

# Monet: Reasoning in Latent Visual Space Beyond Image and Language

## Supplementary Material

### A. Additional Experimental Results

We present the complete results on MME-RealWorld-Lite in Table 6. This benchmark spans a wide range of task domains, such as monitoring, autonomous driving, and complex diagram data. Compared with baseline approaches, Monet achieves consistently strong performance across all data sources, demonstrating its generality in both visual reasoning and perception tasks.

### B. Implementation Details

#### B.1. SFT Training

We use TRL [36] to implement our SFT training. In a training CoT, we allow the model to predict the special latent-start token `<latent>` while masking the latent-end token `</latent>`, the observation-start token `<observation>`, and the observation-end token `</observation>` in the SFT labels.

**Hyperparameters.** We list the training hyperparameters of SFT in Table 7. To avoid OOM, we use a batch size of 1 and set gradient accumulation steps to 16. Additionally, we restrict the total number of pixels (denoted by “max total pixels”) in a training sequence to prevent the OOM issue caused by too many auxiliary images.

**Latent only backpropagation in SFT Stage 2.** To ensure that gradients flow only through the latent representations to the model parameters, we replace  $\mathcal{L}_{\text{align-obs}}$  with the following surrogate loss:

$$\mathcal{L}'_{\text{align-obs}} = \frac{1}{N} \sum_i \text{stop\_grad} \left( \frac{\partial \mathcal{L}_{\text{align-obs}}}{\partial \hat{\mathbf{h}}_{\text{latent}}^{(i,L)}} \right)^\top \hat{\mathbf{h}}_{\text{latent}}^{(i,L)}, \quad (9)$$

where `stop_grad` denotes the stop-gradient operation,  $\hat{\mathbf{h}}_{\text{latent}}^{(i,L)}$  are the generated latent embeddings in the student CoT,  $N$  is the number of the training data and  $L$  is the number of layers in the MLLM’s language model. By differentiating  $\mathcal{L}'_{\text{align-obs}}$  with respect to the model parameters  $\theta$ :

$$\frac{\partial \mathcal{L}'_{\text{align-obs}}}{\partial \theta} = \frac{\partial \mathcal{L}_{\text{align-obs}}}{\partial \hat{\mathbf{h}}_{\text{latent}}^{(i,L)}} \frac{\partial \hat{\mathbf{h}}_{\text{latent}}^{(i,L)}}{\partial \theta}, \quad (10)$$

we can see that minimizing  $\mathcal{L}'_{\text{align-obs}}$  is equivalent to optimizing the parameters  $\theta$  through only the generated latent embeddings.

#### B.2. RL Training

We implement our RL training on EasyR1 [52], an open-source RL training framework for multimodal LLMs.

**Hyperparameters.** We summarize the RL training hyperparameters in Table 8. Note that we apply an “accuracy threshold” to retain only samples whose group accuracy falls below this threshold but remains non-zero, preventing the advantage from vanishing when the responses in a group are all correct or incorrect.

### B.3. Training Data Construction

We provide the prompt for the API models used in the *stage 3* of our training data construction pipeline as below:

#### System Prompt for Training Data Construction

You are a helpful assistant. Your task is to identify which parts of the solution rely on visual observations (marked by `<latent></latent>`). Wrap these essential observations in `<observation>...</observation>` while leaving all other text unchanged. The input contains only a question (with conditions) and a solution chain-of-thought text — no images. Identifying observations does not require access to images. Segments are delimited by `<STEP_i> ... <END_STEP_i>`.

Rules:

1. Do NOT remove, rename, or reorder any markers. Always output all markers exactly as given.
2. Do NOT add new content to the reasoning.
3. Inside `<observation>...</observation>`, include as little as possible — only the key tokens that truly require image observation.
4. Do NOT put information that can be inferred from the question text or earlier reasoning into `<observation>...</observation>`.
5. If you think there’s no observation, just return the input text unchanged.

### C. Detailed Experimental Setup

We use the VLMEvalKit [9] framework for our evaluation. We set the inference-time maximum visual tokens to be  $8192 \times 28 \times 28$ . Below is the system prompt we used for evaluation.

#### System Prompt for Evaluation

You are a helpful multimodal assistant. You are required to answer the question based on the image provided. Put your final answer in `\\boxed{ }`.

Table 6. Full result on MME-RealWorld-Lite.

	MME-RealWorld-Lite										
	Reasoning					Perception					
	Overall	MO	AD	OCR	DT	overall	MO	AD	OCR	DT	RS
Qwen2.5-VL-7B-Instruct	39.73	30.67	27.00	74.00	70.00	49.62	31.03	31.43	90.80	85.00	39.33
+ SFT	46.40	52.00	33.00	73.00	65.00	54.41	40.44	36.29	88.80	<b>86.00</b>	48.00
+ SFT + GRPO	48.13	<b>59.33</b>	33.75	71.00	66.00	55.18	43.26	37.71	89.60	83.00	45.33
DeepEyes	50.53	46.67	40.25	<b>78.00</b>	70.00	56.63	<b>43.89</b>	38.86	<b>90.00</b>	84.00	51.33
Monet-SFT	50.53	48.67	<b>43.00</b>	65.00	69.00	54.06	41.69	44.57	76.40	71.00	54.00
Monet-7B (Monet-SFT + VLPO)	<b>51.07</b>	46.00	41.50	73.00	<b>75.00</b>	<b>58.34</b>	41.07	<b>48.86</b>	85.60	84.00	<b>54.67</b>

Table 7. Hyperparameters for SFT.

