SynerGen-VL: Towards Synergistic Image Understanding and Generation with Vision Experts and Token Folding

Supplementary Material

A. Ablation Study

In this section, we ablate the effectiveness of the two important techniques of SynerGen-VL, *i.e.*, *token folding* and *progressive alignment pre-training with MMoEs*. In this ablation study, we use Qwen2-0.5B-Instruct [4] as the initialized LLM and image size 256 unless otherwise specified.

A.1. Effectiveness of Token Folding

To verify the effectiveness of token folding on high-resolution image understanding, we compare SynerGen-VL with the baseline version without token folding and the dynamic resolution strategy on image understanding tasks. Specifically, the baseline model directly use the tokenized sequence as the input image sequence without token folding, where the input image size is 256×256 and the tokenized sequence length is 1024. Meanwhile, for the model with token folding, we follow InternVL-1.5 [1] to implement the dynamic resolution strategy as stated in the paper to provide high-resolution input images. For fair comparison, we use a token folding ratio of 2×4 and control the maximum number of dynamic image patches so that the average length of image token sequence after token folding is also 1024.

We train the models with a subset of stage 2 (S.2) understanding data, and evaluate the pre-trained models on VQA benchmarks. Results are shown in Tab. 1. On datasets requiring precise understanding of detailed image information such as TextVQA, DocVQA, ChartVQA, and InfographicVQA, the model with token folding achieves significantly better results, demonstrating its advantages of high-resolution image understanding.

Model	TextVQA	GQA	DocVQA	AI2D	ChartQA	InfoVQA
w/o token folding	18.7	45.3	14.7	42.0	20.9	18.7
w/ token folding	35.0	45.1	36.7	42.1	49.7	21.1

Table 1. **Comparison between models with and without token-folding on VQA benchmarks.** The model with token folding demonstrates significant performance improvements with the same image token sequence length.

A.2. Effectiveness of the Progressive Alignment Pre-training with MMoEs

We ablate our proposed visual alignment pre-training strategy on various benchmarks, including visual question answering (VQA), natural language processing (NLP) and text-to-image (T2I) generation, as shown in Tab. 2. To ensure fair comparison, neither token folding nor dynamic resolution strategies are employed. For experimental efficiency, only 1/6 of the training data is used for both stages.

The results show that our progressive strategy matches or exceeds the fully parameter-trained strategy on VQA benchmarks and significantly outperforms it on text-to-image generation benchmarks. Meanwhile, on NLP benchmarks, our model with progressive alignment pre-training delivers results much closer to the pre-trained LLM (Qwen2-0.5B-Instruct) compared with the fully parameter-trained model. This validates that our approach effectively preserves the original knowledge in the pre-trained LLM while learning robust visual representations. Furthermore, the two-stage training strategy outperforms training solely with stage 1 or stage 2, particularly on VQA and text-to-image generation benchmarks. This underscores the importance of learning basic visual concepts and pixel dependencies from large-scale noisy data, as well as enhancing image-text alignment and image aesthetics with high-quality data.

B. Analysis of Relationship Between Image Generation and Understanding

We provide visualization and analysis to understand the relationship between image generation and understanding tasks, *i.e.* how the two tasks might be related in terms of their processing or feature utilization.

B.1. Image Feature Similarity

We first analyze whether the two tasks share similar representations. We use the same input image paired with text instructions of generation or understanding, and compute the cosine similarity between visual features of the two tasks at each layer. As shown in Fig. 1, the two features are nearly identical (0.999) at shallower layers, but the similarity decreases as layers deepen. It finally reaches a near-zero value (0.035) at the last layer, suggesting that the two representations are disentangled. This observation implies that while image generation and understanding may share foundational visual representations in the early stages, they develop task-specific representations based on different instructions of image generation and understanding at deeper layers.

