Supplementary Materials
Advancing Myopia To Holism: Fully Contrastive Language-Image Pre-training

In supplementary materials, we first provide more details about our multi-to-multi contrastive learning; then offer the details to create (image, multi-texts) dataset by VLMs' captioning; finally demonstrate more visualizations.

1. Implementation Details

1.1. Multi-to-Multi Contrastive Learning

The image is much more fine-grained and informative than texts. To fully align the image semantics with text descriptions from various perspectives, we propose one novel contrastive learning paradigm called multi-to-multi contrastive learning (M2M). The objective of M2M is to let the model learn multi-perspective image features under the guidance of typical texts. M2M prevents the model from aligning one summary image feature with multiple text features with different semantic meanings.

1.1.1. Holistic CLIP Model

To automatically output multi-branch image features in one forward pass, we design two types of architectures: class-token based Ψ_{CLS} and mlp-layer based Ψ_{MLP}.

Ψ_{CLS} Architecture. For most VLMs, the image encoder is transformer-based architecture, thus we always a sequence of visual embeddings with one class token to summarize the overall semantics. However, in M2M pipeline, a single class token is not sufficient to cover representations from various perspectives (only one attention map for one class token). Thus, we initialize multiple class tokens, each focusing on specific aspects of an image guided by the assigned text descriptions. In this way, we obtain multiple image features with almost no extra efforts.

Ψ_{MLP} Architecture. Although Ψ_{CLS} is already an effective way to extract features from different views, all the class tokens share the visual token sequence, which may not be always an ideal solution. Inspired by the MoE intuition [4], we also introduce one MLP-based architecture. Specifically, to output H image features, we expand the last 3 MLP layers to H parallel parts. For efficiency, we only expand the second linear FFN. In Figure 1, we show in details the architecture of both Ψ_{CLS} and Ψ_{MLP}.

1.1.2. Multi-to-Multi Matching

Here, after obtaining the H image features from $\{\mathbf{v}_i\}_{i=1}^{H} = \Psi_{\text{CLS}}(\mathcal{I})$ or $\{\mathbf{v}_i\}_{i=1}^{H} = \Psi_{\text{MLP}}(\mathcal{I})$ and M text features from the corresponding multi-perspective texts $\{\mathbf{t}_i\}_{i=1}^{M} = \Phi^{\text{txt}}(\{\mathcal{T}_i\}_{i=1}^{M})$, we match them to jointly learn the alignment. Normally, we set $H = M$ so that we assign each image head a specific type of text to align (object-oriented, background, long caption...). If M is relatively large, *e.g.*, if we prompt 4 different VLMs with 4 different prompts, thus to get 16 distinct texts. In this case we first group the texts into H sets based on their similarities [7]. If some texts are mixed up, we use free match: match the image head with maximum cosine similarity. After matching texts with their corresponding image features, we can train the model by M2M contrastive learning described in the main paper.

1.1.3. Training Details

We train ViT-B-16 model on 8 A100-80G GPUs with batch size 256 for 100 epochs on CC12M [2] and batch size 64 for 200 epochs on CC3M [2], with a learning rate of $5e^{-4}$. As we tested, the above training setup achieves the best result on Conceptual datasets, though the training converges slower than larger batch size setup.

1.1.4. Inference Details

After generating multiple image features, we first normalize them, then take the average to produce one compositional image representation for retrieval-based downstream tasks. To train LLaVA1.5 for visually dense tasks, we replace the visual encoder by our trained model and use the whole output visual sequence as input for projection layer and LLM.

2. (Image, Multi-Texts) Dataset

2.1. Prompts & Examples

In the main paper, we describe the prompt design principle for multi-view/granularity captions: *Focus Guide* indicates whether attention should be laid on foreground objects or general background substances. *Physical or Sensory* refers to describe solid nouns or feeling styles. *Gaze or Glance* define captions are dense long-details or compact short-overview. *Complex Reasoning* distinguishes relationship or sequence for entities. To generate text descriptions from