Hyperparameter	Value
learning rate	0.00001
batch size	1
gradient accumulation steps	16
weight decay	0.01
SFT Stage 1 steps	3884 (3 epochs)
SFT Stage 2 steps	1000 (1.03 epoch)
SFT Stage 3 steps	1000 (1.03 epoch)
latent size	{8, 10, 12}
SFT Stage 1 max total pixels	2000×28×28
SFT Stage 2 max total pixels	1500×28×28
SFT Stage 3 max total pixels	2000×28×28
SFT Stage 1,2,3 max pixels per img	2000×28×28
SFT Stage 2 alignment weight	2.0
SFT Stage 3 alignment weight	2.0

Table 8. Hyperparameters for RL.

Hyperparameter	Value
learning rate	0.000001
batch size	64
weight decay	0.01
rollout size	8
temperature	0.5
max response length	4096
VLPO $\sigma$	10.0
SFT Stage 1 steps	3884 (3 epochs)
SFT Stage 2 steps	1000 (1.03 epoch)
SFT Stage 3 steps	1000 (1.03 epoch)
latent size	10
max pixels per img	2000×28×28
accuracy threshold	0.6

To ensure reliable evaluation, we first apply a rule-based judge (exact match) and then employ DeepSeek-V3.1 or Gemini-2.5-Pro as secondary judges.

## D. Computational Cost Analysis

We report the training runtime in Figure 6 on eight H800 GPUs (note that our training was conducted on eight A100 GPUs). As shown in Figure 6(a), Stages 1, 2, and 3 of Monet SFT require 21.92 h, 10.43 h, and 7.48 h, respectively. Figure 6(b) shows that, compared to vanilla SFT, Monet SFT in Stages 2 and 3 incurs approximately 1.85× and 1.32× higher computational cost per step. However, since Stages 2 and 3 converge within only 1,000 training steps (approximately 1 epoch), the overall computational overhead remains acceptable. Figure 6(c) further demonstrates that VLPO introduces little overhead compared to GRPO.

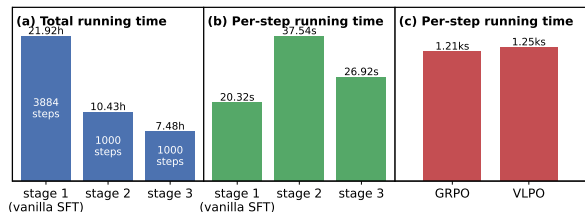


Figure 6. (a) Total running time of Monet SFT stage 1, 2, and 3. (b) Per-step running time of Monet SFT stage 1, 2, and 3. (c) Per-step running time of GRPO and VLPO.

## E. Comparison with “Think-with-images” Approaches

In this section, we will compare Monet and existing “Think-with-images” methods from two perspectives: the cost of obtaining training data and the interpretability of the reasoning process.

### E.1. The Cost of Obtaining Training Data

In this section, we compare the construction of the training data of Monet and that of the existing tool-based approaches from the following aspects: annotation cost, data scalability, and automation level.

**Annotation Cost.** The annotation cost of Monet is **comparable** to that of tool-based and code-based approaches, as all of them require expert models to generate tool-invocation instructions or code to obtain auxiliary images.

**Data Scalability.** Monet exhibits better data scalability than tool-based and code-based methods for two reasons.

**1) Monet does not need to learn specific tool-invocation formats or code libraries.** Instead, it only relies on auxiliary images, which allows it to leverage a large amount of heterogeneous interleaved image–text CoT training data, greatly expanding the pool of available data. In contrast, tool-based methods require training data with a unified invocation format for specific tools, which severely limits data scalability. **2) Scaling up data with simple visual operations is sufficient to incentivize strong OOD generalization of Monet.** In Monet-SFT-125K, 118.6K samples involve only simple visual operations (cropping key regions), while only a small fraction contains more complex operations (e.g., drawing auxiliary lines or sketching). Therefore, it’s easy for Monet to scale up the training data by incorporating simple visual operations. In contrast, tool-based and code-based methods must generate dedicated training data for each visual operation in order for the model to acquire the corresponding capability.

**Automation Level.** The data curation process of Monet can be fully automated by MLLMs.

## E.2. Interpretability of the Reasoning Process

**Monet demonstrates stronger interpretability than pure latent reasoning methods, though it remains less interpretable than “think-with-images” approaches.** Monet adopts an interleaved text–latent reasoning paradigm, where it’s likely that the latent embeddings are followed by explicit observations derived from the latent embeddings. Unlike other visual latent reasoning approaches (e.g., LVR and Mirage), which replace the entire intermediate reasoning process with latent representations, Monet does not fully hide intermediate reasoning in the latent space and therefore retains good interpretability.

Take the chart understanding problem in Figure 11 in the Appendix as an example, the Monet-7B model generates an intermediate conclusion (highlighted in **bold**) immediately after the latent embeddings, which enables us to check the faithfulness of the latent thinking process.

