8. Details of Conversing with ChatGPT

Multi-turn conversation. We use ChatGPT to generate a set of new prompts based on the top and bottom performing prompts (line 10 of Algorithm 2). The exact prompts we use are:

```
Hi ChatGPT, assume you are a pattern
learner.  I have two lists of CLIP
templates:  one with good templates
and the other with bad templates.
There are latent patterns that make
a template good or bad.  Based on
these patterns, give me a better
template for image classification
while avoiding worse template.
Here is the list of good templates:
- good1
- good2
- ...
Here is the list of bad templates:
- bad1
- bad2
- ...
Here are my requirements:
- Please only reply with the
template.
- The template should be fewer than
15 words.
- The template should have a similar
structure to the above templates.

Positive Response (if the new prompt
outperforms the top-k)
The performance of the template
``newTemplate'' improves to X.XX%.
Please give me a better template.

Negative Response
The performance of the template
``newTemplate'' drops to X.XX%.
Please give me a better template.
```

Alternative implementation: sending only the initial prompts (default). Multi-turn conversation requires appending all chat history to ChatGPT's official API at every iteration, which costs more input tokens and money. In Figure 5, we show that one can only send the initial prompts (without any response) to ChatGPT at every iteration to get equivalent and even slightly better performance. However, it is important to also update the top-k and bottom-k prompts at every iteration (**Iterative**) for efficiency. We show that the **Non-Iterative** version that keeps re-using the initial top-k and bottom-k prompts leads to worse performance. Therefore, in our paper, we stick to **Iterative** for all experiments.

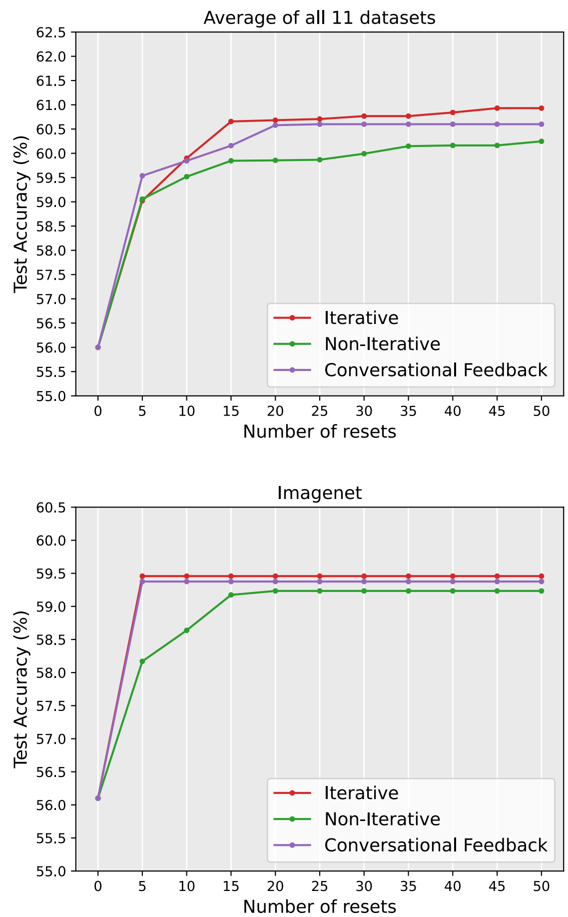


Figure 5. **Updating initial prompts can be as effective as multi-turn conversation.** We ablate different ways of conversing with ChatGPT on all 11 datasets (left) and ImageNet (right). Notably, we find that only updating the top-k and bottom-k prompts (**Iterative**) is as performant and thus a cheaper alternative because sending response to ChatGPT costs more input tokens. On the other hand, reusing the initial prompts (**Non-Iterative**) leads to worse overall performance.

Positive Only (P only). When using only positive prompts, we can remove negative prompts and provide twice as many positive examples:

```
Hi ChatGPT, assume you are a pattern
learner.  I have one list of CLIP
templates:  one with good templates.
There are latent patterns that make
a template good.  Based on these
patterns, give me a better template
for image classification.
Here is the list of good templates:
- good1
- good2
- ...
Here are my requirements:
- Please only reply with the
template.
- The template should be fewer than
15 words.
- The template should have a similar
structure to the above templates.
```

9. Additional Experimental Results

In this section, we present additional experiments to gain further insights into our method.