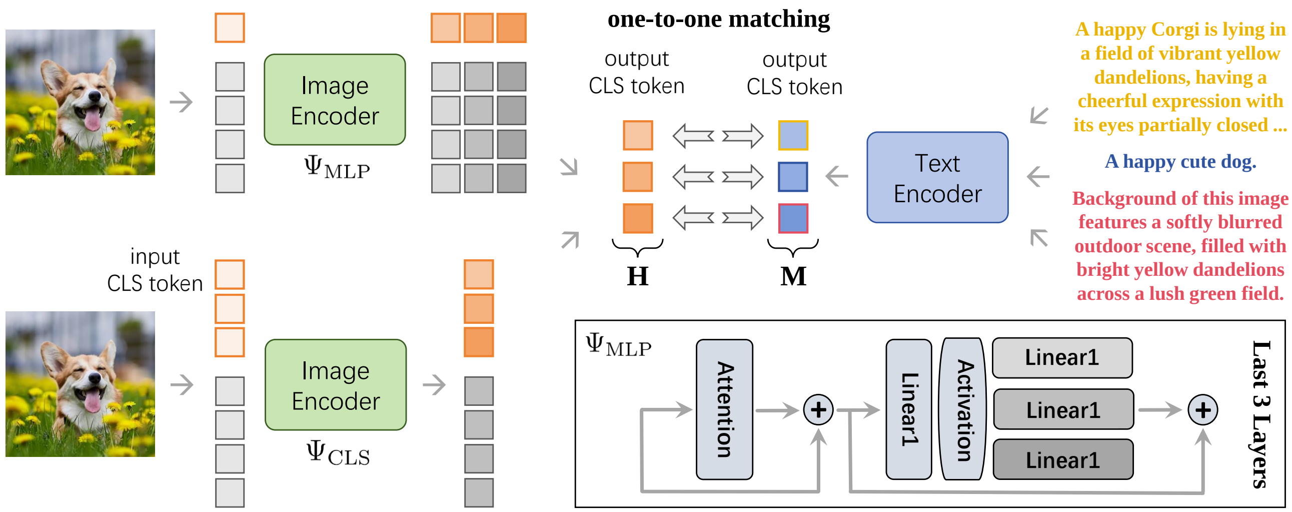


Figure 1. **Architecture Overview of Holistic CLIP.** To generate H image features, we leverage two different structures: Ψ_{CLS} and Ψ_{MLP}. Then we match H image features with M text features. Normally $H = M$ and we apply one-to-one matching.

Table 1. **Multi-Prompts Used for Constructing (Image, Multi-Texts) Data.**

Type	Prompt
Details	<image> Describe the image in detail.
Nouns	Briefly describe the image with few noun words separated by ",".
Main Object	Describe only the one main object in the image, do not say anything about the other objects or background.
Background	Describe only the background of this image, do not say anything about the foreground objects.
Style	Describe the style or your feelings about this image, do not say anything about the objects in the image.

the above aspects, we design 5 different prompts, which are listed in Table 1. We prompt InternVL2-8B to produce captions from various perspectives. In Figure 4 and 5, We offer some examples from CC12M [2] datasets. We can see that these texts describe the image from various perspectives, with different semantic meanings. For common one-to-multi contrastive learning, we align one summary image feature with all the text features, which inevitably pushes texts with different semantic meanings close to each other in the embedding space, making the alignment chaotic and confusing. However, our multi-to-multi contrastive learning fundamentally prevents such chaos, by outputting multiple image representations and learn independently the alignment.

For multi-VLMs data, we adopt the "Details" prompt in Table 1 for all VLMs. In Figure 2 and 3, we provide some caption examples. Overall, the texts generated by different VLMs is rather similar. In the main paper, we choose four captions (without style) from multi-prompts to form data for fairness.

Table 2. **Statistics for Multi-VLMs Data.** We calculate the average number of words and the caption variance of multiple VLMs.

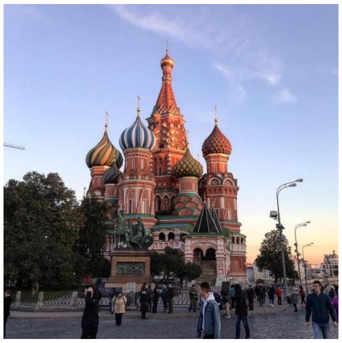
VLM	AVG Word	Var Word
Original	17.3	163.4
InternVL2 [3]	60.8	14.9
Minigpt4 [9]	16.5	24.8
LLaVA1.5 [5]	48.5	286.4
QwenVL2 [1]	58.5	23.6

Table 3. **Statistics for Multi-Prompts Data.** We calculate the caption variance of InternVL2-8B [3] using different prompts.