### Example of the interpretability of Monet

```
“...</latent>The highlighted area in the image clearly shows the Top Sales Countries section...The final answer is D”
```

The “think-with-images” methods typically yield explicit tool-use outputs, such as intermediate images or text, which makes the correctness of the intermediate reasoning process easy to examine. By contrast, although Monet allows assessment of process correctness through its intermediate language reasoning component, the faithfulness latent

embeddings themselves cannot be easily verified for a specific reasoning trajectory<sup>2</sup>.

## F. Limitations

We summarize the limitations of our work as follows. First, Monet relies on a multi-stage SFT pipeline, resulting in increased training complexity and overhead. Second, the impact of different reward designs on latent visual reasoning in MLLMs has not yet been systematically explored, and thus remains an important avenue for future research. Finally, as discussed in E.2, the interpretability of the generated latent embeddings is still limited. However, this is a fundamental limitation of all latent reasoning methods, and we hope future work can develop more effective ways to probe, validate, and improve the faithfulness of latent reasoning processes.

## G. Case Studies

### G.1. Monet-SFT-125K Examples

We select three examples from Monet-SFT-125K to illustrate the data structure and corresponding visual operations. These examples cover three representative types of visual operations: cropping key regions, generating new visual states, and drawing auxiliary lines or bounding boxes. As shown in Figures 7, 8, and 9, each training instance contains an image–text interleaved CoT with intermediate auxiliary images that facilitate the visual reasoning process. Additionally, tokens that encode key observations from the auxiliary images are wrapped in `<observation>` and `</observation>`, which are used in the alignment loss during SFT Stage 2.

### G.2. Inference Examples

We present some inference examples generated by Monet-7B in this section to show the generality of tasks that Monet-7B can solve. For clarity, we don’t map the generated latent embeddings to the language space, as this will produce meaningless tokens. Instead, we use `<latent><latent_embeddings></latent>` to represent the latent reasoning segments.

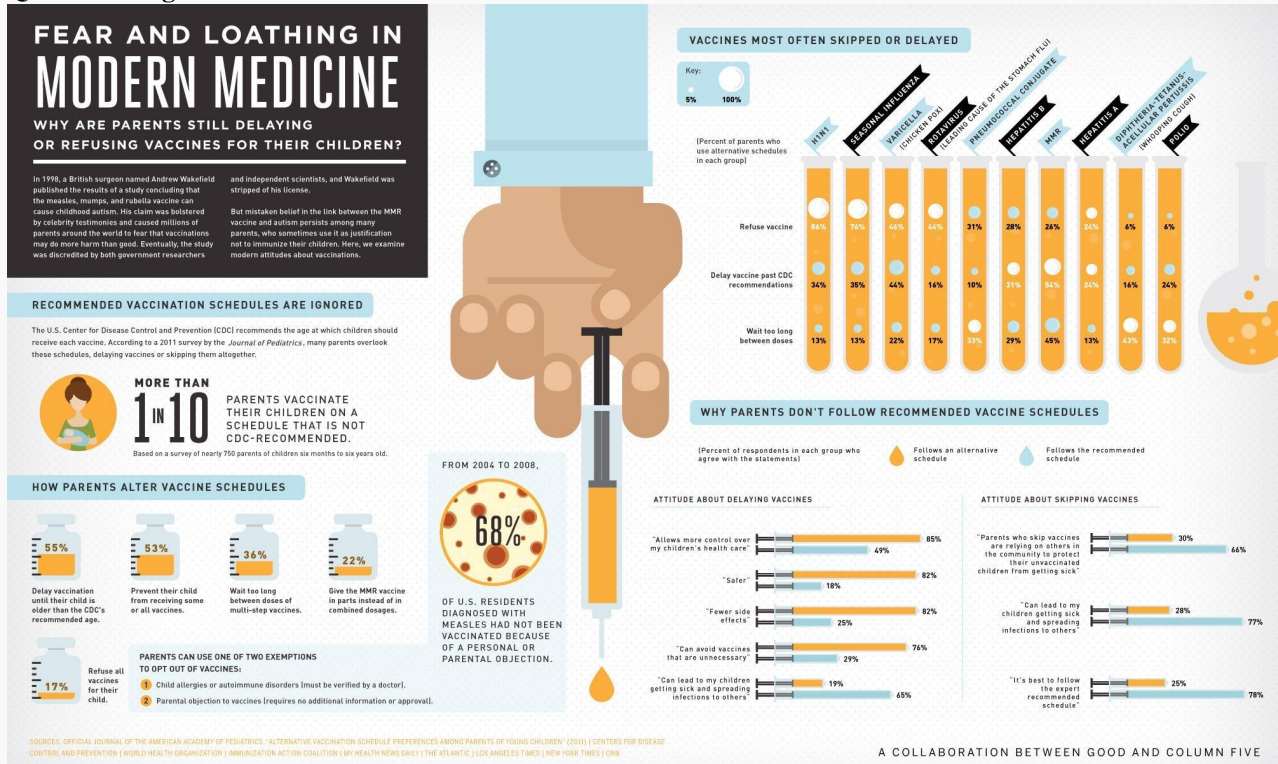
The tasks of the selected examples include 3D Spatial Reasoning (Figure 10), 2D transformation (Figure 11), complex diagram reasoning (Figure 12), commonsense QA (Figure 13), fine-grained OCR (Figure 14), and math reasoning (Figure 15). Note that Monet does not always use latent thinking. For example, for the pure-text math problem in Figure 15, Monet directly relies on text-based reasoning to solve the problem.

