A. Quantitative analysis

A.1. Impact of EmbQ configuration

Table 4 studies how EmbQ's dimensions, layers, and insertion positions affect performance. The first two blocks show that EmbQ works best with 576 dimensions and one layer, matching FlashSloth's efficiency. The third block tests inserting EmbQ at different depths. Shallow insertion limits its self-interaction, while deep insertion hinders information flow.

Table 4. Ablation study on the effects of EmbQ dimensions, layer count, and insertion layers across four benchmarks. The ‡denotes the final selected setting.

Choices	GQA	POPE	MME	MMB
Dimension of EmbQ				
576 ‡	60.8	87.8	1490.7	67.9
768	60.6	86.8	1446.7	66.6
1152	60.6	86.8	1455.5	66.4
2560	60.5	87.1	1420.8	67.1
Number of EmbQ layer				
1 ‡	60.8	87.8	1490.7	67.9
2	60.6	86.9	1488.7	67.1
3	60.6	87.2	1457.5	67.3
EmbQ Insertion Layer				
4	60.7	86.9	1433.3	66.2
8 ‡	60.8	87.8	1490.7	67.9
16	60.4	87.3	1399.2	66.5
24	60.8	87.1	1433.7	67.8
8/16/24	60.8	86.1	1414.5	67.0

A.2. Impact of Feature Fusion Methods

Table 5 compares methods for fusing features learned by EmbQ with Query Token features: direct replacement, addition, and gated fusion [52], which adjusts the contribution of each feature. Results show that direct addition achieves the best performance by effectively integrating EmbQ's features while retaining the original feature information.

Table 5. Ablation experiment results on four benchmarks under different feature fusion method

Feature Fusion	GQA	POPE	MME	MMB
Add	**60.8**	**87.8**	**1490.7**	**67.9**
Replace	60.5	86.3	1443.7	66.4
Gate fusion	60.7	86.1	1468.1	66.1

A.3. Comparison with Other Vision Compression Methods

Table 6 compares our method with various visual feature compression approaches. It can be seen that our scheme of SAP+EmbQ not only outperforms the compared methods, but is also the only solution that can avoid performance drops on most benchmarks.

Table 6. Comparison between different visual compression methods. Baseline refers to no token reduction (729 tokens), while the other methods uses 81 tokens. The training data is LLaVA-665K.

Method	# Params	TFLOPs	GQA	MMB	MME	POPE
Baseline (729)	9.5M	2.384	**61.8**	67.9	1491	87.8
LDPNet-v2 [14]	9.5M	0.672	59.4	65.3	1439	85.6
Pixel Shuffle [11]	33.1M	0.664	58.6	65.8	1444	85.0
C-abstractor [7]	27.2M	0.672	58.5	65.7	1422	83.9
TokenPacker [33]	29.4M	0.672	57.7	63.9	1400	82.4
Our SAP+EmbQ	18.9M	0.688	60.8	**67.9**	**1491**	**87.8**

Table 7. Results of our visual compression designs on LLaVA-1.5-7B. We use the same base model (CLIP&Vicuna-7B) and training paradigm, and reduce visual tokens to 64.

Method	#Tokens	TFLOPS	MMB	MME	MM-VET	GQA	SQA	InfoVQA
LLaVA [39]	576	4.55	64.3	1511	30.5	**62.0**	66.8	25.7
+Ours	**64**	**2.62**	**64.4**	**1524**	**30.6**	61.1	**69.4**	**27.4**

To further validate the versatility of our approach, we applied our compression method to the LLaVA in table 7. It can be seen that our innovative designs barely impede the performance while reducing the token numbers 9 times. In particular, we can even slightly improve the performance of LLaVA-7B on the benchmarks, such as MME and SQA.

A.4. Performance in video tasks.

Table 8. Zero-shot evaluation on video understanding benchmark with 8-frame input sampling

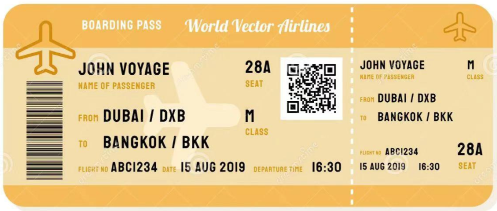
Method	#Tokens	Seed^V	CVRR	AcNetQA	MME^V	TempCom
Video-LLaVA [38]	2048	36.9	15.9	45.3	39.9	49.8
Qwen-vl-chat [4]	1024	34.3	-	-	41.1	52.9
Video-LLaMA-2 [12]	576	-	21.6	49.0	-	43.4
FlashSloth	648	**43.9**	**49.1**	**49.0**	**45.5**	**64.4**

we conduct zero-shot test of our FlashSloth on several video benchmarks [17, 28, 29, 43, 71], which still achieves unexpectedly good results. we conduct **zero-shot test** of FlashSloth on 5 video benchmarks (8 frame inputs, 8 × 81 tokens) in Table 8. Which can even achieve better results than some fully tuned video-MLLMs, well confirming its VL capability.

