

Resize

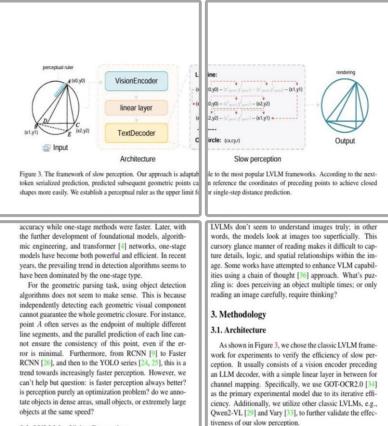
Mode: Tiny||Small

Token: 64||100

|**←** W:1024||1280 → Figure 3. The framework of slow perception. Our approach is adaptable to the most popular LVLM frameworks. According to the nexttoken serialized prediction, predicted subsequent geometric points can reference the coordinates of preceding points to achieve closed shapes more easily. We establish a perceptual ruler as the upper limit for single-step distance prediction. accuracy while one-stage methods were faster. Later, with LVLMs don't seem to understand images truly; in other the further development of foundational models algorith- words the models look at images too superficially. This mic engineering, and transformer [4] networks, one-stage cursory glance manner of reading makes it difficult to capmodels have become both powerful and efficient. In recent ture details, logic, and spatial relationships within the imyears, the prevailing trend in detection algorithms seems to age. Some works have attempted to enhance VLM capabilhave been dominated by the one-stage type. ities using a chain of thought [36] approach. What's puz-For the geometric parsing task, using object detection zling is: does perceiving an object multiple times; or only algorithms does not seem to make sense. This is because reading an image carefully, require thinking? independently detecting each geometric visual component cannot guarantee the whole geometric closure. For instance, point A often serves as the endpoint of multiple different line segments, and the parallel prediction of each line cannot ensure the consistency of this point, even if the er-As shown in Figure 3, we chose the classic LVLM frame ror is minimal. Furthermore, from RCNN [9] to Faster work for experiments to verify the efficiency of slow per-RCNN [26], and then to the YOLO series [24, 25], this is a ception. It usually consists of a vision encoder preceding trend towards increasingly faster perception. However, we an LLM decoder, with a simple linear layer in between for N can't help but question: is faster perception always better? channel mapping. Specifically, we use GOT-OCR2.0 [34] is perception purely an optimization problem? do we annotate objects in dense areas, small objects, or extremely large ciency. Additionally, we utilize other classic LVLMs, e.g. objects at the same speed? Owen2-VL [29] and Vary [33], to further validate the effec-2.2. LVLM for Vision Perception Recently, research on large vision-language models (LVLMs) [16, 2, 37, 7] has been on the rise, and these mod-We render 200k synthetic geometric images as the train els have demonstrated state-of-the-art performance in various visual perception tasks, such as OCR [27, 34, 15, 5] and gine. We stochastically vary multiple parameters to ensure grounding [40, 38]. After more than a year of development, data heterogeneity, including line thickness, line style (solid the framework of these LVLMs has become quite conver-or dashed), and image resolution (DPI). In total, 150k imgent. Specifically, new models often adopt an "encoderperceiver-decoder" architecture and utilize a training aply involving pretraining followed by supervised fine-tuning practical applications. For the corpus of point-line locations (SFT). It is worth noting that the powerful visual knowledge (open-set universal object recognition capability) of lowing generation procedure LVLMs has also left a deep impression on people, leading 1) Selection of substrate. We select the most common us to have very high expectations for LVLMs. quadrilaterals as the rendering base, including squares, rect-However, some works like BlindTest [23] show that angles, parallelograms, rhombuses, trapezoids, isosceles Padding

Mode: Base||Large

Token: 256||400 Valid: (256||400)×R R=1-(H-W)/W



2.2. LVLM for Vision Perception

Recently, research on large vision-language models LVLMs) [16, 2, 37, 7] has been on the rise, and these models have demonstrated state-of-the-art performance in various visual perception tasks, such as OCR [27, 34, 15, 5] and grounding [40, 38]. After more than a year of development, the framework of these LVLMs has become quite convergent. Specifically, new models often adopt an "encoderperceiver-decoder" architecture and utilize a training ap proach similar to large language models (LLMs), primarily involving pretraining followed by supervised fine-tuning (SFT). It is worth noting that the powerful visual knowldge (open-set universal object recognition capability) of LVLMs has also left a deep impression on people, leading us to have very high expectations for LVLMs.