<sup>2</sup>Still, we can validate the effectiveness of latent embeddings by removing them from the reasoning process and measuring the change in performance (see Figure 5).

Monet-SFT-125K Example 1: Cropping the Crucial Region (from Zebra-CoT Visual Search)

**Question:** According to the survey presented in the infographic, what percentage of parents in the U.S. delay the Varicella vaccine for their children beyond the CDC's recommendations?  
 Put your final answer within `\boxed{ }`.

**Question Image:**



**Image-text Interleaved CoT:**

The question asks for the percentage of parents who delay the Varicella vaccine according to CDC recommendations, based on the provided infographic. I need to find the section of the infographic that lists vaccines and the percentages for delaying them past CDC recommendations. I will focus on the Varicella vaccine data.



The cropped image shows the specific data point for the <observation>Varicella vaccine</observation> under the Delay vaccine past CDC recommendations row. Therefore, the final answer is `\boxed{44\%}`.

Figure 7. Example of Monet-SFT-125K: cropping the crucial region.

## Monet-SFT-125K Example 2: Creating New Visual States (from Zebra-CoT 3D Object Counting)

**Question:** Consider the initial arrangement of objects presented in the image below. You are required to perform a series of removal operations in the specified order:

Step 1: Eliminate all objects identified as PYRAMIDS.

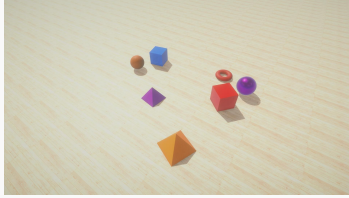
Step 2: Eliminate all objects that are BLUE in color.

Step 3: Eliminate all objects that are RED in color.

After completing these steps, determine the total count of orange objects remaining.

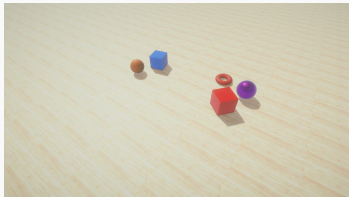
Put your final answer within `\\boxed{}`.”

### Question Image:

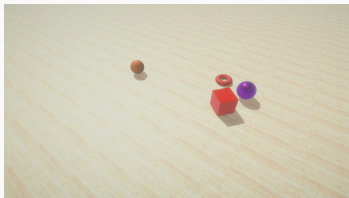


### Image-text Interleaved CoT:

I will begin by analyzing the initial set of objects shown in the problem image. There are `<observation>7</observation>` distinct objects in the initial configuration. My approach will be to systematically apply each removal step and visually track the remaining objects before arriving at the final count. The first operation requires the removal of all pyramid-shaped objects from the scene. This means both the `<observation>purple pyramid</observation>` and the `<observation>orange pyramid</observation>` will be taken away. To visualize the updated set of objects, I will generate an image reflecting this change.



After removing the pyramids, as seen in the previous visual, we are left with `<observation>five</observation>` objects: an `<observation>orange sphere</observation>`, a `<observation>blue cube</observation>`, a `<observation>red torus</observation>`, a `<observation>red cube</observation>`, and a `<observation>purple sphere</observation>`. The next instruction is to remove all objects that are blue. This specifically targets the `<observation>blue cube</observation>`. I will generate a new visual to show the state after this removal.



...

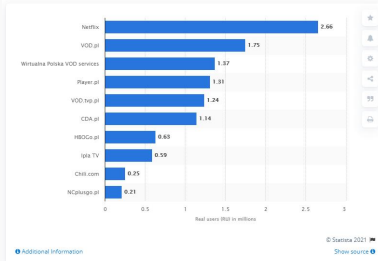
Having completed all specified removal operations, the final visual shows `<observation>two</observation>` objects remaining: an `<observation>orange sphere</observation>` and a `<observation>purple sphere</observation>`. The question asks for the count of orange objects. From the two remaining objects, only one is orange. The final answer is `\\boxed{1}`.

Figure 8. Example of Monet-SFT-125K: creating new visual states.

### Monet-SFT-125K Example 3: Drawing Auxiliary Lines and Bounding Boxes (from CogCoM)

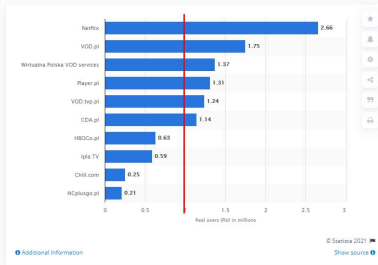
**Question:**What is the sum of real users below 1 million?  
Put your final answer within `\boxed{ }`.”

#### Question Image:

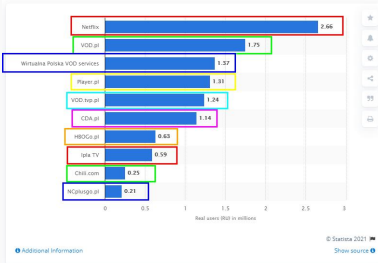


#### Image-text Interleaved CoT:

Draw a line to obtain the new image after the line is drawn.