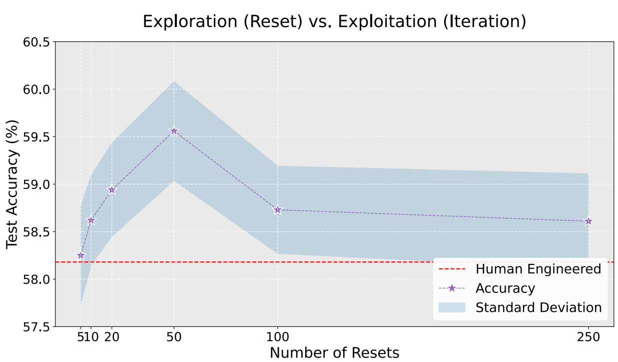


Figure 6. **Balancing exploration and exploitation.** We use a fixed budget of 500 ChatGPT API calls per restart, and ablate the optimal number of resets to use in our algorithm on 1-shot ImageNet. The number of iterations is thus inversely proportional to the number of resets; for example, 10 resets would allow for 50 iterations per reset. We take the average over three runs and also report the standard deviation. We find the optimal balance of exploration and exploitation to be 10 iterations and 50 resets. In contrast, "pure" exploration (2 iterations, 250 resets) leads to 0.9% lower accuracy due to insufficient optimization. On the other hand, when exploitation is overly prioritized (100 iterations, 5 resets), our method gets 1.3% lower accuracy.

Balancing exploration and exploitation can improve the final performance. Our method extensively leverages the ChatGPT API, necessitating an investigation into strategies for minimizing optimization costs. This leads us to examine the classic dilemma of exploration versus exploitation, a foundational concept in reinforcement learning. Specifically, we use a fixed budget of 500 API calls per restart, and investigate the optimal combination of the number of resets and iterations in Figure 6. For example, we can allocate 50 resets with 10 iterations each to encourage more exploration, or 10 resets with 50 iterations each to foster more exploitation. We find that the optimal balance point is 50 resets of 10 iterations each, and note that no other combination is within 1 standard deviation of the optimal performance. As shown in the performance curve, having too much exploration (250 resets), or too little (5 resets) will result in a roughly 1% decrease in performance. In general, we find it is useful to spend more budget on exploration as ChatGPT can be stuck at local minima within one reset.

Backbone	Method		
	Our Approach	Hand-Engineered	Linear Probe
ResNet-50	59.6	58.2	55.9
ResNet-101	61.8	61.6	59.8
ViT-B/32	62.6	62.0	59.6
ViT-B/16	67.8	66.7	65.9

Table 7. **Our method can generalize to various CLIP architectures.** We run our method on 1-shot ImageNet across multiple CLIP backbones, and compare it to the best Human-Engineered prompt and Linear-Probing [54] performance.

Reimplementing (iterative) APE for VLM optimization. We attempt to implement our own version of iterative APE using the given prompts in [78] while making minimal changes such that it fits in our automatic prompt-searching system. For a fair comparison, we reuse exactly the same initial sampled prompts from LAIONCOCO-1M for iterative APE because their "instruction-induction" paradigm cannot be applied to VLM optimization settings. The results are shown in Table 8. We find that iterative APE shows inferior performance to our method, presumably because we leverage more textual feedback for more efficient search. The exact prompt we use is shown below:

```
Hi ChatGPT, generate a single
variation of the following template
while keeping the semantic meaning:
- template
Here is my requirement:
- Please return a single template
starting with '-'
```

Comparison of CLIP backbones. To verify that our method scales properly to other CLIP backbones, we test our method on ImageNet using four different CLIP backbones:

Method	Dataset											Avg
	Caltech	ImageNet	Aircraft	Food	Pets	Cars	SUN	UCF	DTD	EuroSAT	Flowers	
LAIONCOCO-1M	81.4	56.2	17.4	76.5	79.6	51.3	54.9	55.8	43.1	38.6	61.3	56.0
Iterative APE	88.3	58.1	17.0	77.3	85.1	54.8	58.6	57.4	41.2	46.7	65.3	59.0
Ours (P only)	89.0	59.4	17.9	77.8	85.7	55.7	60.4	58.7	43.6	46.7	66.6	60.1
Ours (P+N)	**89.1**	**59.6**	**18.1**	**78.3**	**88.1**	**56.2**	**61.0**	**60.2**	**44.8**	**49.0**	**67.2**	**61.1**

Table 8. **Comparing our method with our own version of iterative APE [78].** Optimized using 1-shot training sets, we find that both iterative APE and our methods can effectively improve upon the initial sampled prompts. However, our method achieves better performance within the same computational budget, presumably because we provide explicit textual feedback to ChatGPT, leading to faster convergence.

