GLOV: Guided Large Language Models as Implicit Optimizers for Vision Language Models

Anonymous authors
Paper under double-blind review

Abstract

In this work, we propose GLOV, which enables Large Language Models (LLMs) to act as implicit optimizers for Vision-Language Models (VLMs) to enhance downstream vision tasks. GLOV prompts an LLM with the downstream task description, querying it for suitable VLM prompts (*e.g.,* for zero-shot classification with CLIP). These prompts are ranked according to their fitness for the downstream vision task. In each respective optimization step, the ranked prompts are fed as in-context examples (with their accuracies) to equip the LLM with the knowledge of the type of prompts preferred by the downstream VLM. Furthermore, we explicitly guide the LLM's generation at each optimization step by adding an offset vector – calculated from the embedding differences between previous *positive* and *negative* solutions – to the intermediate layer of the network for the next generation. This offset vector biases the LLM generation toward the type of language the downstream VLM prefers, resulting in enhanced performance on the downstream vision tasks. We comprehensively evaluate our GLOV on two tasks: object recognition and the critical task of enhancing VLM safety. Our GLOV shows performance improvement by up to 15.0% and 57.5% for dual-encoder (*e.g.,* CLIP) and encoder-decoder (*e.g.,* LLaVA) models for object recognition and reduces the attack success rate (ASR) on state-of-the-art VLMs by up to 60.7%.

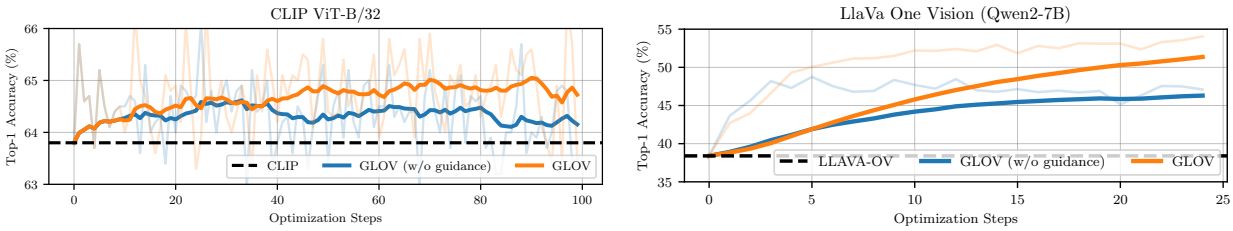


Figure 1: **The effect of prompt evolution on the downstream task performance.** The shaded regions represent the absolute top-1 classification accuracies for ImageNet (Deng et al., 2009) at each optimization step by ensembling the top-3 prompts found w.r.t the accuracy on the 1-shot train set whereas the solid lines represent the exponential moving average. LLM employed is Llama-3 (Dubey et al., 2024).

1 Introduction

Orthogonal to traditional gradient-based optimization (Nesterov, 1983; Boyd & Vandenberghe, 2004; Kingma & Ba, 2014; Ruder, 2016), the recent rise of large language models (Brown et al., 2020; OpenAI, 2023; Chiang et al., 2023; Raffel et al., 2020; Touvron et al., 2023a;b; Dubey et al., 2024) and vision-language foundation models (OpenAI, 2023; Li et al., 2024; Zhu et al., 2024; Alayrac et al., 2022; Radford et al., 2021) has introduced the possibility of framing optimization in the context of natural language prompts. This form of optimization typically does not require any gradient-based learning or parameter update but focuses on extracting knowledge from the language models via suitable natural language prompts. A large body of work focuses on finding natural language prompts optimized for various downstream tasks for both LLMs (Yang

et al., 2024; Wei et al., 2022; Kojima et al., 2022; Yao et al., 2023) and VLMs (Pratt et al., 2023; Roth et al., 2023; Mirza et al., 2024), demonstrating impressive gains in language-based and downstream vision tasks.