Outline each bar column, which are located at Netflix, VoD.pl, Wirtualna Polska VOD services, Player.pl, VoD.tvp.pl, CDA.pl, HBOGo.pl, lpla TV, Chili.com, NCplusgo.pl.



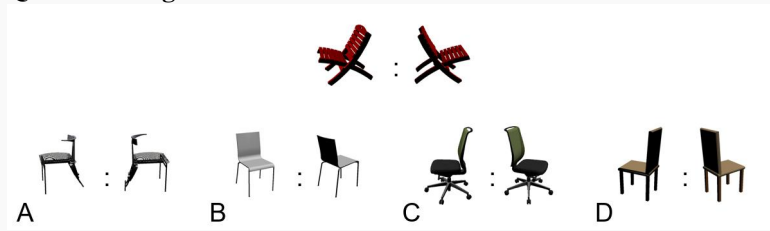
Analyze the value of each bar in relation to  $x=1$ . The values obtained from HBOGo.pl, lpla TV, Chili.com, and NCplusgo.pl are all *less than one million*, with respective values of *0.63, 0.59, 0.25, and 0.21*. Calculate their total sum as  $0.63 + 0.59 + 0.25 + 0.21 = 1.68$ , so the answer is  $\boxed{1.68}$ .

Figure 9. Example of Monet-SFT-125K: drawing auxiliary lines and bounding boxes.

### Monet-7B Inference Example 1: 3D Spatial Reasoning (from VisualPuzzles)

**Question:** Given a pair of chairs arranged at a specific angle, select the option from the four provided that matches the angle of the chairs in the question. Options: Choose from (A) (B) (C) (D) in the image. ”

**Question Image:**



**Response by Monet-7B:**

The problem asks to identify the option that matches the given angle between the chairs. I will analyze the provided image to determine the angle between the chairs and then compare it with the given choices. To accurately assess the angle between the chairs, I will focus on the visual representation of the chairs in the problem image. `<latent><latent_embeddings></latent>` The zoomed-in view of the chairs in the problem image clearly shows the angle between them. This visual confirms the specific spatial relationship between the two chairs. Therefore, the final answer is .

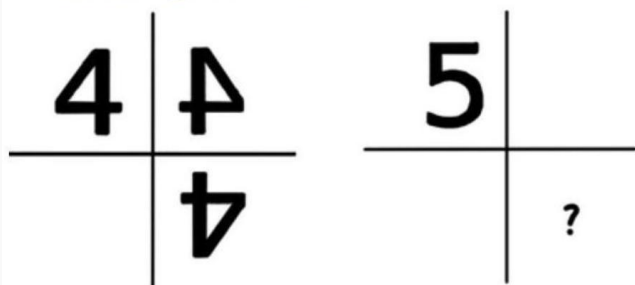
Figure 10. **Inference example: 3D spatial reasoning.** Instead of describing the angles between the chairs in language, Monet directly reasons with latent embeddings before giving the final answer.

Monet-7B Inference Example 2: 2D Transformation (from VisualPuzzles)

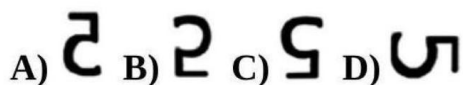
**Question:** The number 4 is reflected twice in <image1>. Which among the options in <image2> should appear in the field with the question mark if we do the same with the number 5? ”

Question Image:

<image1>:



<image2>:



**Response by Monet-7B:**

To determine the correct answer, we need to identify the pattern of reflection for the number 5, similar to how the number 4 is reflected in the provided image. To better understand the reflection pattern, I will generate a zoomed-in view of the relevant section of the image. <latent><latent\_embeddings></latent>The zoomed-in image clearly shows how number 4 is reflected. The reflection pattern for the number 4 is obtained by flipping it twice. Once horizontally, once vertically. Therefore, the final answer is .