Stage	Strategy	VQA Benchmarks ↑						NLP Benchmarks ↑				T2I Benchmark ↓
		TextVQA	GQA	DocVQA	AI2D	ChartQA	InfoVQA	MMLU	CMMLU	AGIEVAL	MATH	MSCOCO
Baseline (Qwen2-0.5B)		-	-	-	-	-	-	42.3	51.4	29.3	12.1	-
S.1 + S.2	Full	14.3	42.9	11.3	24.7	12.4	12.6	23.1	23.0	8.1	0.9	30.7
S.1 only	Progressive	0.1	13.0	0.2	0.3	0.0	0.0	42.3	51.4	29.3	12.1	28.3
S.2 only	Progressive	8.7	36.9	8.6	40.9	11.7	16.2	37.6	45.3	28.9	7.2	34.9
S.1 + S.2	Progressive	13.2	41.2	11.4	41.9	12.8	17.0	39.3	48.2	26.2	8.9	20.2

Table 2. **Zero-shot performance of different pre-training strategies.** "S.1" and "S.2" denote the first and second pre-training stage. "Full" and "Progressive" denote the full parameter tuning and our progressive tuning strategy with MMoEs, respectively. FID [2] is reported for text-to-image generation (T2I) on MSCOCO [3].

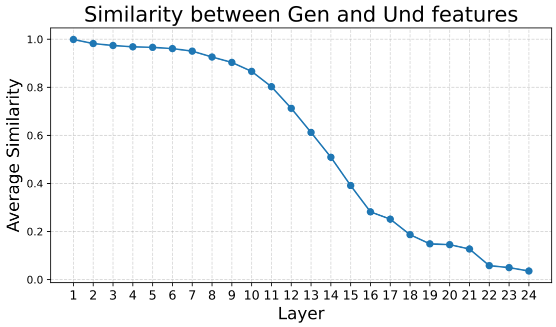


Figure 1. **Cosine similarity of visual features between generation and understanding tasks across different layers**. The representations of the image understanding and generation tasks are similar in shallow layers but disentagle in deeper layers.

B.2. Attention Map Visualization

In Fig. 2, we further investigate whether the two tasks have similar attention map patterns. We discover that in both tasks, locality is present at early layers, where visual tokens only attend to its nearby tokens (*i.e.* near the diagonal). Text tokens and images have few interactions with each other. As layers deepen, longer dependency is observed, and finally global interactions are achieved at the last layer. Text and image also interact more often than at shallower layers. The attention weight also displays a periodicity nature, such as in Layer 4. Visualization in the input image suggests that the period is the number of tokens in each row, validating the locality. When comparing the attention maps in the two tasks, we observe that locality is more obvious in generation than in understanding at the same layer. This can be explained that local details are required to generate a spatially consistent and semantically coherent image, while understanding the whole image requires global context.

C. Detailed Training Configurations

More detailed hyper-parameters used in the training stages are listed in Tab. 3.

D. Visualization

For qualitative evaluation, we visualize examples for image understanding and image generation as follows.