Prompt	AVG Word	Var Word
Details	60.8	14.9
Nouns	10.7	61.8
Main Object	31.7	382.0
Background	42.3	432.0
Style	56.2	53.3

2.2. Dataset Statistics

Table 2 and 3 report the statistics of captions by using different VLMs and prompts. The generated captions have dif-



Original: Red Square is a city square in Moscow, Russia.

Minigpt4: The image shows a large, red and white church with several domes and spires.

Internvl2: The image depicts the iconic Saint Basil's Cathedral, located in Moscow, Russia. This architectural marvel is renowned for its vibrant and colorful onion domes, which are a hallmark of Russian architecture. The cathedral features a total of 60 domes, each painted in different colors and patterns, creating a visually stunning and intricate design.

LLaVA1.5: The image features a large group of people gathered around a beautiful, ornate building. The building is a stunning red and green structure, possibly a church, with a tall steeple and a clock tower. The crowd is walking around the building, admiring its architecture and taking in the view. There are numerous people in the scene.

QwenVL2: The image depicts a bustling city square with a prominent, ornate cathedral in the background. The cathedral features multiple onion-shaped domes, each adorned with colorful patterns and intricate designs. The central dome is particularly striking, with a golden base and a red and blue striped pattern. Surrounding this central dome are several smaller domes, each with its own unique color scheme and design.

Original: 'Lighting Catalog The Home Depot 2018 Home Depot, Led, Interior Lighting, Your Space, Pendant Lighting, Cabinet, Storage, Modern, Table.

Minigpt4: The image shows a modern dining room with a long wooden table in the center.

Internvl2: Lighting Catalog The Home Depot 2018 Home Depot, Led, Interior Lighting, Your Space, Pendant Lighting, Cabinet, Storage, Modern, Table', 'The image depicts a modern, minimalist interior setting, likely a kitchen or dining area. The focal point is a sleek, horizontal LED light fixture mounted on the ceiling. The light fixture has a rectangular shape with a smooth, matte finish and is illuminated, casting a warm, ambient glow.

LLaVA1.5: The image features a modern and stylish interior space featuring a sleek, cylindrical pendant light with an illuminated textured pattern. Below the light is a minimalist wooden table with a glass of water, an open book, and clean lines. A table with a glass of wine on it.

QwenVL2: The image depicts a modern, minimalist interior space, likely a dining or office area. The focal point is a sleek, linear pendant light fixture hanging from the ceiling. The light fixture has a metallic frame with a frosted glass or crystal-like surface, emitting a soft, warm glow that illuminates the surrounding area. The wall behind the light fixture is made of large, rectangular tiles.

Figure 2. **Examples of (Image, Multi-Texts) Data from Multi-VLMs.**

Original: Woman exercising by skipping a rope.

Minigpt4: The image shows a woman in black leggings and a colorful top, standing in a room with wooden floors and a potted plant in the corner.

Internvl2: The image depicts a person engaged in a fitness activity, specifically jump rope. The individual is captured mid-air, with their body slightly tilted forward and their arms extended, holding the jump rope. They are wearing a black sleeveless top, colorful geometric-patterned leggings, and white sneakers. The person has short hair and appears focused on their exercise.

LLaVA1.5: A woman is jumping in the air with a jump rope in her hand. She is wearing colorful leggings and a black shirt. The jump rope is black and she is holding it with both hands. The woman is in the middle of a jump, and she is the main focus of the image.

QwenVL2: The image depicts a person jumping rope in a room with a minimalist and modern aesthetic. The individual is wearing a black t-shirt and brightly colored leggings with a geometric pattern featuring shades of orange, green, blue, and yellow. They are also wearing white sneakers. The person is captured mid-jump, with the rope stretched out in front of them.

Original: 'A bathroom at Staybridge Suites Toronto - Vaughan South

Minigpt4: The image shows a modern hotel room with a kitchenette and dining area.

Internvl2: The image depicts a well-appointed studio apartment with a kitchenette and a bedroom area. The bedroom is furnished with a single bed that has a white bedspread and a single pillow. The headboard is upholstered in a dark wood material, adding a touch of elegance to the room.

LLaVA1.5: The kitchen is a modern space with a stainless steel refrigerator, a sink, a microwave, and a stove. The refrigerator is located on the right side of the kitchen, while the sink is positioned in the middle. The microwave is placed above the stove, which is situated on the far right side.