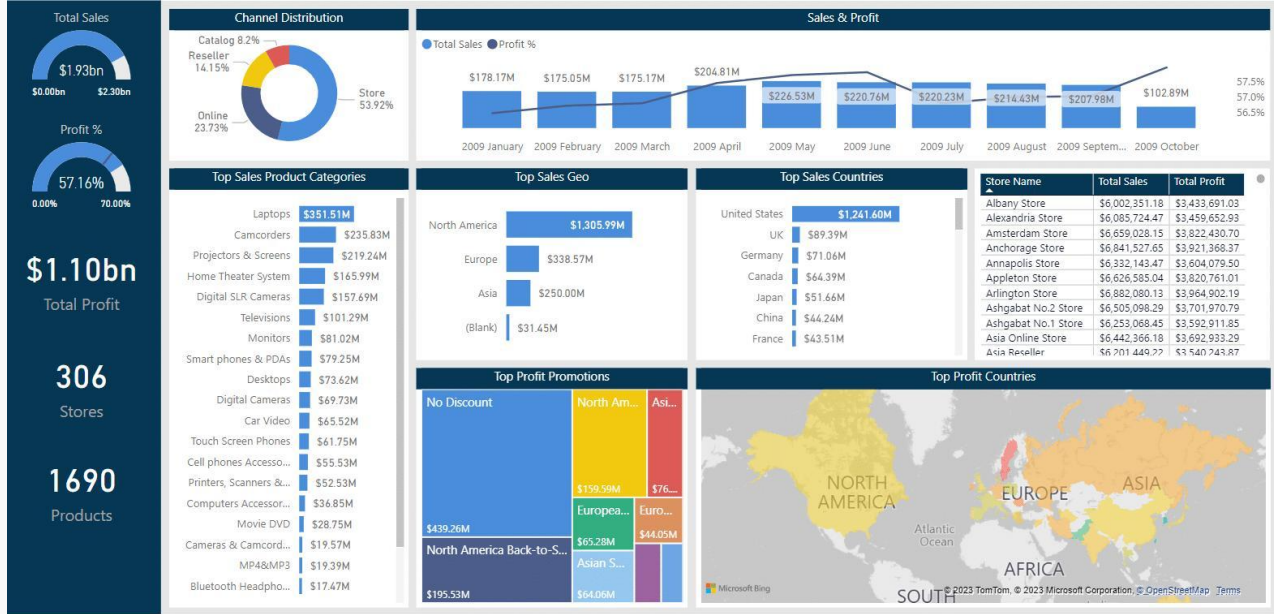
Figure 11. **Inference example: 2D transformation.** By generating latent embeddings, Monet successfully identifies the flipping rule of the number.

**Monet-7B Inference Example 3: Complex Diagram Reasoning (from MME-RealWorld-Lite)**

**Question:** Which country is the top sales country? The choices are listed below:

- (A) Canada
- (B) Germany
- (C) UK
- (D) United States
- (E) The image does not feature the number.

**Question Image:**



**Response by Monet-7B:**

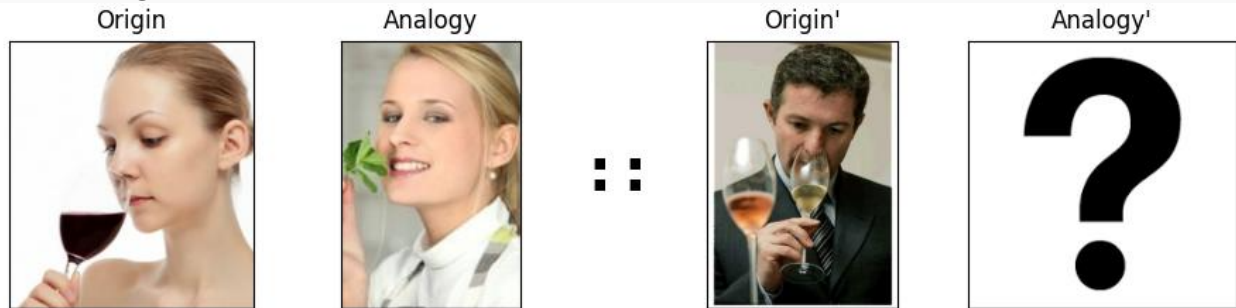
To determine the top sales country, I need to locate the section of the image that displays sales by country. <latent><latent\_embeddings></latent>The highlighted area in the image clearly shows the "Top Sales Countries" section. It lists various countries along with their respective sales figures. The sales figures are in millions of dollars. From the highlighted section, I can see that the United States has the highest sales at 1,241.60M, followed by the UK, Germany, and other countries. This directly answers the question about the top sales country. The final answer is D.

Figure 12. **Inference example: complex diagram reasoning.** Monet-7B exhibit a **hierarchical** reasoning pattern. Firstlt, it focus on the relevant section in the image by generating latent embeddings: "The highlighted area in the image clearly shows the "Top Sales Countries" section." Then, it accurately identifies the contents in the "Top Sales Countries" section and gives the correct answer.

Monet-7B Inference Example 4: Commonsense QA (from VisualPuzzles)

**Question:** Based on the relationship between Origin and Analogy, select the option from the candidates that best completes this analogy.  
Options: Choose from (A) (B) (C) (D) in the image.

**Question Image:**



Candidates



**Response by Monet-7B:**

To identify the correct image that completes the analogy, I will analyze the relationship between the "Origin" and "Analogy" images in the provided problem image. The goal is to find a similar relationship between the "Origin'" image and the missing "Analogy'" image. To better understand the relationship, I will focus on the "Origin" and "Analogy" images to identify the specific action being performed by the person in each image.

<latent><latent\_embeddings></latent>

The zoomed-in view of the "Origin" image clearly shows a person drinking wine, while the "Analogy" image depicts a person smelling a flower. This indicates that the action involves the sense of smell. Therefore, the final answer is D.

Figure 13. **Inference example: commonsense QA.** Monet-7B correctly identifies the connection between the "Origin" and the "Analogy" image, showing its commonsense reasoning capability.

# Monet-7B Inference Example 5: Fine-grained OCR (from MME-RealWorld-Lite)

**Question:** Where does Joseph work? The choices are listed below:

- (A) Boston.
- (B) Miami.
- (C) Chicago.
- (D) New York.
- (E) The image does not feature the related information.