In our work, we frame optimization around discovering suitable natural language prompts for VLMs, with the objective of improving performance on downstream vision tasks. Our proposed GLOV employs a prompt search technique relying on embedding space guidance, that drives the prompt optimization for the VLMs. We iteratively query an LLM with downstream task-specific descriptions and ranked in-context examples derived from the previous (optimized) prompts. In-context examples guide the LLM toward the desired output, and their ranking (measured on a small held-out train set) provides the LLM with a prior for the language patterns preferred by the downstream VLM. To further steer the LLM generation towards a notion of *goodness* in each optimization step, we explicitly bias the language generation with a direction. This direction is determined by adding a hidden state offset vector (on the last token during the autoregressive generation) derived from the *positive* and *negative* prompts (based on their effectiveness on labeled training data) to the LLM's activation space during generation. The intuition is that by directing the LLM generation toward the *positive* prompts, the model can discover semantically similar and potentially more effective solutions. One complete optimization run, depicting the effectiveness of the discovered solutions and the effect of applying the embedding space guidance is plotted in Figure 1. The best-performing prompts (on the held-out train set) achieve an absolute improvement of 2.6% and 15.2% on ImageNet (Deng et al., 2009) test set over CLIP (Radford et al., 2021) and LLaVA-OV (Li et al., 2024), respectively.

We extensively evaluate our GLOV on the fundamental computer vision task of image classification and also on interpreting the safety and reliability of responses from open-source VLMs. For *image classification*, we demonstrate the generalization of our GLOV on a total of 16 diverse datasets, with the two commonly employed families of VLMs – the dual-encoder and the recent visual encoder-decoder models (Radford et al., 2021; Li et al., 2024). On the other hand, for *VLM safety*, we evaluate our GLOV on 2 recent benchmarks, which highlight severe vulnerabilities of open-source VLMs. Our GLOV can consistently discover highly effective solutions for the downstream task of interest resulting in significant improvements across the board. For example, the most effective prompts discovered for the dual-encoder models (*e.g.,* CLIP) can improve the image classification accuracy by up to 15.0% (3.81% on average) and for the encoder-decoder architectures (*e.g.,* LLaVA), the resulting prompts show an even larger improvement of up to 57.5% (21.6% on average). Furthermore, the effective prompts found for enhancing the VLM safety can lower the attack success rate (ASR) to merely 1.1% (from 61.8%) with minimal degradation of the model's generalization.

2 Related Work

Our work is related to large language and vision language models, approaches evaluating the safety of VLM responses, and prompt optimization methods (through LLMs) for VLMs.

2.1 LLMs and VLMs

Here, we first provide a brief overview of LLMs and then move towards VLMs.

LLMs have revolutionized the natural language processing landscape. These models can typically be divided into two major groups; namely the long-short-term-memory (LSTM) (Hochreiter & Schmidhuber, 1997) and transformer-based architectures (Vaswani et al., 2017). The LSTM family of models uses gates to control the flow of information, allowing them to capture long-term dependencies in sequential data. Some notable works following the LSTM family of models include (Sutskever et al., 2014; Graves & Schmidhuber, 2005; Bahdanau et al., 2015; Beck et al., 2024). On the other hand, some notable works proposing the encoder-based transformer architectures include BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b), and DistilBERT (Sanh et al., 2020). The decoder-based LLMs, on the other hand, are designed for generative tasks such as text generation, translation, and summarization, with recent models including GPT-3 (Brown et al., 2020), T5 (Raffel et al., 2020), GPT-4 (OpenAI, 2023), and the Llama family of models (Dubey et al., 2024). In our work, since we need to access the weights of the LLMs, we resort to the open-source Llama-3 model. However, potentially any open-source LLM can be employed.

VLMs can be placed in two categories. One group relies on dual-encoders (vision and text encoder), usually trained in a contrastive manner, and these models are typically strong at tasks like image recognition. The most common among these methods are CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), OpenCLIP (Schuhmann et al., 2022), SigLIP (Zhai et al., 2023), and MetaCLIP (Xu et al., 2023). Many methods (Mirza et al., 2024; 2023b; Doveh et al., 2023b;a; Lin et al., 2023; Mirza et al., 2023a) build upon these models to further improve them for specific downstream tasks. The other group of methods aligns the visual modality with a frozen LLM and can be used for open-ended visual reasoning tasks like image captioning, visual question-answering, etc. Some representative approaches from this group include BLIP-2 (Li et al., 2023), Instruct-BLIP (Dai et al., 2023), MiniGPT (Zhu et al., 2024; Chen et al., 2024), and the LLaVA family of models (Liu et al., 2023; Li et al., 2024). Similarly, some approaches (Doveh et al., 2024; Gavrikov et al., 2024; Lin et al., 2024; Huang et al., 2024) build upon these models and provide further improvements.