QwenVL2: The image depicts a modern, well-furnished hotel room or studio apartment. The room is divided into two main sections: a sleeping area and a kitchenette. Sleeping Area: The bed is positioned against the left wall, with a wooden headboard. The bed is neatly made with white bedding. Kitchenette: The stainless steel appliances, including a refrigerator, a microwave, and a small stovetop are equipped.

Figure 3. **Examples of (Image, Multi-Texts) Data from Multi-VLMs.**

ferent length distributions, forming long/short counterparts. More recent VLMs like Internvl and QwenVL offer long captions steadily while Minigpt4 tends to give a short description. For multi-prompt datasets, the captions for main object and background descriptions vary much in length, highly depending on the image semantics (object oriented or complex scene).

2.3. Caption Details

For multi-VLM captioning, we utilize LLaVA1.5-13B [5], Minigpt4-7B [9], QwenVL2-7B [1] and InternVL2-8B [3]. For multi-prompt captioning, we use InternVL2-8B [3] for its superior instruction-following performance. We set the max number of output tokens to be 77 due to the length limitation of CLIP [6]. On CC12M [2], it takes around 2000-2400 A100-80G GPU hours (depending on the used VLM and prompt) to generate one typical caption (500-700 GPU hours on CC3M [2]). After captioning, we filter captions by setting a minimum length (string length$>$10) and threshold on clip score to filter out bad cases, then we re-caption those unqualified data until they meet the standard.

3. Visualizations

Here we offer more visualizations on the attention distribution among different image class tokens in Ψ_{MLP}. In Figure 6, we demonstrate that our M2M contrastive learning paradigm allows to produce image features that summarize visual semantics from various aspects.

4. Limitations & Future Work

Due to limited computing, we do not fully explore the case where M is large (multi-VLMs and multi-prompts) or conduct experiments on larger datasets (*e.g.* Laion400M [8]). Nevertheless, our exploration on holistic alignment is anticipated to bring significant benefits for the VLM domain. In reality, there exist naturally occurring datasets comprising (image, multi-text) pairs, such as those found in commodity scenario (commodity attribute/merchant description/buyer comments), where our idea is expected to show the superior capability.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2, 4

[2] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pretraining to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021. 1, 2, 4

[3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2, 4

[4] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2022. 1

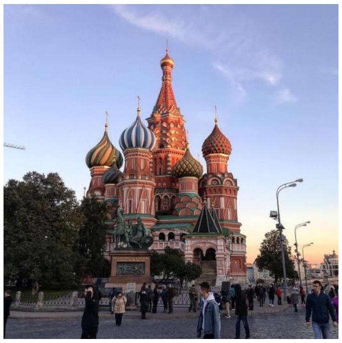
[5] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2, 4

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021. 4

[7] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*, 2019. 1

[8] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 4

[9] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2, 4



Original: Red Square is a city square in Moscow, Russia.

Nouns: Church, St. Basil's, Moscow, Russia, tourists, dome, sunset, sky, street, people.

Style: The image exudes a sense of grandeur and historical significance. The vibrant colors and intricate details of the architecture evoke a feeling of awe and admiration. The play of light and shadow adds depth and dimension, creating a visually captivating scene. The overall atmosphere is one of majesty and cultural richness.

Main Object: The main object in the image is the colorful Saint Basil's Cathedral. It features multiple onion-shaped domes, each adorned with vibrant colors and intricate patterns. The central dome is the largest and most prominent, with a golden top. The structure is surrounded by smaller domes in various colors, including blue, green, and red.

Background: The background of the image features a stunning, vibrant building standing out against a clear sky with a gradient of colors transitioning from blue to warm hues near the horizon, suggesting either sunrise or sunset.

Original: 'Lighting Catalog The Home Depot 2018 Home Depot, Led, Interior Lighting, Your Space, Pendant Lighting, Cabinet, Storage, Modern, Table.

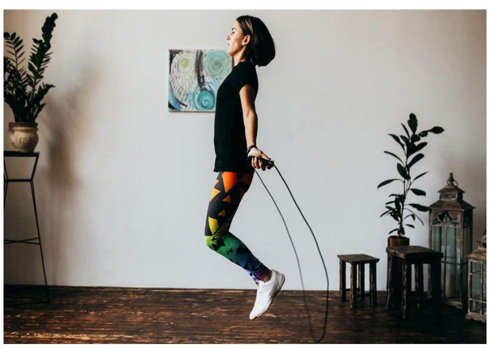
Nouns: Lamp, table, books, window, cityscape, glass, smartphone, magazine.