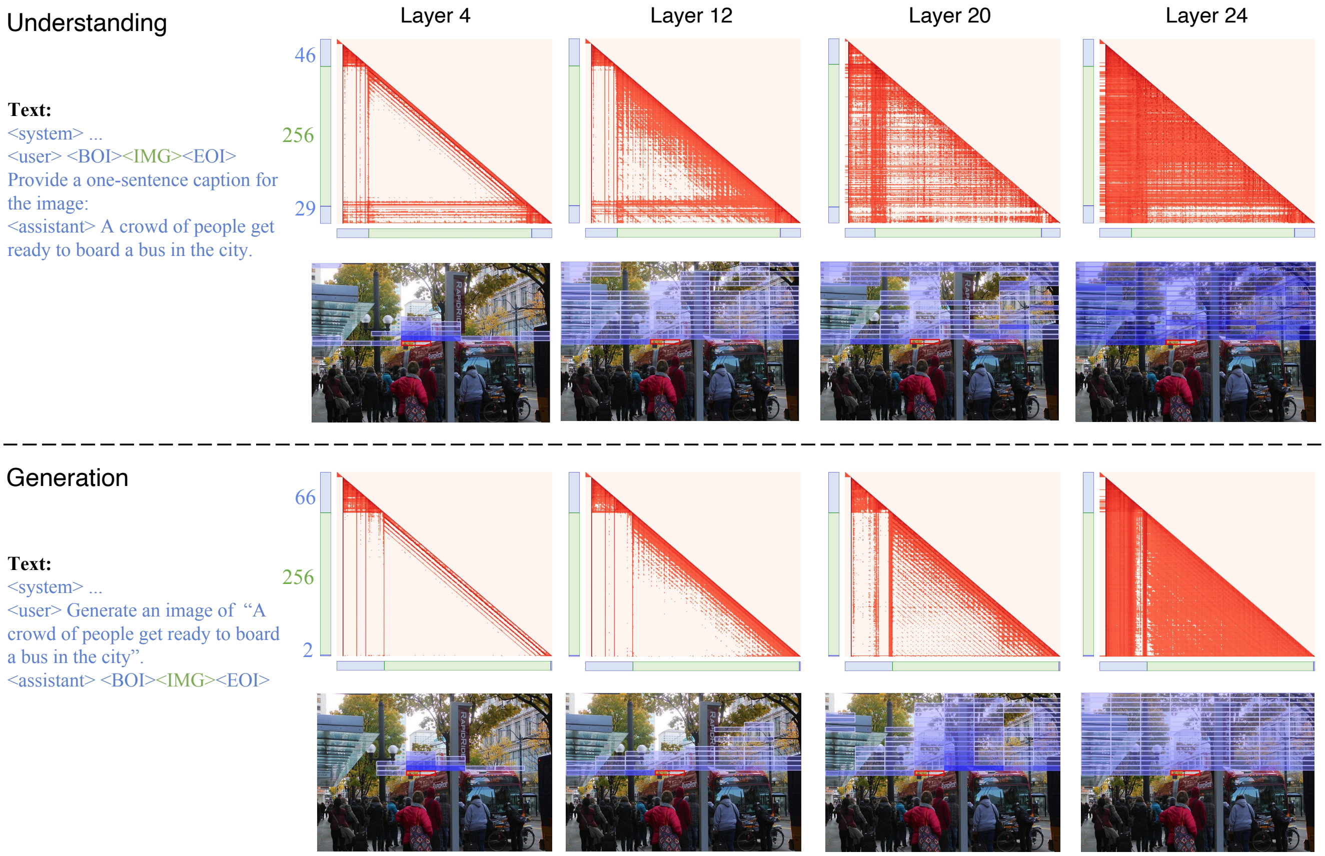
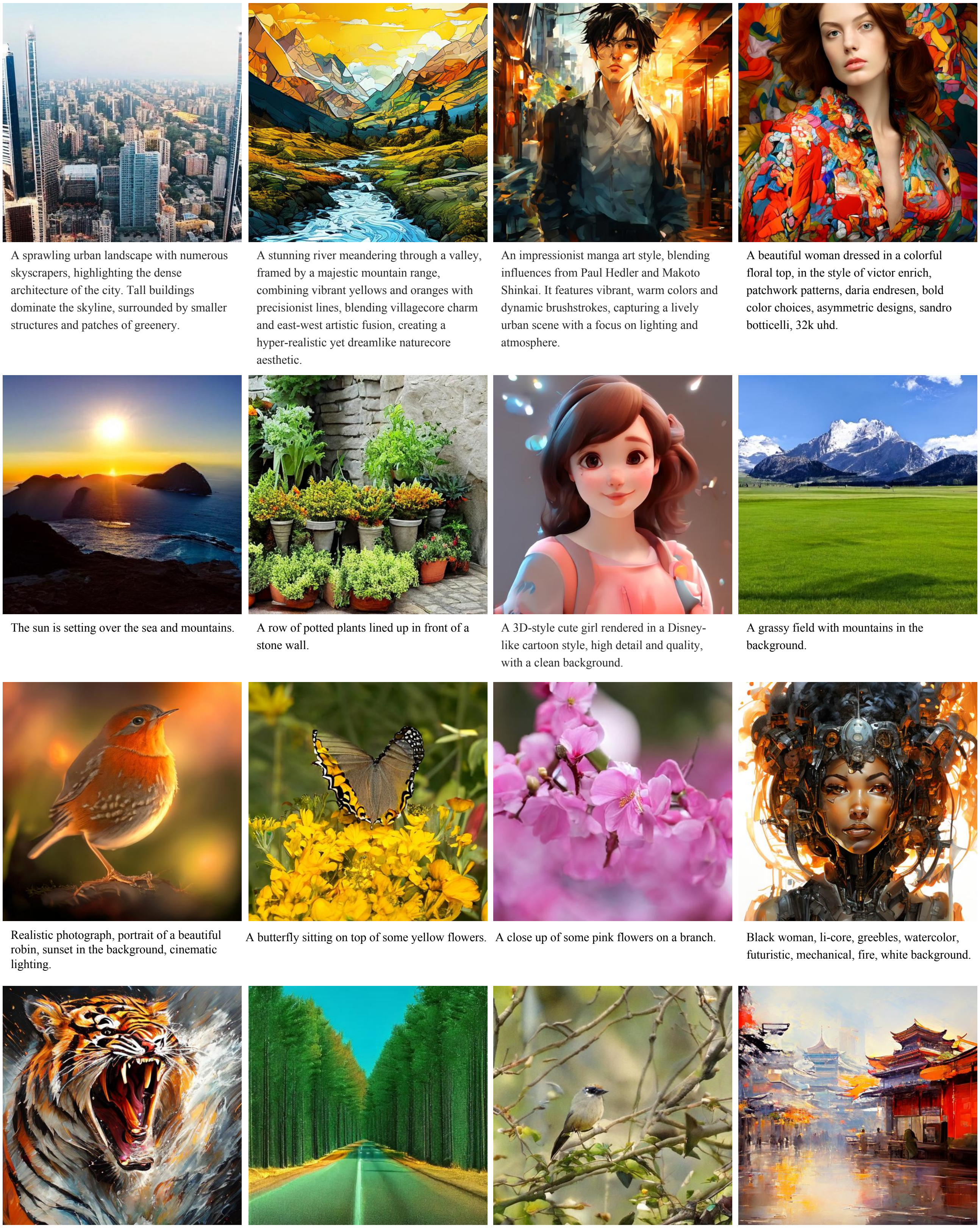