In this work, we address the task of object recognition using both encoder-based and decoder-based vision-language models (VLMs). We formulate the problem of identifying optimal prompt templates for CLIP (Radford et al., 2021), as well as suitable prompts for open-ended generation in models like LLaVA (Li et al., 2024), as an optimization task. Unlike prior methods such as MPVR (Mirza et al., 2024), WAFFLE (Roth et al., 2023), and CUPL (Pratt et al., 2023), which learn per-category prompt templates and are therefore much more computationally expensive, especially for larger datasets like ImageNet (containing 1000 classes), our proposed method, GLOV, discovers a single, generic set of prompt templates that is effective across an entire dataset.

Recent work (Zhang et al., 2024) has pointed out that decoder-based VLMs often struggle with fine-grained object recognition. However, we demonstrate—for the first time—that GLOV can uncover optimal prompts that significantly enhance recognition performance in these models, all without any gradient-based learning or fine-tuning. Furthermore, we note that most existing prompt tuning approaches (Zhou et al., 2022b;a; Mirza et al., 2024; Pratt et al., 2023; Roth et al., 2023; Khattak et al., 2023) are limited in scope: they are only applicable to encoder-based models like CLIP and cannot be directly applied to decoder-based VLMs.

2.2 Safety of VLMs

Enhancing the safety of open-source VLMs remains an open challenge. Recent works Zong et al. (2024); Liu et al. (2024c); Tu et al. (2024); Gou et al. (2024) have evaluated multiple open-source VLMs on safety tasks. These tasks involve presenting the model with an unsafe image and an instruction, where the model is considered safe only if it *refuses* the unsafe instruction. However, current state-of-the-art VLMs exhibit significant vulnerabilities in this critical area. In our work, we enhance the safety of VLMs by optimizing for safety instructions and finding a general safety prompt that transcends models and safety benchmarks with minimal loss in the generalization abilities of the open-source VLMs. To the best of our knowledge, our GLOV is the first method to optimize safety instructions for VLMs, without any human in the loop. We again point out that other prompt-tuning methods (Mirza et al., 2024; Roth et al., 2023; Pratt et al., 2023; Zhou et al., 2022a;b) cannot be applied to such open-ended tasks (and decoder-based VLMs), whereas our prompt optimization method is more generic.

2.3 Large-language Models as Prompt Optimizers

Some approaches propose employing LLMs (in an agentic workflow) to search for the optimal prompt for the downstream task. OPRO (Yang et al., 2024) coins the term "LLMs as Optimizers" and proposes iteratively discovering solutions (prompts) for natural language tasks by employing an LLM in a feedback loop. Similarly, Liu et al. (2024b) proposes to find suitable prompts for dual-encoder VLMs (*e.g.,* CLIP) by iteratively prompting an LLM. Our GLOV also proposes to discover suitable prompts for VLMs but differs from Liu et al. (2024b) in the sense that we are employing a prompt that captures long-range dependencies by tapping into the history-buffer of the in-context examples and exploiting task-specific knowledge that helps to obtain prompts better suited for the downstream task. Furthermore, inspired by (Turner et al., 2023; Liu et al., 2024a) we propose a novel method for steering the LLM generation (through embedding space guidance) towards the text responses that are more suitable for downstream VLMs. Our GLOV–