Style: The image exudes a modern and minimalist aesthetic, characterized by clean lines, a neutral color palette, and a focus on functionality. The overall atmosphere is one of sophistication and contemporary elegance, with a touch of industrial design elements. The lighting fixture adds a warm, inviting glow, enhancing the serene and calm ambiance of the space.

Main Object: The main object in the image is a sleek, modern pendant light fixture. It features a rectangular shape with a frosted glass cover that diffuses the light, creating a soft and ambient glow. The fixture is suspended from the ceiling by a simple, elegant cord, and it appears to be mounted on a wall or ceiling.

Background: The background of the image features a modern, minimalist interior with a large window that offers a view of a cityscape at night. The city lights are visible, creating a vibrant and bustling atmosphere outside. The interior walls are made of a light-colored, textured material, possibly concrete or stone, adding to the contemporary feel of the space.

Figure 4. **Examples of (Image, Multi-Texts) Data from Multi-Prompts.**



Original: Woman exercising by skipping a rope.

Nouns: Woman, jump rope, room, plants, furniture, wall art.

Style: The image exudes a sense of calm and serenity. The soft, natural lighting and the minimalist decor create a tranquil atmosphere. The person's posture and expression suggest a moment of peaceful reflection or relaxation. The overall aesthetic is clean and uncluttered, evoking a feeling of simplicity and harmony.

Main Object: The main object in the image is a person performing a jump rope exercise.

Background: The background of the image features a plain, light-colored wall. On the left side, there is a potted plant with long, slender leaves. Next to the plant, there is a small, black, geometric-shaped table. On the right side, there is another potted plant with broader leaves. Near the center of the wall, there is a framed piece of abstract.

Original: 'A bathroom at Staybridge Suites Toronto - Vaughan South

Nouns: Bedroom, kitchen, refrigerator, microwave, sink, coffee maker, dining table, chairs, wall art, bed, pillows, nightstand, lamp.

Style: The image exudes a modern and clean aesthetic, with a harmonious blend of neutral tones and subtle earthy accents. The overall ambiance is one of comfort and functionality, creating a welcoming and practical space. The design elements are well-coordinated, providing a sense of cohesion and order.

Main Object: The main object in the image is a kitchen area with a granite countertop.

Background: The background of the image features a wall with a large, abstract painting that has a circular design with intricate patterns. The wall is painted in a light color, and there is a small, square-shaped vent near the top. The ceiling is also light-colored, and there is a small, rectangular light fixture installed.

Figure 5. **Examples of (Image, Multi-Texts) Data from Multi-Prompts.**

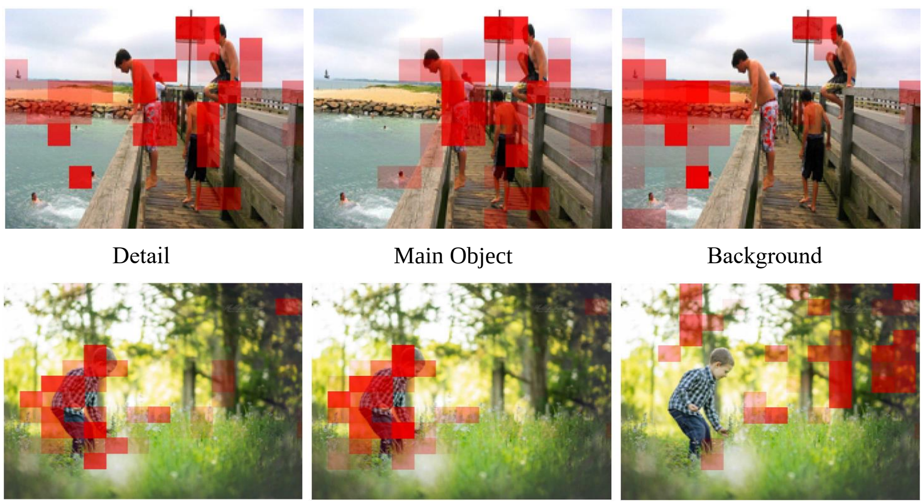


Figure 6. **Visualization of Attention Maps Among Different Visual Class Tokens.**