However, some works like BlindTest [23] show that

←640||1024**→**|

We render 200k synthetic geometric images as the train ata, wherein Matplotlib is employed as the rendering engine. We stochastically vary multiple parameters to ensure ata heterogeneity, including line thickness, line style (solid r dashed), and image resolution (DPI). In total, 150k imges are generated with DPI values randomly distributed beween 36 and 300, while the remaining 50k are uniformly set to 96 DPI, reflecting a commonly used resolution in ractical applications. For the corpus of point-line locations d relationships that make up geometry, we devise the folwing generation procedure

) Selection of substrate. We select the most common uadrilaterals as the rendering base, including squares, rectangles, parallelograms, rhombuses, trapezoids, isosceles

Figure 3. The framework of slow perception. Our approach is adaptable to the most popular LVLM frameworks. According to the nexttoken serialized prediction, predicted subsequent geometric points can reference the coordinates of preceding points to achieve closed shapes more easily. We establish a perceptual ruler as the upper limit for single-step distance prediction accuracy while one-stage methods were faster. Later, with LVLMs don't seem to understand images truly; in other the further development of foundational models, algorith- words, the models look at images too superficially. This mic engineering, and transformer [4] networks, one-stage cursory glance manner of reading makes it difficult to cap models have become both powerful and efficient. In recent ture details, logic, and spatial relationships within the im vears, the prevailing trend in detection algorithms seems to age. Some works have attempted to enhance VLM capabil-

algorithms does not seem to make sense. This is because reading an image carefully, require thinking? independently detecting each geometric visual component cannot guarantee the whole geometric closure. For instance,

point A often serves as the endpoint of multiple different line segments, and the parallel prediction of each line cannot ensure the consistency of this point, even if the er-As shown in Figure 3 we chose the classic LVLM frame ror is minimal. Furthermore, from RCNN [9] to Faster work for experiments to verify the efficiency of slow per-RCNN [26], and then to the YOLO series [24, 25], this is a ception. It usually consists of a vision encoder preceding

For the geometric parsing task, using object detection zling is: does perceiving an object multiple times; or only

an LLM decoder, with a simple linear layer in between for channel mapping. Specifically, we use GOT-OCR2.0 [34 is perception purely an optimization problem? do we annociency. Additionally, we utilize other classic LVLMs, e.g.,

Owen2-VL [29] and Vary [33], to further validate the effect

ities using a chain of thought [36] approach. What's puz

We render 200k synthetic geometric images as the train els have demonstrated state-of-the-art performance in various visual perception tasks, such as OCR [27, 34, 15, 5] and gine. We stochastically vary multiple parameters to ensure grounding [40, 38]. After more than a year of development, data heterogeneity, including line thickness, line style (solid the framework of these LVLMs has become quite conver-or dashed), and image resolution (DPI). In total, 150k imgent. Specifically, new models often adopt an "encoderperceiver-decoder" architecture and utilize a training approach similar to large language models (LLMs), primar-set to 96 DPI, reflecting a commonly used resolution in practical applications. For the corpus of point-line locations and relationships that make up geometry, we devise the fol-

edge (open-set universal object recognition capability) of lowing generation procedure:

quadrilaterals as the rendering base, including squares, rect However, some works like BlindTest [23] show that angles, parallelograms, rhombuses, trapezoids, isosceles

have been dominated by the one-stage type

objects at the same speed?

2.2. LVLM for Vision Perception

trend towards increasingly faster perception. However, we

can't help but question: is faster perception always better?

tate objects in dense areas, small objects, or extremely large

Recently, research on large vision-language models

ily involving pretraining followed by supervised fine-tuning

(SFT). It is worth noting that the powerful visual knowl-

us to have very high expectations for LVLMs

(LVLMs) [16, 2, 37, 7] has been on the rise, and these mod-

|**←** W:1024||1280 —

Mode: Gundam||Gundam (Master)

Token: n×(100||256) + (256||400) Valid: n×(100||256) + (256||400)×R n∈[2:9]