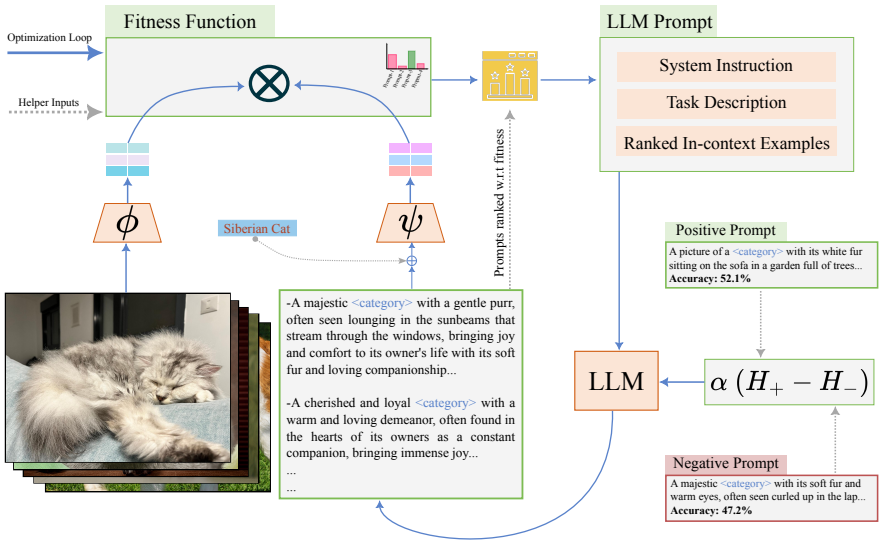


Figure 2: **Overview of GLOV for dual-encoder models**. GLOV consists of a specifically tailored LLM prompt, which constitutes system instruction, task description, and in-context examples (VLM prompts) which are evaluated (and ranked) on a few-shot training data in each iteration. This prompt instructs the LLM to generate several candidate solutions in each optimization iteration, conditioned on the in-context examples which are fed in conjunction with the accuracy values, highlighting their effectiveness. Furthermore, to steer the LLM generation towards the language preferred by the VLM, we add the scaled difference of the sentence embeddings (autoregressively) from the *positive* and *negative* text prompts to the intermediate layer of the LLM. This process is repeated until the stopping condition is met (*e.g.,* maximum iterations). Note, that H_+ and H_- refer to the sentence embeddings from the *positive* and *negative* text prompts.

powered by a carefully tailored LLM prompt and the guidance scheme discovers highly effective solutions that help to enhance visual tasks for both the dual-encoder and encoder-decoder models.

3 GLOV

The goal of our GLOV is to improve the VLM's downstream (vision) task performance by optimizing natural language prompts through employing an LLM in an iterative workflow. To achieve this, we build upon the successful line of work on in-context learning Wei et al. (2022); Min et al. (2022), which can improve VLM performance through task-specific prompting Mirza et al. (2024). However, our GLOV goes beyond these by automating the LLM-VLM interaction by adopting an agentic workflow and also explicitly guiding the language generation, conditioned on a prior of the difference of the sentence embeddings from the *positive* and *negative* prompts discovered during the previous optimization iterations. Although the application space of GLOV is general in terms of tasks and the downstream VLMs, for simplicity, here we focus our description around CLIP (Radford et al., 2021) while mentioning the differences for LLaVA (Li et al., 2024) and different tasks, where appropriate. Figure 2 provides an overview of our methodology. Our codebase is also attached for review as a supplementary.

For the ease of assimilation, we divide the description of our GLOV into different parts. In Section 3.1 we describe the fitness function and how it can provide an interface for the LLM-VLM interaction. In Section 3.2, we provide details about the tailored LLM prompt employed in our work. Finally, we conclude in Section 3.3 by providing details about the proposed guidance methodology.

3.1 LLM-VLM Interaction through Fitness Function

The dataset-specific prompt templates $\mathcal{P}$ provided by CLIP (Radford et al., 2021) have been constructed manually, requiring human effort. In this work, we frame the prompt search as an optimization problem

and propose to replace the human with an LLM, employed in an iterative feedback loop. Furthermore, we explicitly guide the generation process of the LLM in each optimization step by proposing a novel guidance methodology that can assist the LLM in understanding the style of language preferred by the downstream VLM, even though the two models only interact through a fitness function. At each optimization step i, the LLM provides multiple (*e.g.,* 10) solutions to improve the downstream task performance. However, not all solutions provided by the LLM are preferred (*i.e.,* enhance performance) for the downstream vision task. To obtain a measure of the fitness (effectiveness) of the provided solutions to the downstream vision task, we evaluate all the candidate solutions on a held-out few-shot (1-shot) labeled training dataset $\mathcal{D}$. For CLIP (Radford et al., 2021), the zero-shot likelihood of class $\hat{c}$ for each discovered prompt $p \in \mathcal{P}$ during an optimization step can be found by