**Question Image:**

**The Terrible Toll Of Art Anxiety**

From Page 1. This Section now than there used to be, and that their prices are so much higher — at least for the moment — exacerbates the problem. “Four or five years ago, Chelsea seemed like maybe 30 galleries,” said Bernard Lampton, a producer at MTV News, who began collecting around that time. “Now it’s more like 300. Why would you want to jump in now, when it feels like you’re going to drown?”

Joseph Higgins, a 43-year-old portfolio manager in New York with a \$800,000 mansion in west Setao and a house in the suburbs, is one of the rare sufferers who will speak openly about his art paralysis. He blames it on galleries, and overcame it, he said, by breaking from their grasp.

“You’re going into an intimidating space and having a curator or a gallery owner ask you ‘Do you like this style or this art’ when you have no idea what the price tag is,” he said. “It’s hard to say, I’m browsing, after someone spends time with you in a gallery and tells you ‘I’ll put it under a light for you’ and sets you up in a little room and brings you a cognac cocktail.”

Mr. Higgins started out by using paintingdirect.com, a Web site that sells the work of hundreds of artists from around the world in categories

“I walk into those snooty galleries in Chelsea and feel like I don’t belong.”

...ranging from landscape to “fantasy.” He has bought 14 paintings there and has little patience with those who would sneer at such a site. New York may be a world capital of modern and contemporary art, he said, but he finds the same “edginess” online that he does in the galleries of Chelsea, at much lower prices.

Mitchell Gold, the co-founder of Mitchell Gold & Bob Williams furniture, shares Mr. Higgins’s aversion. “I can’t stand going into galleries,” he said. “They don’t put prices on, you get all worked up, you don’t know the price is \$20,000 and you think, Gee, I don’t want to spend that.” (His own stores sell photographs by Tipper Gore, for \$750 to \$2,500, and the prices, he pointed out, are right beside the work.)

James Dantzig, who runs the Danger Project gallery in Chelsea, which specializes in photography, recently took a walk in the neighborhood and saw something he liked. When he asked to see a price list he was told there wasn’t one, which happens to be a violation of the city’s Truth in Pricing Law.

That, Mr. Dantzig said, is a tactic you will find in most of the upper-end Chelsea galleries. “They don’t put out price lists,” he said. “I said, ‘Can I ask the price of this work?’ They said it was \$2,500. I said, ‘Oh, we’re dealing in pounds now?’ It was a German artist based in London, but I think if you were a nervous Nellie and asked for a price, and they said no, and finally they gave you a price in pounds, you’d have every reason not to buy.”

Art paralysis takes many forms. In addition to the would-be buyers who are intimidated by galleries, there are those worried about making an unfashionable

choice: those obsessed with investment value and those who return to a gallery for months, even years, never buying a thing. (Some of these suffer from a form of art paralysis that Stephen Norflingen, the president of the Foundry Gallery in Washington, calls red dot syndrome — a desperate longing only for those pieces bearing the red dots that show they’ve been sold.) And then there are the people whose reasons make no sense at all, at least to those doing the selling.

“I had a whole series of these great chocolate, 10-by-10 inches on boards, \$600 a piece,” said Armand Bolling, the director of Karen Lyme Galleries, of Boca Raton, Fla., and Beverly Hills, Calif., which specializes in large-scale figurative oil paintings. “I held it up on the wall and this lady says, ‘It makes me depressed.’ I’m like, ‘Is this lady out of her mind? It’s a piece of chocolate.’”

Mr. Bolling warmed to his theme. “You go in a house, you see 16 empty walls and two pieces, and one they probably got at an auction for Crohn’s or colitis for \$500, and the other was given to them. These people aren’t buying art.”

Amy Marx, a Maryland painter whose work features images of violent storms and tornadoes, and who currently has a show at OK Harris in SoHo, where her large pieces are listed at between \$8,500 and \$12,500, remembers protracted discussions with a family in Maryland a few years ago. “I stretched all these paintings specifically for them,” Ms. Marx said. “He — the man of the house — said they love them, they want to buy them. Then he called up the next day and said, ‘Well, since 9/11, you know.’ We’re not talking Manhattan, we’re talking Prince George’s County in Maryland; nothing happened.”

One major reason for art paralysis, many experts say, is the feeling that when buying art you are exposing yourself to ridicule. “Art has always been a barometer of class,” said Jonathan Santolof, a Chelsea painter and novelist. “If you buy the wrong thing and people come to your house, you’re screwed.”

Are New Yorkers really that critical? Mr. Santolof talks a story that sounds like the beginning of an art joke: an artist (himself), an art writer and a dealer visit the home of a collector. After they leave — the period of time that is the

most fraught with danger for New York hosts — the guests tell each other what they really think. “The art writer said, ‘Now there’s an obvious collection,’” he said. “I said, ‘I could forgive that if the examples he bought were better.’ The art dealer said he liked 1 out of 40 — and he would totally forgive that if he had sold them the art.”

Even people who have been serious collectors in the past may suffer bouts of paralysis, particularly if they’re looking at a large investment. Melissa Fink and her husband, Steven Fink, live in Bedford, N.Y., in “a fabulous, contemporary, 6,000-square-foot house,” as Ms. Fink described it. “We just spent a fortune doing additions on this fabulous

house, it’s huge and beautiful and kind of like we have all these empty spaces.” They’ve worked with art consultants in the past, spending about \$200,000, and are prepared to spend another \$100,000, according to Ms. Fink.

