2024.2359574.

Gao, P., Geng, S., Zhang, R., et al., 2024. CLIP-adapter: better vision-language models with feature adapters. Int. J. Comput. Vis. 132 (2), 581–595. https://doi.org/10.1007/s11263-023-01891-x.

Guo, Z.Q., Qin, X.Y., Liu, C., 2023. Quantitative characterization of tight gas sandstone reservoirs using seismic data via an integrated rock-physics-based framework. Pet. Sci. 20 (6), 3428–3440. https://doi.org/10.1016/j.petsci.2023.09.003.

Gupta, D., Kharbanda, S., Zhou, J., et al., 2023. CLIPTrans: transferring visual knowledge with pre-trained models for multimodal machine translation. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, pp. 2863–2874 https://doi.org/10.1109/ICCV51070.2023.00269.

Holtzman, A., Buys, J., Du, L., et al., 2020. The Curious Case of Neural Text Degeneration. https://doi.org/10.48550/arXiv.1904.09751.

Hu, X., Al-Olimat, H.S., Kersten, J., et al., 2022. GazPNE: annotation-free deep learning for place name extraction from microblogs leveraging gazetteer and synthetic data by rules. Int. J. Geogr. Inf. Sci. 36 (2), 310–337. https://doi.org/10.1080/13658816.2021.1947507.

Kanfar, R., Alali, A., Tonellot, T.L., et al., 2025. Intelligent seismic workflows: the power of generative AI and language models. Lead. Edge 44 (2), 142–151. https://doi.org/10.1190/tle44020142.1.

Lee, S.C., Lee, D.G., Seo, Y.S., 2024. Determining the best feature combination through text and probabilistic feature analysis for GPT-2-based mobile app review detection. Appl. Intell. 54 (2), 1219–1246. https://doi.org/10.1007/s10489-023-05201-3.

Li, J., Li, D., Savarese, S., et al., 2023a. In: Blip-2: Bootstrapping language-image Pre-training with Frozen Image Encoders and Large Language Models. arXiv: 2301.12597. https://doi.org/10.48550/arXiv.2301.12597.

Li, J., Vo, D.M., Sugimoto, A., et al., 2024. Evcap: retrieval-augmented image captioning with external visual-name memory for open-world comprehension. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 13733–13742. https://doi.org/10.1109/CVPR52733.2024.01303.

Li, S., Cao, J., Ye, P., et al., 2025. ClipSAM: CLIP and SAM collaboration for zero-shot anomaly segmentation. Neurocomputing 618, 129122. https://doi.org/10.1016/j.neucom.2024.129122.

Li, X., Li, B., Liu, F., et al., 2023b. Advances in the application of deep learning methods to digital rock technology. Adv. Geo-Energy Res. 8 (1), 5–18. https://doi.org/10.46690/ager.2023.04.02.

Liu, H., Ren, Y., Li, X., et al., 2024. Research status and application of artificial intelligence large models in the oil and gas industry. Petrol. Explor. Dev. 51 (4), 1049–1065. https://doi.org/10.1016/S1876-3804(24)60524-0.

Liu, M., Mukerji, T., 2022. Multiscale fusion of digital rock images based on deep generative adversarial networks. Geophys. Res. Lett. 49 (9), e2022GL098342. https://doi.org/10.1029/2022GL098342.

Liu, W., Chen, S., Guo, L., et al., 2021a. CPTR: Full Transformer Network for Image Captioning. https://doi.org/10.48550/arXiv.2101.10804.

Liu, Y., Zhang, Q., Zhang, N., et al., 2022. Enhancement of thin-section image using super-resolution method with application to the mineral segmentation and classification in tight sandstone reservoir. J. Pet. Sci. Eng. 216, 110774. https://doi.org/10.1016/j.petrol.2022.110774.

Liu, Z., Lin, Y., Cao, Y., et al., 2021b. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022. https://doi.org/10.48550/arXiv.2103.14030.

Luo, X., Sun, J., Zhang, R., et al., 2024. A multi-condition denoising diffusion probabilistic model controls the reconstruction of 3D digital rocks. Comput. Geosci. 184, 105541. https://doi.org/10.1016/j.cageo.2024.105541.