Shot	Method	Dataset											Avg
		Caltech	ImageNet	Aircraft	Food	Pets	Cars	SUN	UCF	DTD	EuroSAT	Flowers	
1 shot	Ours (P only)	89.0	59.4	17.9	77.8	87.8	55.7	60.4	58.7	43.6	46.7	66.6	60.1
	Ours (P+N)	89.1	59.6	18.1	78.3	88.1	56.2	61.0	60.2	44.8	49.0	67.2	61.1
16 shot	Ours (P only)	89.3	59.6	17.7	77.9	86.6	56.2	61.0	60.2	44.0	49.0	66.0	60.6
	Ours (P+N)	89.5	59.9	18.1	78.3	88.3	56.8	60.8	60.5	44.9	51.4	67.4	61.4

Table 9. **Higher-shot performance.** We report the 16-shot performance of our method in this table. It is important to note that as the number of shots increases, the role of the natural language prompt diminishes because it will be more effective to tune the visual representations (which requires white-box access to VLMs).

GPT version	Dataset											Avg
	Caltech	ImageNet	Aircraft	Food	Pets	Cars	SUN	UCF	DTD	EuroSAT	Flowers	
gpt-turbo-3.5-0301	89.1	59.6	18.1	78.3	88.1	56.2	61.0	60.2	44.8	49.0	67.2	61.1
gpt-4-0314	89.1	59.6	17.9	78.5	87.7	56.2	60.3	59.9	45.0	48.0	67.6	60.9

Table 10. **ChatGPT versus GPT4.** Our approach is equally effective using other versions of ChatGPT.

ResNet-50, ResNet-101, ViT-B/32, and ViT-B/16. We compare our method with hand-engineered prompts, and a linear probe (linear classification on the visual embeddings). Table 7 shows the results of the experiment, where we see that our method outperforms the baselines consistently. Thus, our method scales appropriately with larger and more powerful models.

Results on higher shots. We additionally test the generalization ability of our method given more data (16 shots), with results shown in Table 9. We observe that our method gains small but incremental improvements given more data, and using both top-k and bottom-k prompts (P+N) consistently outperforms top-2k prompts (P only).

Results using GPT4. We run our approach using the same hyperparameters and initial prompts using GPT4 in Table 10. It shows that our approach is equally effective using other versions of ChatGPT, but interestingly, there is no performance benefit of using GPT4. This may be because our hyperparameters were optimized on ChatGPT, and are suboptimal for GPT4.

Cost analysis. We use GPT3.5 which costs \$0.0015 per 1000 tokens. In our default setup, we use an average of 500 tokens per API call. We use a total of 500 API calls (50 resets and 10 iterations) for a total of 250,000 tokens per restart, and thus each run costs around 50 cents. Since we use 20 restarts per dataset, the total cost over the suite of 11 datasets is around \$100 for each of the three folds.

10. T2I Experimental Details

In this section, we include implementation details and more qualitative results for T2I generation experiments.

Image generation using DALL-E 3. We use the below template to generate images without changing the prompts.

```
Create this exact image without any
changes to the prompt:  {prompt}.
```

T2I generation (Figure 3). We use DALL-E 3 to expand the query text to a longer prompt for the first image. Next, we send generated image, query text, and current prompt to GPT4-V for prompt optimization.

```
Prompt for DALLE-3 (first round):
Create an image that shows {query
text}.

Prompt for GPT-4V: Do you think this
image {generated image} correctly
depicts {query text}?  If not,
briefly explain why and suggest
modifications.  Then, help me adjust
the prompt to make it correct:
{prompt}.  Please provide a response
in a JSON file format containing:
(1) "feedback" summarizing the key
points, and (2) "new_prompt" with
the revised text.
```

Prompt inversion (Figure 4). We use GPT4-V to generate the initial prompt given the query image. Next, we send query image, generated image, and current prompt to GPT4-V for prompt optimization.

```
Prompt for GPT-4V (first round):
Generate a detailed text prompt to
recreate the attached image {query
image} using an image generator.

Prompt for GPT4-V: Compare the
original image {query image} and
generated image {generated image},
analyze their differences, and
then propose changes to update the
original prompt in-place:  {prompt}.
Please provide a response in a
JSON file format containing:  (1)
"feedback" summarizing the key
points, and (2) "new_prompt" with
the revised text.
```

Failure cases. We show some failure cases of our method in Table 13 and Table 14. We note that these queries are especially challenging even for state-of-the-art VLMs because they require complex reasoning abilities. We expect better performance of our framework using stronger generative models in the future.

Prompt inversion on natural images. In addition to sampling queries from DiffusionDB [68], we also attempt at prompt inversion with natural images, as shown in Table 15.