Figure 2. **Attention map visualization of understanding and generation tasks.** In the second and fourth rows, we visualize a query token (red) and its attended tokens (blue) in the input image. Each token corresponds to a horizontal rectangular area in the original image due to the 2 × 4 token folding. Darker blue indicates larger attention weights.

Configuration	Alignment Pre-training		Instruction
	S.1	S.2	Tuning
Maximum number of image tiles	1	6	12
LLM sequence length	$4,096$	$8,192$	$16,384$
Use thumbnail	✗	✓	✓
Global batch size (per-task)	$6,988$	$5,090$	$1,760$
Peak learning rate	$1e^{-4}$	$5e^{-5}$	$5e^{-5}$
Learning rate schedule	constant with warm-up	cosine decay	cosine decay
Weight decay	0.05	0.05	0.01
Training steps	95k	35k	12k
Warm-up steps	200		
Optimizer	AdamW		
Optimizer hyperparameters	$\beta_1 = 0.9, \beta_2 = 0.95, eps = 1e^{-8}$		
Gradient accumulation	1		
Numerical precision	`bfloat16`		

Table 3. **Hyper-parameters used in the alignment pre-training and instruction tuning stages.**



A sprawling urban landscape with numerous skyscrapers, highlighting the dense architecture of the city. Tall buildings dominate the skyline, surrounded by smaller structures and patches of greenery.

A stunning river meandering through a valley, framed by a majestic mountain range, combining vibrant yellows and oranges with precisionist lines, blending villagecore charm and east-west artistic fusion, creating a hyper-realistic yet dreamlike naturecore aesthetic.