B. Qualitative Analysis

This section showcases practical examples of FlashSloth and FlashSloth-HD, demonstrating their real-world performance in multidisciplinary question answering, code generation, real-world scene reasoning, fine-grained text information extraction, and chart analysis reasoning. FlashSloth delivers accurate responses and exceptional performance across these tasks.

OCR


BOARDING PASS
World Vector Airlines

JOHN VOYAGE
NAME OF PASSENGER
28A
SEAT
FROM DUBAI / DXB
M
CLASS
TO BANGKOK / BKK
FLIGHT NO ABC1234 DATE 15 AUG 2019 DEPARTURE TIME 16:30

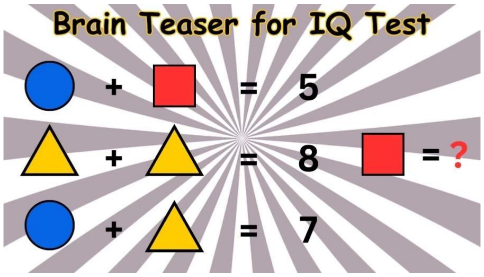
JOHN VOYAGE
NAME OF PASSENGER
M
CLASS
FROM DUBAI / DXB
TO BANGKOK / BKK
FLIGHT NO ABC1234
15 AUG 2019 16:30
28A
SEAT

USER: Please read the text in this image and return the information in the following JSON format (note xxx is placeholder, if the information is not available in the image, put 'NONE' instead).{"airport": xxx, "flight_number": xxx, "from": xxx, "destination": xxx, "class": xxx, "seat": xxx, "date": xxx, "boarding_time": xxx, "gate": xxx, "name": xxx, "ticket_number": xxx}

FlashSloth: {"airport": " World Vector Airlines ", "flight_number": "28", "from": "Bangkok", "destination": "Dubai", "class": "economy", "seat": "S-224", "date": "15 AUG 2019", "boarding_time": "16:00", "gate": "A-23", "name": "JOHN VOGAGE", "ticket_number": "NONE"}

FlashSloth-HD: { "airport": "World Vector Airlines", "flight_number": "ABC1234", from": "DUBAI / DXB", destination": "BANGKOK / BKK", "class": "M", "seat": "28A", "date": "15 AUG 2019", "boarding_time": "16:30", "gate": "NONE", "name": "JOHN VOYAGE", "ticket_number": "ABC1234",}

Mathematical Problem Solving



USER: Find the value of the square in the figure.

FlashSloth: To solve this brain teaser for the IQ test, we need to use the given information and the properties of shapes to find the value of the square.
Given information:
- The figure contains a blue circle, a red triangle, and a red square.

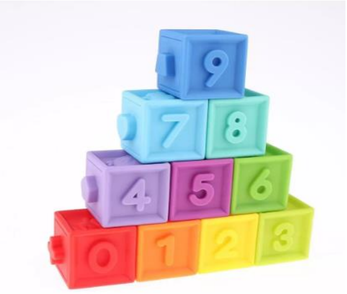
To find the value of the square, we can use the following steps:
1. Observe the pattern: The sum of the values of the blue circle and the red triangle is 5.
2. The sum of the values of two yellow triangle is 8.
3. The sum of the values of the blue circle and the yellow square is 7.
4. Since the sum of the values of two yellow triangle is 8 , each yellow triangle has a value of 4.
5. From step 3, the sum of the blue circle and the yellow square is 7.
6. From step 1, the sum of the blue circle and the red triangle is 5.

Therefore, the value of the square in the figure is 2.