Human studies. We hire two graphical designers who have one year of experience using AI content creation tools such as Stable Diffusion and Midjourney to manually design the prompts for DALL-E 3. We also hire two volunteers to assign a Likert scale score between the generated image and user query according to Table 16.

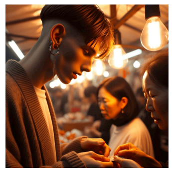
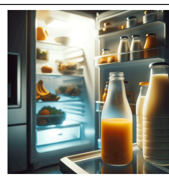
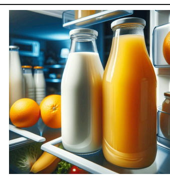
User Query	Init. Image	Final Image	Final Prompt
		Text-to-image generation	
The unmasked wrestler hits the masked wrestler.			Photo of a wrestling ring where an unmasked male wrestler with a muscular physique is in the midst of delivering a powerful blow to a masked male wrestler donning a lucha libre style mask. The spectators in the background are on the edge of their seats, watching the action closely.
The person with earrings pays the person without earrings			Photo of a person with a short haircut and noticeable earrings in the process of paying a long-haired vendor without earrings at a market stall, with warm lighting.
A bird eats a snake			Photo of a vast desert landscape under a clear blue sky. In the foreground, a large, powerful eagle with brown feathers and piercing eyes is perched confidently on a tall, green cactus. The eagle tightly clenches a rattlesnake in its strong talons. The snake's rattle is visible, and it appears to be struggling. The eagle's beak is wide open, showing its sharp beak, indicating it's about to consume the snake.
A shorter person is covering the eyes of a taller person.			A shorter individual reaching up to cover the eyes of a taller person standing in front of them. The shorter person is on their tiptoes, trying to reach the taller person's eyes. Both of them are smiling and seem to be enjoying the moment. The backdrop is simple and unobtrusive to maintain focus on the subjects.
There is less milk than orange juice.			A kitchen scene showing a fridge door open. Inside, there's a clear glass bottle of milk and a larger bottle of orange juice. Both the milk and orange juice bottles are almost full, with a nearly equal amount visible at the top. The fridge shelves are well-lit, and the rest of the fridge contains a variety of food items including fruits and vegetables. The focus is on the milk and orange juice bottles that are prominently displayed on the shelves.
Getting a horse wet.			A scene depicting a person using a hose to gently spray water on a horse in an open field. The horse appears calm and enjoys the water, with droplets of water glistening on its coat. The person is smiling, dressed in casual outdoor attire. The background features a clear blue sky and a few trees, creating a serene and peaceful setting. The horse is a beautiful chestnut color, and the person is Caucasian with short brown hair.
Some are parking in a train.			A whimsical scene depicting a train where some of the carriages are designed as parking spaces, with various types of cars parked inside them. The train is moving through a picturesque landscape, with mountains in the background and a clear blue sky overhead. The cars in the train's parking carriages include a red sports car, a green SUV, and a yellow compact car. The train itself is a classic steam locomotive with a touch of modern design, emitting a puff of steam as it chugs along the tracks.
The white wall will soon be painted blue.			A white wall in a room, with a paint can and a paintbrush beside it. The can is open and filled with blue paint, ready for use. A painter, a middle-aged Caucasian man wearing a white painter's outfit and a cap, is dipping the brush into the blue paint, preparing to start painting the wall. The room has a window with daylight coming through, casting a bright ambiance over the scene.

