We sincerely appreciate the constructive feedback provided by the reviewer NZ7M, which is invaluable for enhancing the quality of our paper. Below are our responses to the raised concerns. We also commit to accurately reflecting these points in the revised manuscript.

W1:

Regarding the motivation and novelty of using two-scale images:
Indeed, many existing works have adopted two-scale image features, as evidenced by LLaVA-Next, which serves as a baseline for our VLSA, concatenating features from high and low-resolution images as input to capture multi-scale information. However, they always bring considerable computational overhead since introducing high-resolution features boosts the LLM's input sequence length. Our contribution lies in improving the efficiency of utilizing two-scale image features. We propose to leverage low-resolution images for **(1)** *modeling the relationships between high-resolution sub-images* and **(2)** *compressing the length of their feature sequence*, thereby improving the performance while alleviating computational burden.

Specifically, we design SA-Perceiver, which comprising four $\mathbb{R}^{1024\times1024}$ linear layers and one $\mathbb{R}^{1024\times4096}$ linear layer (save 60% parameters compared with the projection module in LLaVA-Next, which consists of a $\mathbb{R}^{1024\times4096}$ and a $\mathbb{R}^{4096\times4096}$ linear layer), to integrate high-resolution image information into low-resolution image features at a lower cost. Only the low-resolution features are then utilized as input to the LLM. Given that the projection module (including MLP, Q-former, and our SA-Perceiver) has a significantly lower parameter count and computational complexity than the LLM, the overall system latency is predominantly dictated by the computation delay of the LLM. As SA-Perceiver enables a reduction of visual sequence length up to four-fold, and the time complexity of LLM is $O\left(n^2\right)$, our method can theoretically achieve a maximum reduction in latency by a factor of 16. However, system latency is also affected by factors such as the number of input text tokens, the length of the generated sentences, and other intricate system dynamics. To more accurately assess the impact of SA-Perceiver in reducing computational overhead, we randomly select 1,000 images, remove their textual instructions, resize them to various resolutions, and compare the latency and FLOPs of our method and LLaVA-Next during the feedforward process.

Latency of processing 1000 images (seconds):

Method	**336x336**	**672x336**	**1008x336**	**672x672**
LLaVA-Next	449	475	564	738
VLSA	**373**	**377**	**385**	**399**

FLOPs in processing 1000 images (GFLOPs):

Method	**336x336**	**672x336**	**1008x336**	**672x672**
LLaVA-Next	18798.9	27444.3	36462.3	45840.6
VLSA	**9842.3**	**9884.7**	**10190.8**	**10539.3**

Besides, as demonstrated in the main text, our compressive image encoding coupled with reconstructive training achieves performance comparable to or even better than other high-resolution methods. The authors believe this renders our work highly competitive compared to related efforts.

Regarding the motivation and novelty of cognition alignment:
The work [a] mentioned by the reviewer and some related works can, to a certain extent, facilitate MLLM's cognition abilities by improving the understanding of objectives and some pre-defined interrelations. However, these methods have some limitations. (1) Most additional fine-tuning tasks rely on manually annotated data (or a semi-automatic annotation process with human participation). (2) Existing tasks focus on certain attributes, making it challenging to comprehensively cover images' visual semantics. In contrast, our cognition alignment is constructed in a self-supervised manner, eliminating the need for manual annotation and significantly enhancing the model's scalability. Furthermore, given that VQ-VAE functions as an autoencoder, its image embeddings can preserve a broad spectrum of semantic information within the image, extending beyond merely the semantics associated with specific human-defined categories. Utilizing the codebook indices of VQ-VAE as learning targets can facilitate a more comprehensive understanding of visual semantics by the model.

We promise to bring more comparisons with related works in the revised manuscript.

W2:

In the revised manuscript, we will incorporate the following evaluation:

In response to the reviewer's comments regarding to SA-Perceiver:
The SA-Perceiver is designed to integrate features from high-resolution images into the features of low-resolution images. To achieve this, we

first implement a cross-attention layer that collects information from high-resolution images. As these high-resolution images are divided into multiple sub-images during preprocessing, we have subsequently incorporated a self-attention layer to enhance the modeling of interrelationships among the sub-image information. To ensure parameter efficiency, we have omitted certain projection layers typically found in the standard attention mechanism. To demonstrate the effectiveness of these design choices, we conducted the following ablation experiments on SA-Perceiver: (ex1) removing the self-attention layer and (ex2) retaining all linear projections in cross and self-attention (including those for key, query, value, and output).

Variant	GQA	SQA-I	DocVQA
(ex1) w/o. self-attn	65.1	76.9	72.4
(ex2) full projections	65.0	77.3	**75.3**
VLSA	**65.3**	**77.5**	75.2

In response to the reviewer's comments regarding the use of LDMs:
We previously outlined two intuitive reasons in Line260 for adopting LDMs instead of conventional AEs (AutoEncoder) for image reconstruction:

(1) Employing pretrained text-to-image LDMs can mitigate information loss during image encoding while facilitating the alignment of visual and textual representations.
(2) Pretrained text-to-image LDMs excel at accomplishing reconstructive tasks from a semantic perspective, thus assisting in the extraction of rich visual semantics during image encoding.