FlashSloth-HD: The image shows a brain teaser for an IQ test. The question asks to find the value of the square in the figure.
To solve this, we need to look at the relationships between the different shapes and numbers in the image.
The image shows a series of shapes - a blue circle, a red square, a yellow triangle, and a red square. The numbers next to these shapes are 5, 8, and 7.
The question states that the blue circle plus the red square equals 5. This means that the value of the blue circle is 5.
The yellow triangle plus the red square equals 8. This means that the value of the yellow triangle is 8 - 5 = 3.
Finally, the red square plus the blue circle equals 7. This means that the value of the red square is 7 - 5 = 2.
the value of the square in the figure is 2.

Figure 5. FlashSloth's Performance in Ticket OCR Recognition and Mathematical Question Answering.

Multi-object Recognition

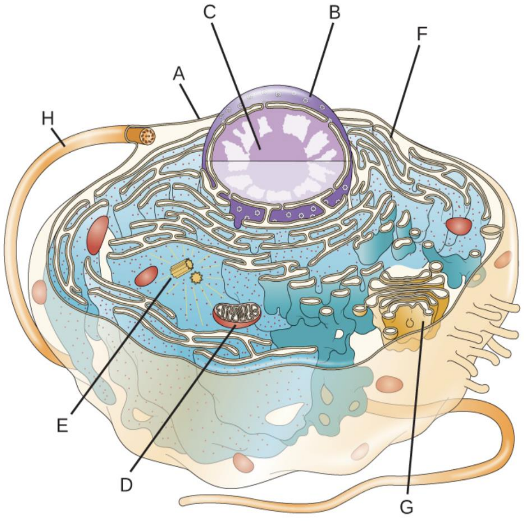


USER: Output the color and number of each box.

FlashSloth: The image shows a set of colorful blocks arranged in a pyramid-like structure. Each block has a number printed on it, ranging from 0 to 9. The colors of the blocks are red, orange, yellow, green, blue, purple, pink, and gray

FlashSloth-HD: The image shows a stack of blocks with numbers and colors. The colors of the blocks are: - Red: 0 - Orange: 1 - Yellow: 2 - Green: 3 - Purple: 4 - Blue: 5 - Teal: 6 - Cyan: 7 - Light Blue: 8 - Dark Blue: 9
Each block has a number printed on it, ranging from 0 to 9. The numbers are arranged in ascending order from the bottom to the top of the stack. The colors are arranged in a rainbow order, starting with red at the bottom and ending with blue at the top.

Biological Problem Solving



USER: What is the part of the cell labeled 'B' in image?

FlashSloth: The part of the cell labeled 'B' in the image is the nucleus. The nucleus is the control center of the cell, containing the genetic material (DNA) that directs the cell's activities. It is typically located in the center of the cell and is surrounded by a nuclear membrane. The nucleus is responsible for regulating gene expression, cell division, and other important cellular processes.

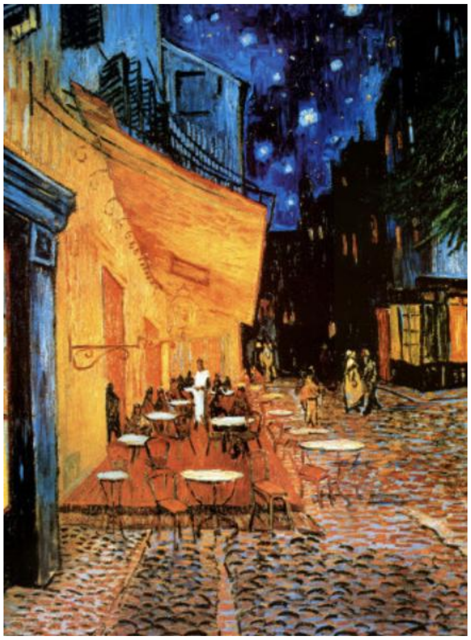
To arrive at this answer, I carefully examined the labeled diagram of the cell and identified the structure labeled 'B', which is clearly the nucleus. I then provided a brief explanation of the nucleus's role as the control center of the cell, based on my understanding of cellular biology.

When answering questions about scientific diagrams or images, it's important to closely observe the details provided and match them to the relevant biological concepts. In this case, the label 'B' directly corresponds to the nucleus, so I was able to confidently identify and explain this key cellular component.

FlashSloth-HD: The part of the cell labeled 'B' in the image is the nucleus. The nucleus is typically the largest organelle in a eukaryotic cell and is often referred to as the "control center" of the cell. It contains the cell's DNA and is responsible for directing the cell's activities.

Figure 6. FlashSloth's Performance in Recognizing the Colors and Content of Multiple Objects and on Biological Question Answering Tasks.