Table 11. **More results of T2I optimization.**

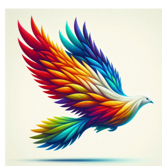
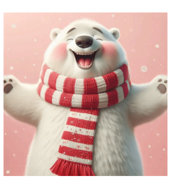
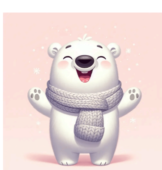
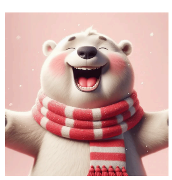
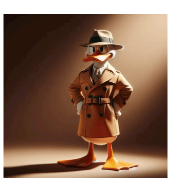
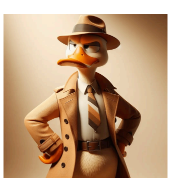
User Query	Init. Image	Final Image	Final Prompt
			**Prompt inversion** Create a digital artwork of a stylized, geometric rhinoceros head with a dynamic array of sharp, crystalline facets in a monochromatic palette of black, white, and gray. The design should feature intricate shadows and highlights to produce a three-dimensional illusion, with a focus on accurately representing the creature's contours and muscle structure. Adjust the composition to show the rhinoceros head from a frontal perspective, ensuring that both the horn and the ears are symmetrically aligned in the center. Emphasize the geometric nature of the facets by making them more pronounced and varied in shape, creating a complex mosaic that captures the interplay of light and shadow. Add a slight glow to the edges of the facets to enhance the three-dimensional effect and the metallic quality of the artwork. Display the rhinoceros head against a pitch-black background, with a light source positioned to cast dramatic, high-contrast illumination that emphasizes its multifaceted texture. Incorporate a subtle reflective sheen on the surface to suggest a sleek, metallic finish, and ensure the rhinoceros's eye is detailed and expressive, contributing to the overall lifelike appearance of the artwork.
			A hyper-realistic full slice of an orange with intricate details, including the textured pulp and clearly defined rind, positioned off-center on a reflective gradient surface transitioning from white to dark. The orange's juicy texture is accentuated by a dynamic splash of juice, with droplets captured mid-air, creating an energetic and lively scene. The lighting is dramatic and contrasting, with a spotlight effect casting a pronounced shadow to one side to enhance the three-dimensional effect and emphasize the vibrant orange color. Include a clear reflection on the surface and a small stem attached to the orange slice to underscore the realism and freshness. Enhance the composition by ensuring the orange slice is angled slightly, with the splash of juice originating from the lower right side, to add a sense of motion and vitality.
			A medieval knight in full armor stands with a shield, the dark background highlighting his silhouette against a subtle warm glow. His helmet features a visor with a single vertical slit, and his armor includes a chainmail coif beneath a segmented plate gorget and articulated plate gauntlets, with layered plate armor and flared ridged pauldrons. The knight's shield is centered and bears a detailed, embossed golden fleur-de-lis on a field of weathered steel, surrounded by rivets. The vibrant orange cloak drapes over both shoulders and behind his back, adding a touch of regal color to the composition. His stance is grounded and balanced, with his left arm extended, presenting the shield, and his right hand resting on the pommel of his sword, exuding a calm and noble demeanor.
			Create a stylized illustration of a dove in flight, with feathers that transition smoothly through a spectrum of colors including red, orange, yellow, green, blue, indigo, and violet. The dove's plumage should resemble a dynamic, three-dimensional arrangement of vibrant, overlapping feathers, giving a sense of movement and freedom. The style should be a fusion of semi-realistic and digital art, with a focus on vivid colors and a clean, light background that emphasizes the artwork's lively and spirited nature. Adjust the feather arrangement to be more structured and flame-like, with the feathers at the tips being more elongated and pointed to enhance the sense of elegance and flow.
			Create an illustration of a stylized, geometric dinosaur with a textured body in two shades of green: a lighter green for the main body and a darker green for the spiky plates along its back. The dinosaur should have a friendly demeanor, with a long, curved tail and a smooth, rounded head featuring two small, circular white eyes with black pupils. It should stand on two legs with small, rounded feet, each with three visible toes. The background should be a flat, light beige color, with a simple, elongated shadow extending to the right of the dinosaur, indicating a soft light source to the left.
			Generate an image of a cartoon-style polar bear with gleefully closed eyes and a wide, toothy grin, revealing just a hint of its tongue. The bear should look exuberant, standing on its hind legs with arms open wide as if ready for a hug. The bear's fur should appear extremely soft and fluffy, with a pronounced blush of rosy pink on both cheeks and belly, enhancing its charm. Adorn the bear with a cozy, chunky-knit scarf, vibrant red with prominent, horizontal white stripes, stylishly wrapped around its neck and draping with a dense tassel fringe at the ends. Situate the bear against a gentle pastel pink backdrop, scattered with delicate, small snowflakes, conveying the splendor and coziness of festive winter cheer.
			An anthropomorphic duck standing confidently with hands on hips, styled as a classic film noir detective. The duck has a calm and cool expression, wearing a tan detective's fedora hat and a matching double-breasted trench coat, buttoned up, with a broad collar, epaulets, and a belted waist. The character has a white shirt and a patterned tie with a diagonal stripe design underneath. The character has orange webbed feet and a large, prominent beak. The lighting is dramatic, with a strong contrast between light and shadow, creating a focused shadow on the background that mimics the character's silhouette. The overall color palette is warm with a gentle light source coming from the side, casting the background in a gradient from warm beige to shadows, giving the image a mysterious and dramatic appearance.