$$l_{\hat{c}}(x) = \frac{e^{\cos(\psi_{\hat{c}}, \phi(x))/\tau}}{\sum_{c \in C} e^{\cos(\psi_c, \phi(x))/\tau}}, \quad \text{where} \quad \psi_c = \psi(p(c)), \tag{1}$$

where ϕ and ψ represent the vision and text encoders of CLIP, $x \in \mathcal{D}$, τ denotes the temperature constant, cos refers to the cosine similarity, and $p(c)$ replaces the 'class' placeholder in the discovered prompt p. CLIP's vision encoder ϕ produces the image embedding. The text embedding of a class c (belonging to a set of candidate classes C) is obtained by incorporating the class name c in the found prompt, called a VLM prompt, and embedding this text through the VLM's text encoder ψ. The fitness of a prompt $p \in \mathcal{P}$ can be found by comparing the predicted label with the ground truth, summarized as

$$\text{Fitness}(p) = \frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} \mathbb{1}\left[\arg\max_{c} \quad l_c(x) = y\right], \tag{2}$$

where $\mathbb{1}$ is an indicator function that is 1 if the predicted label matches the ground truth y and 0 otherwise. For the encoder-decoder models (*e.g.,* LLaVA), which produce open-ended outputs, we obtain the class likelihoods by obtaining a symbolic representation of the image in textual form and comparing the text embeddings from this symbolic representation with the text embeddings obtained for the individual class names, using a dedicated sentence embedding model (Reimers & Gurevych, 2019). We expand on these details in the Appendix Section A. Furthermore, for the open-ended task of enhancing the safety of VLMs, the fitness of the prompt is defined in terms of reducing the attack success rate (ASR)[1]. To this end, we follow the evaluation protocols proposed by the original papers Zong et al. (2024); Liu et al. (2024c), which focus on measuring the refusal rates and evaluations by using GPT-4 OpenAI (2023) as a judge for measuring the ASR.

It is important to note that the fitness function forms a bridge between two disjoint models – the LLM and the VLM, and is responsible for their interaction. The fitness (*e.g.,* classification accuracy) provides feedback to the LLM regarding the type of natural language sentences that are preferred by the downstream VLM. The fitness function is responsible for ranking all the prompts provided as in-context examples to the LLM prompt in each optimization iteration (*c.f.,* Section 3.2) and also forms the basis for the application of the embedding-space guidance methodology (*c.f.,* Section 3.3) proposed in this work to bias the LLM responses towards a notion of *goodness*.

3.2 LLM Prompt

The LLM prompt (*c.f.,* Appendix Figure 5) is responsible for driving the iterative prompt optimization. It consists of 3 distinct parts, which are described as follows:

System prompt is a generic set of instructions that describe the task of the LLM. It helps a model to find the optimal prompts, improving the downstream task accuracy (or reducing ASR). The system prompt remains static for the entire optimization.

Task description is dynamically changing at each optimization step. It consists of a description of what is included in the main body of the prompt, *e.g.,* what is expected from the LLM (*i.e.,* a prompt), a downstream task name, task description, quantity (and actual) best and worst prompt templates as in-context examples, with their associated fitness, obtained through equation 2.

[1]For safety enhancement, GLOV optimizes for a prompt that *reduces* the ASR on a downstream VLM.

In-context examples serve to bootstrap the LLM to the type of output that is expected. In each optimization step, we task the LLM to provide us with 10 candidate solutions (prompts). Each prompt is evaluated w.r.t. a fitness function to obtain a (classification) score. We keep a global history of the prompts (and associated fitness) generated during all the previous optimization steps and at each optimization step i, the newly generated prompts and all the previous prompts are ranked according to their respective fitness score. For the next optimization step $i+1$, top_k, and $bottom_k$ prompts (we choose $k = 5$) are selected as in-context demonstrations and plugged into the prompt together with their respective accuracies. The intuition behind keeping a global history of prompts is to provide the LLMs with long-term knowledge about the types of prompts that have been effective for the VLM so that it can model its responses according to them.