An impressionist manga art style, blending influences from Paul Hedler and Makoto Shinkai. It features vibrant, warm colors and dynamic brushstrokes, capturing a lively urban scene with a focus on lighting and atmosphere.

A beautiful woman dressed in a colorful floral top, in the style of victor enrich, patchwork patterns, daria endresen, bold color choices, asymmetric designs, sandro botticelli, 32k uhd.

The sun is setting over the sea and mountains.

A row of potted plants lined up in front of a stone wall.

A 3D-style cute girl rendered in a Disney-like cartoon style, high detail and quality, with a clean background.

A grassy field with mountains in the background.

Realistic photograph, portrait of a beautiful robin, sunset in the background, cinematic lighting.

A butterfly sitting on top of some yellow flowers.

A close up of some pink flowers on a branch.

Black woman, li-core, greebles, watercolor, futuristic, mechanical, fire, white background.

Illustration paintings roaring tiger of successful unique wall paintings.

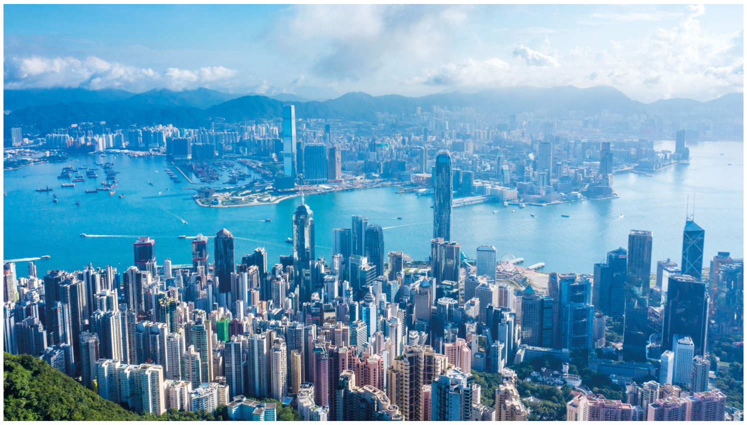
An empty road with trees in the background.

A small bird sitting on a branch of a tree.

Impressionist painting, Beijing, vibrant colors, brushstrokes, bustling streets, traditional architecture, hazy atmosphere, soft light, Claude Monet.

Figure 3. **Qualitative results of image generation.** The images are of size 512×512.

Image Captioning



Prompt: Describe this image in detail.

Response: The image depicts a panoramic view of a bustling urban cityscape, likely a major metropolitan area. The foreground is dominated by numerous high-rise buildings, showcasing a dense concentration of skyscrapers. These buildings vary in height and architectural style, with some featuring modern glass facades and others displaying more traditional designs. The city appears to be a hub of economic activity, given the density and variety of the structures.

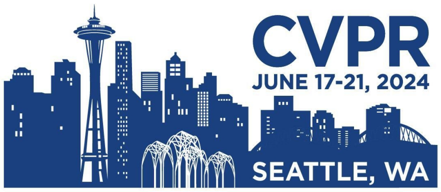
In the midground, the city extends into a large body of water, likely a harbor or bay, with numerous boats and ships. This waterway is likely a significant transportation route, facilitating the movement of goods and people. The presence of numerous boats suggests a thriving maritime industry.

The background features a vast expanse of water, likely a bay or a coastal area, with more high-rise buildings and urban development extending into the distance. The horizon is lined with a range of mountains or large hills, adding a natural backdrop to the otherwise man-made environment. The sky above is mostly clear with a few scattered clouds, indicating good weather conditions.

The overall scene is one of economic prosperity and urban development, with the combination of modern skyscrapers and the expansive waterway suggesting a city that is both economically vibrant and strategically located for maritime trade. The presence of numerous boats and the extensive waterway indicate a well-developed infrastructure, likely supporting a significant population and economy.