Table 12. **More results of prompt inversion.**

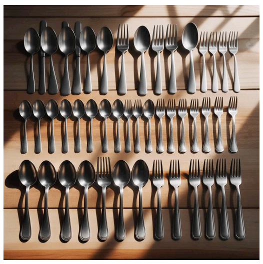
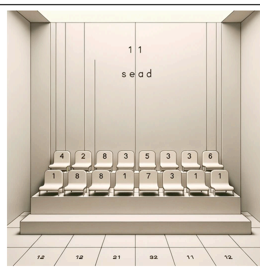
User Query	Final Image	Final Prompt	Reason for Failure
		Text-to-image generation	
There are fewer forks than spoons.		Photo of a wooden dining table with a collection of spoons and forks laid out. It's clear upon counting that there are fewer spoons than forks. The spoons are stainless steel and placed on the left while the forks, also stainless steel, are on the right. Sunlight filters from a nearby window, casting soft shadows.	Failure to compare the exact number of forks and spoons in an image.
The blue balloon is being lifted by the weight.		A surreal scene where a blue balloon appears to be lifting a weight. The weight is realistically depicted, with clear markings indicating its heaviness, while the balloon is oversized and vibrant blue, floating in a clear sky. The scene plays with the concept of reverse gravity, where the balloon, typically light and airy, is paradoxically lifting the heavy weight, creating a visually intriguing and thought-provoking image.	Failure to incorporate unusual / counterintuitive situations.
Seat numbers increasing from right to left.		Illustration of a series of seats in a minimalist room. The seats on the right start with the number '1' and the numbers increase as we move to the left. The room has a neutral color palette to ensure focus on the seat numbers.	Failure to sort the chair numbers in increasing order.

Table 13. **Failure cases of T2I generation.** We note that some Winoground queries that involve commonsense reasoning (e.g., mathematical reasoning, counting) are still too challenging even for DALL-E 3. We expect better results with stronger generative models in the future.

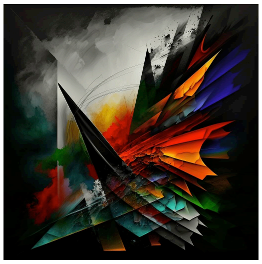
User Query	Final Image	Final Prompt	Reason for Failure
		Prompt Inversion	
		Create an abstract composition with a dynamic array of shattered, angular shapes emanating from a central point towards the edges of the image. Intensify the contrast by incorporating a deep black void at the core, surrounded by a gradient of vivid colors like red, orange, yellow, green, blue, and indigo transitioning from warm to cool tones to represent this burst of shapes. Add a contrasting background with subtle grayscale gradients, smudges, and paint splatters to enhance the sense of explosion and movement. Include sharp, crisp edges on the shapes to give a sense of three-dimensionality and depth. Ensure the overall effect is of a high-contrast, visually impactful piece that combines both geometric and organic elements, with a clear distinction between the vibrant center and the muted, textured periphery. Adjust the composition to have a more chaotic arrangement of shapes with varying sizes and directions, and incorporate a mix of both soft and hard edges to add complexity. Emphasize a more pronounced use of shadows and highlights to give the shapes a more tangible feel and enhance the illusion of depth and volume.	Highly challenging abstract details with complex atypical shapes are difficult to describe in detail, even for GPT-4V.
		Create a collection of highly detailed, anthropomorphic bird knights with meticulously crafted medieval armors and heraldic shields, standing in a 3x3 grid formation against a smooth, gradient background. Each bird should display intricate feather patterns and vibrant colors true to real bird species, with helmets thoughtfully designed to accommodate their beaks and crests. The armor should be complete with ornate shoulder plates, breastplates, gauntlets, and greaves, while the shields are to be kite or tower shield shaped, adorned with elaborate coat of arms featuring mythical creatures. Armaments will include finely wrought swords, lances, and axes. Aim for a high-fidelity 3D rendering style with a sophisticated color palette and dynamic lighting to accentuate the textures and metallic sheen of the armors, ensuring each knight is posed in a stately and dignified manner.	Failure to determine the exact number of objects in an image, especially if the values are greater than 10 or are not in a uniform pattern (grid-shaped).
		Create a 3x3x3 cube arrangement of light grey pumice stones with visible pores and rough texture, each stone equally sized and cube-shaped, on a gradient dark to light gray background with soft focused shadows and a glossy surface reflective of studio lighting. Adjust the lighting to create a more pronounced contrast, highlighting the top edges of the cubes and casting a subtle shadow on the right side, ensuring the image is sharp and high-resolution at a close-up angle to showcase the detail of the stones' textures, with the topmost center stone slightly brighter as if catching more light.	DALLE-3 fails to understand the meaning of a 3x3x3 set of cubes multiple times, and performs poorly on geometry and patterns in 3 dimensions.
		Create a collage of six images with a cosmic gastronomy theme: top-left depicts an assortment of cookies and chocolates arranged to mimic a galaxy on a space-like background; top-middle features a swirl of soft-serve ice cream in a dark cup, resembling a nebula against a starry sky; top-right displays a stack of golden brown waffles with a dusting of powdered sugar, resembling a celestial body; bottom-left shows a large, detailed moon looming over a twilight horizon; bottom-middle captures various sweets and snacks cascading onto a shadowy surface, evoking a meteor shower; bottom-right presents a cup filled with popcorn and a straw, giving the illusion of a galaxy-themed beverage, set against a backdrop of floating popcorn and sparkling stars.	DALLE-3 fails to disambiguate details between panels, eventually confusing GPT-4V as well from the comparison.