Artistic Problem Solving

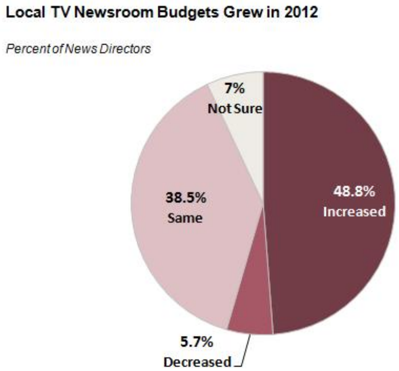


USER: Who painted this image?

FlashSloth: Vincent van Gogh

FlashSloth-HD: The painting in the image is by Vincent van Gogh, a Dutch post-impressionist painter. Van Gogh is known for his expressive and emotive style, which often included bold colors and dynamic brushstrokes. This particular painting is titled "The Starry Night Over the Rhone at Saint-Rémy-de-Provence" and was painted in 1888. It is one of his most famous works and is celebrated for its vivid depiction of the night sky and the swirling patterns of the stars. Van Gogh's work is often associated with the Post-Impressionist movement, which sought to break away from the naturalism of Impressionism and instead focus on more symbolic and emotional content.

Table Analysis



USER: Describe in detail the percentage of each section.

FlashSloth: The pie chart shows the percentage of news directors who reported changes in their local TV news budgets in 2012. The chart is divided into three sections, each representing a different category of change.

The largest section, colored in a dark red hue, represents 48.8% of news directors who reported an increase in their local TV news budgets. This indicates that nearly half of the news directors surveyed experienced an increase in their news budgets.

The second largest section, colored in a light pink hue, represents 38.5% of news directors who reported a "Same" budget. This suggests that a significant portion of news directors did not see any change in their news budgets.

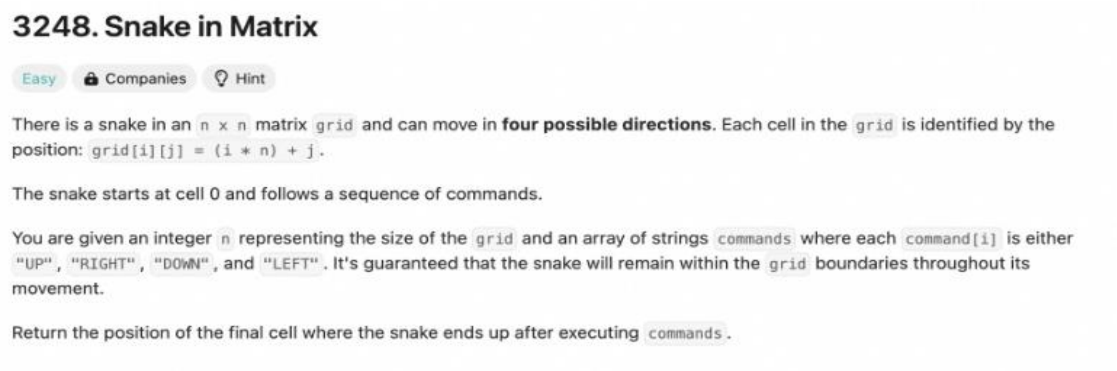
The smallest section, colored in a light purple hue, represents 5.7% of news directors who reported a "Decreased" budget. This indicates that a small fraction of news directors experienced a decrease in their news budgets.

FlashSloth-HD: The pie chart shows the percentage of news directors who reported changes in their local TV newsroom budgets in 2012. The chart is divided into four sections, each representing a different category of change:

1. Increased: 48.8% - This section represents the news directors who reported an increase in their local TV newsroom budgets in 2012.
2. Same: 38.5% - This section represents the news directors who reported that their local TV newsroom budgets remained the same in 2012.
3. Decreased: 5.7% - This section represents the news directors who reported a decrease in their local TV newsroom budgets in 2012.
4. Not Sure: 7% - This section represents the news directors who were unsure about the change in their local TV newsroom budgets in 2012.

Figure 7. FlashSloth's Performance in Artistic Question Answering and Chart Content Analysis.

Code Generation



3248. Snake in Matrix

There is a snake in an `n x n` matrix `grid` and can move in **four possible directions**. Each cell in the `grid` is identified by the position: `grid[i][j] = (i * n) + j`.

The snake starts at cell 0 and follows a sequence of commands.