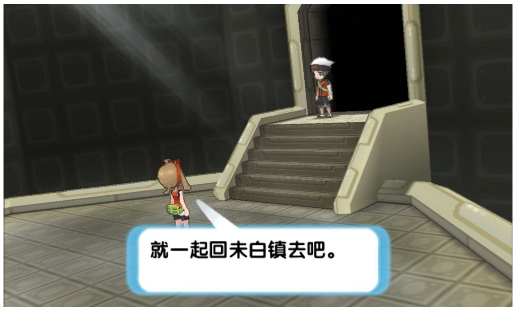
Prompt: Where is this most likely to be?

Response: This image is most likely of Hong Kong. The presence of the distinctive twin-towered skyscrapers, the hilly terrain, the bay with numerous boats, and the urban skyline with numerous high-rise buildings are characteristic features of Hong Kong's skyline.



Prompt: Tell me about this logo.

Response: This is a logo for the CVPR conference, standing for the Conference on Computer Vision and Pattern Recognition, scheduled from June 17 to June 21, 2024, in Seattle, WA. The logo features a stylized blue cityscape of Seattle, prominently displaying notable landmarks such as the Space Needle. The text "CVPR" is displayed in large, bold letters to the right of the cityscape, with the dates and location information below it. The design is simple yet recognizable, effectively communicating the event's location and timing.



Prompt: What is the content in this image?

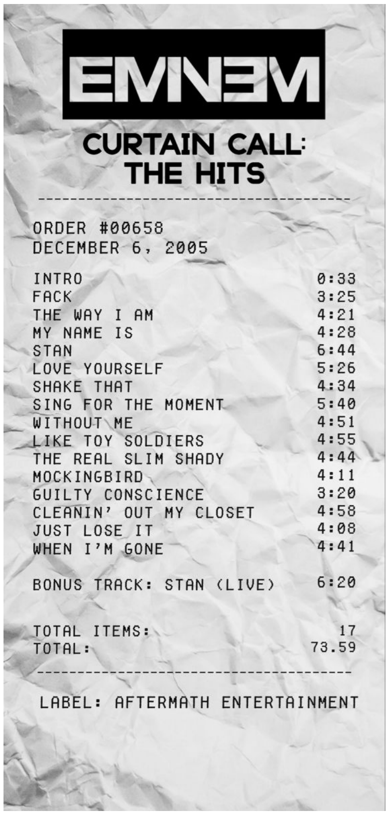
Response: The image appears to be a screenshot from a video game. The scene is set in a dimly lit, stone-walled environment that resembles a castle or fortress. The walls are made of large stone blocks, and there is a staircase leading up to a doorway on the right side of the image. The doorway is open, revealing a dark interior.

In the foreground, there is a character with brown hair, wearing a red and white outfit. The character is facing away from the camera, looking up towards the stairs.

In the background, another character is standing on the stairs, facing the camera. This character has a white cap, a red and white outfit, and is carrying a backpack. The character is looking down the stairs towards the open doorway.

At the bottom of the image, there is a dialogue box with Chinese text. The text reads: “就一起回末白镇去吧。”.

OCR



EMINEM

CURTAIN CALL:
THE HITS

ORDER #00658
DECEMBER 6, 2005

INTRO	0:33
FACK	3:25
THE WAY I AM	4:21
MY NAME IS	4:28
STAN	6:44
LOVE YOURSELF	5:26
SHAKE THAT	4:34
SING FOR THE MOMENT	5:40
WITHOUT ME	4:51
LIKE TOY SOLDIERS	4:55
THE REAL SLIM SHADY	4:44
MOCKINGBIRD	4:11
GUILTY CONSCIENCE	3:20
CLEANIN' OUT MY CLOSET	4:58
JUST LOSE IT	4:08
WHEN I'M GONE	4:41
BONUS TRACK: STAN (LIVE)	6:20
TOTAL ITEMS:	17
TOTAL:	73.59