Ma, H., Han, G., Peng, L., et al., 2021. Rock thin sections identification based on improved squeeze-and-excitation networks model. Comput. Geosci. 152, 104780. https://doi.org/10.1016/j.cageo.2021.104780.
Mokady, R., Hertz, A., Bermano, A.H., 2021. ClipCap: CLIP Prefix for Image Captioning. https://doi.org/10.48550/arXiv.2111.09734.
Papineni, K., Roukos, S., Ward, T., et al., 2002. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. https://doi.org/10.3115/1073083.1073135.
Pham, N., Di, H., Zhao, T., et al., 2025. SeisBERT: a pretrained seismic image representation model for seismic data interpretation. Lead. Edge 44 (2), 96–106. https://doi.org/10.1190/tle44020096.1.
Qi, H., Ba, J., Müller, T.M., 2021. Temperature effect on the velocity-porosity relationship in rocks. JGR Solid Earth 126 (1), e2019JB019317. https://doi.org/10.1029/2019JB019317.
Radford, A., Kim, J.W., Hallacy, C., et al., 2021. Learning transferable visual models from natural language supervision. arXiv:2103, 00020. https://doi.org/10.48550/arXiv.2103.00020.
Radford, A., Narasimhan, K., Salimans, T., et al., 2018. Improving Language Understanding by Generative Pre-training. https://openai.com/research/language-unsupervised.
Radford, A., Wu, J., Child, R., et al., 2019. Language models are unsupervised multitask learners. OpenAI blog 1 (8), 9. https://api.semanticscholar.org/CorpusID:160025533.
Raffel, C., Shazeer, N., Roberts, A., et al., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv:1910, 10683. https://doi.org/10.48550/arXiv.1910.10683.
Ramos, R., Martins, B., Elliott, D., et al., 2023. Smallcap: lightweight image captioning prompted with retrieval augmentation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, pp. 2840–2849. https://doi.org/10.1109/CVPR52729.2023.00278.
Saad, O.M., Chen, Y., Savvaidis, A., et al., 2022. Real-time earthquake detection and magnitude estimation using vision transformer. JGR Solid Earth 127 (5), e2021JB023657. https://doi.org/10.1029/2021JB023657.
Saxena, N., Day-Stirrat, R.J., Hows, A., et al., 2021. Application of deep learning for semantic segmentation of sandstone thin sections. Comput. Geosci. 152, 104778. https://doi.org/10.1016/j.cageo.2021.104778.
Schneider, E.T.R., De Souza, J.V.A., Gumiel, Y.B., et al., 2021. A GPT-2 language model for biomedical texts in portuguese. IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), Aveiro, Portugal, 2021, pp. 474–479. https://doi.org/10.1109/CBMS52027.2021.00056.
Shrestha, P., Kandel, J., Tayara, H., et al., 2024. Post-translational modification prediction via prompt-based fine-tuning of a GPT-2 model. Nat. Commun. 15 (1), 6699. https://doi.org/10.1038/s41467-024-51071-9.
Si, X., Wu, X., Sheng, H., et al., 2024. SeisCLIP: a seismology foundation model pretrained by multimodal data for multipurpose seismic feature extraction. IEEE Trans. Geosci. Rem. Sens. 62, 1–13. https://doi.org/10.1109/TGRS.2024.3354456.
Sotelo, E., Barbosa, N.D., Solazzi, S.G., et al., 2023. Homogenization of porous thin layers with internal stratification for the estimation of seismic reflection coefficients. JGR Solid Earth 128 (11), e2023JB026782. https://doi.org/10.1029/2023JB026782.
Tolstikhin, I., Houlsby, N., Kolesnikov, A., et al., 2021. MLP-Mixer: an all-MLP Architecture for Vision. arXiv: 2105.01601. https://doi.org/10.48550/arXiv.2105.01601.
Ulmer-Scholle, D.S., Scholle, P.A., Schieber, J., et al., 2015. A Color Guide to the Petrography of Sandstones, Siltstones, Shales and Associated Rocks. AAPG Memoir. https://doi.org/10.1306/M1091304.
Vedantam, R., Zitnick, C.L., Parikh, D., 2015. CIDEr: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566–4575. https://doi.org/10.48550/arXiv.1411.5726.
Vinyals, O., Toshev, A., Bengio, S., et al., 2015. Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164. https://doi.org/10.48550/arXiv.1411.4555.
Wang, J., Wang, K., Yu, Y., et al., 2024a. Self-improving generative foundation model for synthetic medical image generation and clinical applications. Nat. Med. 31 (2), 609–617. https://doi.org/10.1038/s41591-024-03359-y.
Wang, K., Liu, C., Wei, H., et al., 2024b. RFNET: refined fusion three-branch RGB-D salient object detection network. In: 2024 IEEE International Conference on Image Processing (ICIP), pp. 741–746. https://doi.org/10.1109/ICIP51287.2024.10647308.
Wu, H., Liu, H., Qiao, Y., et al., 2024. DIBS: enhancing dense video captioning with unlabeled videos via pseudo boundary enrichment and online refinement. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 18699–18708. https://doi.org/10.1109/CVPR52733.2024.01769.
Wu, Z., Weng, Z., Peng, W., et al., 2023. Building an open-vocabulary video CLIP model with better architectures, optimization and data. IEEE Trans. Pattern Anal. Mach. Intell. 46 (7), 4747–4762. https://doi.org/10.1109/TPAMI.2024.3357503.
Xu, K., Ba, J., Kiros, R., et al., 2016. Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning, pp. 2048–2057.
Yan, J., Xie, Y., Zou, S., et al., 2025. EntroCap: zero-shot image captioning with entropy-based retrieval. Neurocomputing 611, 128666. https://doi.org/10.1016/j.neucom.2024.128666.
Zhang, Z.H., Zhang, T., Liu, H.Q., et al., 2024. Control and prediction of bedding-parallel fractures in fine-grained sedimentary rocks: A case from the Permian Lucaogou Formation in Jimusar Sag, Junggar Basin, Western China. Pet. Sci. 21 (6), 3815–3838. https://doi.org/10.1016/j.petsci.2024.09.003.