A further explanation for (2): This is attributed to the pretraining objectives of LDMs, which equip them with the capability to generate images from abstract, high-level semantics, whereas traditional AEs are not explicitly constrained to extract or comprehend such high-level visual semantics.
To further substantiate the necessity of using LDMs, we have conducted new ablations where we implemented reconstruction training using pretrained VAE (ex3) and VQ-VAE (ex4) models. The revised manuscript will include detailed information on these two AE encoders.

Variant	GQA	SQA-I	DocVQA
(ex3) Reconstruction by VAE	64.7	75.0	74.4
(ex4) Reconstruction by VQ-VAE	65.2	74.2	69.39
VLSA	**65.3**	**77.5**	**75.2**

In response to the reviewer's comments regarding VQ-VAE code indices:
We acknowledge that the task of Predicting Objects indeed facilitates the model's understanding of object-specific attributes. However, utilizing codebook indices promotes a more comprehensive semantic alignment, which could be more advantageous for improving performance across various downstream tasks. This advantage arises from the fact that codebook indices encapsulate a greater amount of information, as it enables a visual decoder to reconstruct most details in images from them. To further support this viewpoint, we report the performances of predicting codebook indices and predicting the bounding boxes provided by grounding DINO during fine-tuning. The implemental details will be included in the revised manuscript.

W3:

(1) VLSA's generality:
We have applied VLSA to Qwen-VL[1] and LLaMA-Adapter V2[2], and reported the preliminary experimental results. More results and implementation details will be included in the revised manuscript.

Variant	GQA	ChartQA	DocVQA	SEED-Bench	MME	COCO Cap
Qwen-VL	59.3	65.7	65.1	56.3	-	-
(ex5) Qwen-VL + VLSA	**62.1**	**66.4**	**65.4**	**62.0**	-	-
LLaMA-Adapter V2	-	-	-	32.7	1221	122.2
(ex6) LLaMA-Adapter V2 + VLSA	-	-	-	41.3	**1475**	**143.1**

[1] Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond
[2] LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model

(2) Fairness in comparisions:

Since our model is developed based on the LLaVA-Next, we directly reference the models it has compared and strive to ensure a fair comparison with LLaVA-Next. To clarify, all the results for LLaVA-Next presented in our tables were reproduced using the same CLIP encoder and batch size as VLSA, guaranteeing a completely equitable comparison. To further address the reviewer's concerns, we provide the performance of VLSA using SigLIP as the encoder.

Variant	Vision Encoder	Batchsize	AI2D	MMB	DocVQA
LLaVA-Next (Reported)	SigLIP	512	71.6	72.1	78.2
(ex7) LLaVA-Next (Reproduced)	SigLIP	128	71.4	71.7	78.3
(ex8) VLSA	SigLIP	128	**73.1**	**72.9**	**80.1**

Indeed, achieving a completely fair comparison with other methods is quite challenging. To the best of our effort, we have included the performance of VLSA without the OCR training dataset (LLaVAR) and with the replacement of the backbone model to LLaMA2-7B and Vicuna1.5-7B as references.

Variant	LLM	OCR Data	GQA	AI2D	DocVQA
(ex9) VLSA	LLaMA2-7B	✓	63.0	68.4	73.2
(ex10) VLSA	Vicuna1.5-7B	✓	63.6	67.5	74.6
(ex11) VLSA	LLaMA3-8B	×	65.1	69.7	73.9
VLSA	LLaMA3-8B	✓	**65.3**	**71.4**	**75.2**

(3) Mixed results in Table 3:

Our earlier interpretation of Table 3 may have been somewhat misleading. A comparison of experiments (2) and (3) reveals that compressive image encoding significantly alleviates the perception performance degradation associated with lowering the actual input resolution. Furthermore, experiment (4) demonstrates that, with reconstructive training, low-resolution inputs can achieve perception performance that even surpasses that of high-resolution inputs in experiment (1), leading to substantial improvements across most tasks. The performance drop observed in AI2D in the experiment (4) suggests that when providing LLMs with richer visual semantic information, they may struggle to comprehend it accurately, potentially resulting in negative outcomes. This highlights the need for our proposed perception alignment to work in conjunction with cognition alignment. Therefore, the experimental results presented in the table are still in line with our expectations.

(3)Missing Ablations:

In our previous responses, we have provided ablations regarding the structure of SA-Perceiver (ex1, ex2), the vision backbone (ex8), the language model (ex9, ex10), and the type of image generation model used in reconstructive training (ex3, ex4). Here, we report the ablation study on the two tasks of fine-tuning: (ex12) predicting Codebook indices and (ex13) predicting RGB pixel values.

Variant	Predict RGB	Predict Codebook indices	AI2D	SQA-I	ChartQA
w/o. Cognition Alignment	×	×	68.2	74.1	67.4
(ex12) w/o. RGB	×	✓	69.8	77.2	67.7
(ex13) w/o. Codebook	✓	×	68.4	71.6	67.5
VLSA	✓	✓	**71.4**	**77.5**	**67.9**

By comparing the results without cognition alignment to (ex9) and (ex10), we observe that predicting RGB values alone enhances the model's fine-grained cognitive capabilities, benefiting document understanding tasks. However, it may overly emphasize low-level semantics, detrimentally affecting the understanding of high-level semantics and leading to a significant performance drop on SQA. In contrast, predicting codebook indices alone consistently improves performance across various tasks. This might be attributed to VQ-VAE, as an autoencoder, being able to balance semantics at different levels. Furthermore, combining both tasks yields further improvements.