LABEL: AFTERMATH ENTERTAINMENT

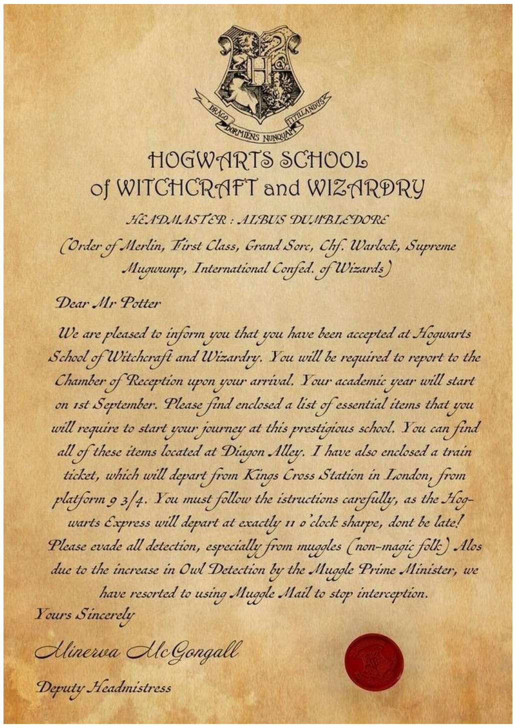
Prompt: Describe this image from @albumreceipt.

Response: This image appears to be a tracklist for a music release titled “Curtain Call: The Hits” by EMINEM, in a receipt-like style. The tracklist is as follows:

1. Intro - 0:33
2. Fack - 3:25
3. The Way I Am - 4:21
4. My Name Is - 4:28
5. Stan - 6:44
6. Love Yourself - 5:26
7. Shake That - 4:34
8. Sing For The Moment - 5:40
9. Without Me - 4:51
10. Like Toy Soldiers - 4:55
11. The Real Slim Shady - 4:44
12. Mockingbird - 4:11
13. Guilty Conscience - 3:20
14. Cleanin’ Out My Closet - 4:58
15. Just Lose It - 4:08
16. When I’m Gone - 4:41
17. Stan (live, bonus track) - 6:20

The total tracklist includes 17 songs. The total length of the songs in this tracklist is 73:59. The release date is December 6, 2005. The label is “Aftermath Entertainment”.

Document Understanding



DRACO DORMIENS NUNQUAM TITILLANDUS

HOGWARTS SCHOOL
of WITCHCRAFT and WIZARDRY

HEADMASTER : ALBUS DUMBLEDORE

(Order of Merlin, First Class, Grand Sorc, Chf. Warlock, Supreme Mugwump, International Confed. of Wizards)

Dear Mr Potter

We are pleased to inform you that you have been accepted at Hogwarts School of Witchcraft and Wizardry. You will be required to report to the Chamber of Reception upon your arrival. Your academic year will start on 1st September. Please find enclosed a list of essential items that you will require to start your journey at this prestigious school. You can find all of these items located at Diagon Alley. I have also enclosed a train ticket, which will depart from Kings Cross Station in London, from platform 9 3/4. You must follow the istructions carefully, as the Hogwarts Express will depart at exactly 11 o'clock sharpe, dont be late! Please evade all detection, especially from muggles (non-magic folk) Alos due to the increase in Owl Detection by the Muggle Prime Minister, we have resorted to using Muggle Mail to stop interception.

Yours Sincerely

Minerva McGongall

Deputy Headmistress

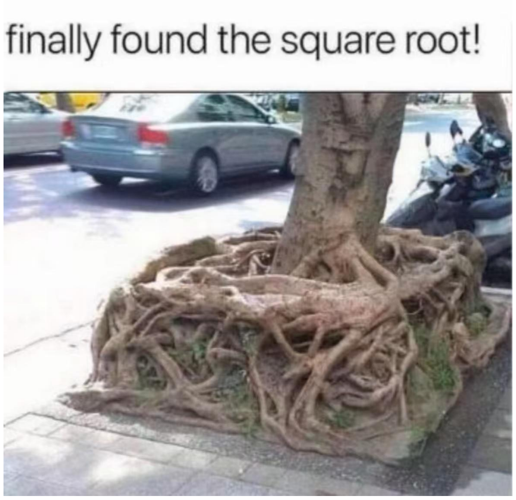
Prompt: Which platform will the train depart from?