You are given an integer `n` representing the size of the `grid` and an array of strings `commands` where each `command[i]` is either `"UP"`, `"RIGHT"`, `"DOWN"`, and `"LEFT"`. It's guaranteed that the snake will remain within the `grid` boundaries throughout its movement.

Return the position of the final cell where the snake ends up after executing `commands`.

USER: Solve the algorithmic problem in the image.

FlashSloth-HD: The image describes a snake in a matrix, which is a common problem in computer science and mathematics. The snake is represented by a sequence of commands, where each command is either "RIGHT", "DOWN", or "LEFT". The goal is to determine the final position of the snake after executing all the commands.

To solve this problem, we can use a simple algorithm that iterates through the commands and updates the snake's position accordingly. Here's a Python code that implements this algorithm:

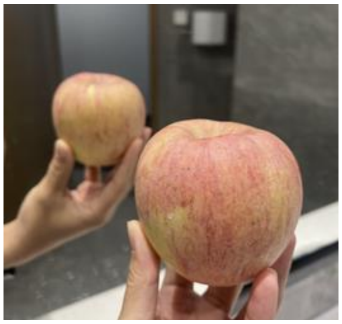
```python
def snake_in_matrix(matrix, commands):
    # Initialize the snake's position
    snake_position = [0, 0]

    # Iterate through the commands
    for command in commands:
        # Update the snake's position based on the command
        if command == "RIGHT":
            snake_position[0] += 1
        elif command == "DOWN":
            snake_position[1] += 1
        elif command == "LEFT":
            snake_position[0] -= 1

    # Return the final position of the snake
    return snake_position
```

This function takes two arguments: the matrix and the sequence of commands. The matrix is a 2D list representing the grid, and the commands are a list of strings representing the snake's movements.

The function initializes the snake's position to the top-left corner of the grid. It then iterates through the commands and updates the snake's position based on the command. If the command is "RIGHT", the snake moves one cell to the right. If the command is "DOWN", the snake moves one cell down. If the command is "LEFT", the snake moves one cell to the left.
Finally, the function returns the final position of the snake.

Figure 8. FlashSloth's Performance in Solving Algorithmic Problems Using Code.

Vision Reasoning



USER: How many apples in the picture?

FlashSloth: 2

FlashSloth-HD: 2

USER: Are they the same apple and why?

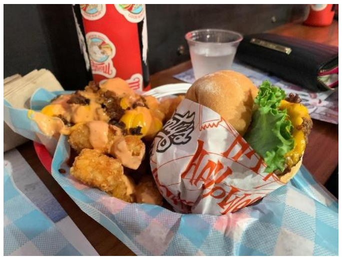
FlashSloth: Yes, they are the same apple. This can be inferred from the visual similarity in their color and texture, as well as the fact that they are both held in the same hand, suggesting they are being compared or are of the same variety. The slight variations in color and texture are common in apples and do not necessarily indicate different varieties.

FlashSloth-HD: The apples appear to be the same due to their similar coloration and size. The apple in the foreground has a more pronounced blush and a slightly different pattern of red and yellow spots, which could be due to the angle of the photograph or the natural variation in the apple's skin. However, without a closer examination or additional information, it is not possible to definitively determine if they are the same variety or if they are simply apples of similar appearance.

USER: Where is the water compared to the tater tots?

FlashSloth: To the right

FlashSloth: The water is behind to the right of the tater tots.



USER: I have high cholesterol. Is it suitable for me to eat these foods? Why?

FlashSloth: It is generally not recommended for individuals with high cholesterol to consume foods that are high in saturated fats and cholesterol, as these can contribute to elevated cholesterol levels. The burger and fries in the image are likely to be high in saturated fats, which can increase LDL (bad) cholesterol levels. Additionally, the cheese and mayonnaise on the burger are also sources of saturated fats.

However, it is important to note that the impact of these foods on cholesterol levels can vary depending on the individual's overall diet and lifestyle. If you have high cholesterol, it is advisable to consult with a healthcare professional or a registered dietitian to determine the best dietary choices for your specific needs. They can provide personalized recommendations and guidance on managing your cholesterol levels through diet and other lifestyle modifications.

In general, it is recommended to choose healthier options such as lean proteins, whole grains, fruits, and vegetables, and to limit the intake of saturated fats and cholesterol-rich foods. This can help maintain a balanced diet and support overall health, including cholesterol management.

Figure 9. FlashSloth's Performance in Real-World Scene Reasoning Tasks.