Table 14. **Failure cases of prompt inversion.** We find that our method produces suboptimal results for challenging query images. These involve images that are too abstract to describe, contain too many objects, require geometric reasoning, or involve multiple panels.

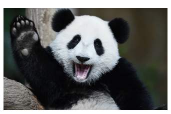
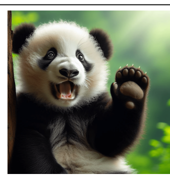
User Query	Init. Image	Final Image	Final Prompt
			**Prompt inversion** A young golden retriever puppy with a soft, fluffy coat and gentle eyes, tenderly nuzzling a small American Shorthair kitten with a curious and attentive expression. Both animals are sitting close together on a sunlit cobblestone path with patches of vibrant green moss, with the puppy's paw affectionately resting on the kitten, in the warm ambiance of a backyard during the golden hour. The background features a soft bokeh of lush greenery and the warm tones of a wooden fence, evoking a serene garden or park. The scene captures a moment of affection and camaraderie, showcasing the endearing connection between the two different species, with the sunlight casting a gentle glow and creating soft shadows on their fur.
			A person in a luxurious crimson kimono, embossed with bold indigo floral patterns, stands diminutive at the lower center of a photorealistic Japanese street as dusk settles in. Facing a grand five-tiered pagoda that ascends into the hazy sky, they hold an expansive crimson paper umbrella aloft, masking their upper body and creating an arresting visual anchor. The alley, bathed in the soft glow from the traditional wooden buildings' lanterns, stretches around them, while the cobblestone path gleams under the ambient light. In the background, life continues as silhouettes of pedestrians engage in subdued conversations or pause to photograph the scene, adding layers of depth and motion to the tranquil tableau. The pagoda, a silhouette against the misty heavens, invites the viewer's gaze upward, reinforcing the composition's sense of depth and perspective.
			Create an image of a juvenile giant panda with a striking black and white fur pattern, perched on a tree branch. The panda's mouth is agape as if mid-vocalization, and it is raising its left paw in a greeting gesture, showcasing its prominent claws. Its eyes are round and expressive, reflecting a sense of wonder. The background is a soft-focus portrayal of lush greenery, evoking a dense, misty forest atmosphere. The lighting is diffuse, with a subtle emphasis on the panda's facial features to highlight its endearing and playful demeanor.

Table 15. **Prompt inversion for natural images.** We show that our framework can also reverse engineer prompts for natural photos.

Score	Meaning
1	**Not Aligned.** The generated image shows a significant divergence from the user query. Key elements, attributes, or relationships differ notably, indicating a clear mismatch.
2	**Mildly Similar.** The generated image shows a basic level of resemblance to the user query. It includes some of the requested elements or themes, but there are significant inaccuracies or omissions, making it only loosely related.
3	**Moderately Similar.** The generated image is moderately aligned with the user query. Most of the key elements and attributes are present, but there may be some minor inaccuracies or missing details.
4	**Highly Similar.** The generated image closely aligns with the user query, accurately representing most main objects, attributes, and their relations, with only minor discrepancies.
5	**Perfect.** The generated image perfectly matches the user query in every aspect. All main objects, attributes, and their relations are exactly as requested, representing an ideal, precise match.

Table 16. **Likert scale for human evaluation.**

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1, 3

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 1

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Joyce Zhuang, Juntang andLee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiang, and Aditya Ramesh. Improving image generation with better captions. *Note on Dalle-3*, 2023. 1, 2, 3, 7, 8

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 6, 7

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[7] Yekun Chai, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Clip-tuning: Towards derivative-free prompt learning with a mixture of rewards. *arXiv preprint arXiv:2210.12050*, 2022. 3

[8] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 7

[9] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *ACL*, 2023. 3, 4

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 5, 6, 7

[11] Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P. Xing, and Zhiting Hu. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*, 2022. 3