“We had our art consultant come two different times,” Ms. Fink said. “When it comes to artwork, I don’t know that I know the value. Are you buying it for enjoyment, or are you buying it for investment? If you buy some Yves Saint Laurent or Balmain, you know that you are getting value.”

Did she have a bad experience? Yes, she did.

“I was getting a manicure at Canyon Ranch, it was the end of seven days of manicures and hiking and there was this piece of art behind me and I said, ‘It’s so lovely, I’ll take it.’”

She bought it impulsively without telling her husband and had it shipped. The moment it arrived, they agreed it was awful. It now lives in a storage room in the fabulous contemporary home. Cost of the painting: \$12,000. Lesson, as explained by the art consultant: “Whatever you do in your life, if you want to buy art on vacation, please call us and

**HOW TO KNOW**  
**WHAT YOU LIKE**  
Joseph Higgins, above, a portfolio manager in New York who was intimidated by art galleries, began buying art online. Jonathan Arnold, left, feared making a mistake until he began working with Marcia Levine, an art consultant.

let us do some research.”

Jonathan Arnold, a 45-year-old Chicago economist, also turned to a consultant for help. He dabbled with collecting in his early 20s, he said, and last year thought about starting again, but feared making a financial mistake. “If you don’t know what you’re doing,” he said, “it’s an invitation to be taken down.”

Mr. Arnold asked friends whose collections he admired how they acquired their work, and came up with the name of an art consultant, Marcia G. Levine, who runs Fine Art in Manhattan, who once sold a Modigliani to the Vatican Museums and creates collections for people around the world.

Her commission, should you be new to this world, comes from the gallery, unless she is on retainer from a client. If you’re not prepared to make a commitment to creating a collection — which, for a young person, can start with a \$5,000 drawing — don’t bother her. Ms. Levine arranged for Mr. Arnold to go to the Marborough gallery in New York before an opening to see the work of an American artist named Michael Anderson. Mr. Arnold was especially taken with “Strange Party,” a collage created from torn bits of street posters. He bought it for \$20,000. He also saw the work of another artist, Steven Charas, and liked it so much that he will be buying a piece that has not yet been finished, for a price he does not know.

“Having just bought a piece that wasn’t even hung, I was seduced into buying a piece that wasn’t even made,” Mr. Arnold said, laughing. His art anxieties, you might say, were over.

**Julie Scelzo**

## Response by Monet-7B:

To determine where Joseph works, I need to locate his name or relevant information within the article. I will zoom in on the section of the article that mentions Joseph to clearly identify his location.

<latent><latent\_embeddings></latent>

The zoomed-in image clearly shows that Joseph Higgins is a portfolio manager in New York. This directly answers the question about where Joseph works. Therefore, the final answer is **D**.

Figure 14. Inference example: fine-grained OCR. Monet-7B accurately identifies the key information, which is located in the middle of the rightmost region of the image.

### Monet-7B Inference Example 6: Math Reasoning (from VisualPuzzles)

**Question:** Every night the wizard Tilim makes the weather forecast for the king. When Tilim gets it right he gets 3 gold coins, but when he makes a mistake, he pays a fine of 2 gold coins. After making the prediction for 5 days, Tilim did the math and discovered that he neither won nor lost coins. How many times did he get the weather forecast right in those 5 days?

Options:

- (A) 1
- (B) 2
- (C) 3
- (D) 4

#### Response by Monet-7B:

To determine the number of correct weather forecasts, we can set up an equation based on the given conditions. Let's define:

-  $x$  as the number of correct forecasts (where he earns 3 gold coins per correct forecast).

-  $y$  as the number of incorrect forecasts (where he loses 2 gold coins per incorrect forecast)

The wizard made a total of 5 forecasts, so:

$$x + y = 5$$

He neither won nor lost any gold, which means the total gain from correct forecasts equals the total loss from incorrect forecasts:

$$3x = 2y$$

We now have a system of two equations:

1.  $x + y = 5$

2.  $3x = 2y$

We can solve this system by expressing  $y$  in terms of  $x$  from the first equation:

$$y = 5 - x$$

Next, substitute  $y = 5 - x$  into the second equation:

$$3x = 2(5 - x)$$

Solve for  $x$ :

$$3x = 10 - 2x$$

$$3x + 2x = 10$$

$$5x = 10$$

$$x = 2$$

So, Tilim made 2 correct forecasts. To verify, we calculate  $y$ :

$$y = 5 - x = 5 - 2 = 3$$

Now, check the total gold:

- Correct forecasts:  $2 \times 3 = 6$  gold

- Incorrect forecasts:  $3 \times (-2) = -6$  gold

- Total gold:  $6 + (-6) = 0$  gold. The calculations confirm that Tilim neither won nor lost any gold. Therefore, the number of correct forecasts is  $\boxed{2}$ . The final answer is  $\boxed{B}$ .

Figure 15. **Inference example: math reasoning.** Monet-7B can perform pure-text math reasoning. Since the problem is not visual-related, it doesn't activate latent thinking mode.