Response: The train will depart from King's Cross Station in London, specifically Platform 9 3/4.

Prompt: Where should Mr. Potter buy the items needed for school?

Response: Based on the image, Mr. Potter should buy the items needed for school from the Diagon Alley, which is mentioned as a requirement for the school trip.

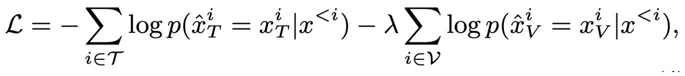
Humor Understanding



Prompt: What is the punchline?

Response: The punchline to the joke in the image is the unexpected and humorous twist on the common phrase "finding the square root of a number." Instead of finding the square root, the image humorously shows someone finding a tree root in the square shape. The root structure is exaggerated to the point where it appears to be an infinite loop, adding a layer of absurdity and humor to the joke.

Image to Latex



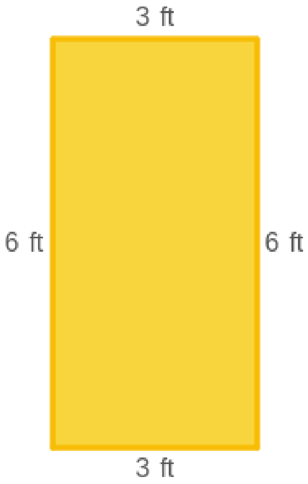
$$\mathcal{L} = -\sum_{i\in\mathcal{T}} \log p(\hat{x}^i_T = x^i_T|x^{<i}) - \lambda \sum_{i\in\mathcal{V}} \log p(\hat{x}^i_V = x^i_V|x^{<i}),$$

Prompt: Convert the formula in the figure to latex format.

Response: Certainly, I'll convert the formula in the figure to LaTeX format for you.

```
\begin{equation}
\mathcal{L} = - \sum_{i \in \mathcal{T}} \log p(\hat{x}^i_T = x^i_T | x^{<i})
- \lambda \sum_{i \in \mathcal{V}} \log p(\hat{x}^i_V = x^i_V | x^{<i}),
\end{equation}
```

Math



Prompt: What is the perimeter of the rectangle?

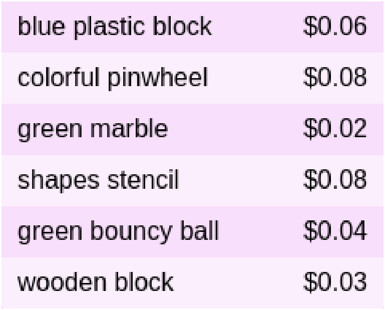
Response: To calculate the perimeter of the rectangle, we need to add up the lengths of all four sides. In this case, the rectangle has two sides that are each 6 feet long, and two sides that are each 3 feet long.

The perimeter of a rectangle is calculated by adding the lengths of all four sides. So, the formula is:
Perimeter = 2(length + width)

For this rectangle, the length is 6 feet and the width is 3 feet.

Plugging these values into the formula, we get:
Perimeter = 2(6 + 3) = 2(9) = 18

Therefore, the perimeter of the rectangle is 18 feet.



blue plastic block	$0.06
colorful pinwheel	$0.08
green marble	$0.02
shapes stencil	$0.08
green bouncy ball	$0.04
wooden block	$0.03

Prompt: Lena has $0.13. How much money will Lena have left if she buys a shapes stencil and a green bouncy ball? (Unit: $)

Response: To find out how much money Lena will have left after buying a shapes stencil and a green bouncy ball, we need to subtract the total cost of these items from the amount she has.

The total cost of the shapes stencil and the green bouncy ball is $0.08 + $0.04 = $0.12.

To find out how much money Lena will have left, we subtract the total cost from the amount she has:

$0.13 - $0.12 = $0.01

Therefore, Lena will have $0.01 left after buying a shapes stencil and a green bouncy ball.

References

[1] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv:2404.16821*, 2024. 1

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2

[3] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014. 2

[4] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 1