[12] Shizhe Diao, Xuechun Li, Yong Lin, Zhichao Huang, and Tong Zhang. Black-box prompt learning for pre-trained language models. *arXiv preprint arXiv:2201.08531*, 2022. 3

[13] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 1

[14] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3-4):163–352, 2022. 3

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[16] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. *arXiv preprint arXiv:2211.11559*, 2022. 3

[17] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, 2017. 5, 7

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[19] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1): 411–420, 2017. 5

[20] Or Honovich, Uri Shaham, Samuel R Bowman, and Omer Levy. Instruction induction: From few examples to natural language task descriptions. *arXiv preprint arXiv:2205.10782*, 2022. 3

[21] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 1

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1

[23] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 1

[24] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 1

[25] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 7

[26] Li, Andreeto, Ranzato, and Perona. Caltech 101, 2022. 6, 7

[27] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 9

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 3

[29] Shuang Li, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, and Igor Mordatch. Composing ensembles of pre-trained models via iterative consensus. *arXiv preprint arXiv:2210.11522*, 2022. 3

[30] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020. 1

[31] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 3

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[33] Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. *arXiv preprint arXiv:2306.01879*, 2023. 9

[34] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramana. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. *arXiv preprint arXiv:2301.06267*, 2023. 1, 3, 5, 6

[35] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. 9

[36] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023. 3

[37] Vivian Liu. Beyond text-to-image: Multimodal prompts to explore generative ai. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2023. 7

[38] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*, 2023. 3

[39] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1

[40] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 5, 7

[41] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54 (6):1–35, 2021. 9

[42] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *ICLR*, 2023. 1, 2, 3, 5, 6

[43] Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. Reframing instructional prompts to gptk's language. *arXiv preprint arXiv:2109.07830*, 2021. 3

[44] Melanie Mitchell, John Holland, and Stephanie Forrest. When will a genetic algorithm outperform hill climbing. *Advances in neural information processing systems*, 6, 1993. 3

[45] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 7

[46] Changdae Oh, Hyeji Hwang, Hee-young Lee, YongTaek Lim, Geunyoung Jung, Jiyoung Jung, Hosik Choi, and Kyungwoo Song. Blackvip: Black-box visual prompting for robust transfer learning. In *CVPR*, 2023. 3, 5, 6, 9

[47] OpenAI. Gpt-4 technical report. 2023. 1, 2, 3, 7, 8

[48] Yassine Ouali, Adrian Bulat, Brais Matinez, and Georgios Tzimiropoulos. Black box few-shot adaptation for vision-language models. In *ICCV*, 2023. 3, 5, 6

[49] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 1, 2, 3, 4

[50] Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails of vision-language models. *arXiv preprint arXiv:2401.12425*, 2024. 9

[51] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 7

[52] Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. Grips: Gradient-free, edit-based instruction search for prompting large language models. *arXiv preprint arXiv:2203.07281*, 2022. 3

[53] Sarah Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. *ICCV*, 2023. 1, 2, 3

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 6, 11

[55] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 1

[56] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2, 9

[57] Stuart J Russell. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010. 3, 4

[58] 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. 1, 4, 5

[59] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 1, 3

[60] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Zhe Gan, Lijuan Wang, Lu Yuan, Ce Liu, et al. K-lite: Learning transferable visual models with external knowledge. *Advances in Neural Information Processing Systems*, 35:15558–15573, 2022. 3

[61] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023. 3

[62] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020. 3

[63] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6, 7

[64] Tianxiang Sun, Zhengfu He, Hong Qian, Yunhua Zhou, Xuan-Jing Huang, and Xipeng Qiu. Bbtv2: Towards a gradient-free future with large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3916–3930, 2022. 3

[65] Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*, pages 20841–20855. PMLR, 2022. 3

[66] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 2, 7, 8

[67] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 1

[68] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022. 9, 13

[69] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022. 1

[70] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021. https://arxiv.org/abs/2109.01903. 2, 5, 6

[71] Jianxiong Xiao, Krista A. Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *Int. J. Comput. Vision*, 119(1): 3–22, 2016. 7

[72] Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Yanggang Wang, Haiyu Li, and Zhilin Yang. Gps: Genetic prompt search for efficient few-shot learning. *arXiv preprint arXiv:2210.17041*, 2022. 3

[73] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 1

[74] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1

[75] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 3

[76] Mingkai Zheng, Xiu Su, Shan You, Fei Wang, Chen Qian, Chang Xu, and Samuel Albanie. Can gpt-4 perform neural architecture search? *arXiv preprint arXiv:2304.10970*, 2023. 3

[77] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 1, 2, 3, 5, 6

[78] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *ICLR*, 2023. 2, 3, 11, 12