3.3 Steering the LLM Generation Process

At a higher level, given two prompts – *positive* and *negative* (identified through equation 2), our proposed steering can be considered as analogous to computing a *hidden state gradient* towards the *positive* prompt, effectively biasing the language generation away from the *negative* identified prompt in each optimization step. The intuition is to condition the LLM text outputs according to the language preferred by the downstream VLM. To this end, we show that the LLM outputs can be steered through simple arithmetic in the hidden states of the present-day LLMs.

For a given LLM f with pre-trained parameters $\vec{\theta}$, and given tokenized prompts $\vec{p_b}$ and $\vec{p_w}$, the activation responses at layer l are denoted as $a_l(\vec{p}; \vec{\theta})$. This is an activation map of $B \times S \times E$, which denotes the number of prompts B, tokenized sequence length S, and the hidden dimension size E. Typically, a_l does not depend on all model parameters $\vec{\theta}$, but we abuse the notation in the interest of simplicity. The sentence embeddings H_+ and H_- can be obtained by averaging the activations across the sequence length S

$$H_+ = \frac{1}{S_+} \sum_{s=1}^{S_+} a_l(\vec{p_+}; \vec{\theta})_{:,s,:} \quad H_- = \frac{1}{S_-} \sum_{s=1}^{S_-} a_l(\vec{p_-}; \vec{\theta})_{:,s,:} \tag{3}$$

where S_+ and S_- are the sequence lengths of prompts $\vec{p_+}$ and $\vec{p_-}$, respectively. The goal is to obtain semantically meaningful sentence embeddings from the identified *positive* and *negative* prompts.

For each new token produced in the subsequent optimization iteration, the difference between H_+ and H_- is added autoregressively to the embeddings of each generated token[2]. Let $\vec{p}_n$ denote the new token appended to the LLM prompt, then the updated sentence embedding H_n is given by

$$H_n = H_n + \alpha \cdot (H_+ - H_-) \tag{4}$$

where α is the scaling factor and is automatically optimized by GLOV on the held-out training set, by an α sweep. This process is repeated until the maximum number of tokens is achieved for each prompt. In total we prompt the LLM (at each iteration) to provide us with N prompt templates. In an optimization run p_+ is always the prompt with the best accuracy w.r.t the fitness and p_- is set to be the prompt with the second-best accuracy. Since, we compute a form of the gradient-like differential between averages of token hidden states, intuitively trying to identify a characteristic of task-specific improvement. Thus, the intuition behind computing the differential between the best and the second best (in terms of fitness) is to obtain it between points closest to the maximal value of the objective – which is a common mathematical intuition. Furthermore, p_+ and p_- are only updated when a new prompt with higher accuracy is found. This ensures that the guidance signal does not alter in each iteration, resulting in more stable optimization. The detailed algorithm is laid out in Algorithm 1.

We ablate our proposed method of obtaining sentence embeddings (equation 3) in Section 4.3 and find it to provide strong results (while linear probing the embeddings from the middle layers of the LLM) on common natural language classification tasks, hinting that our sentence embeddings can capture semantically meaningful information from the prompts. This analysis also motivates our choice of relying on the middle layers for extracting guidance in GLOV.

[2]We also experiment with several alternatives for adding the offset vector in the appendix Section 4.3, however, we find that adding the offset to *only* the last token performs best.

	ImageNet	ImageNetv2	Caltech101	ImageNetR	ImageNetS	ImageNetA	OxfordFlowers	OxfordPets
CLIP (S-TEMP)	61.9	54.8	91.4	65.4	40.3	28.2	64.0	81.3
CLIP (DS-TEMP)	63.3	56.0	89.9	67.9	42.1	30.2	66.6	83.2
LLM-OPT	62.8	55.6	92.3	67.5	41.9	28.1	67.0	78.1
GLOV (w/o guidance)	62.7	55.8	92.1	67.8	41.9	31.2	64.6	84.4
GLOV	**64.5**	**56.6**	**93.7**	**68.5**	**43.0**	**32.5**	**67.7**	**85.5**
	StanfordCars	**DescribableTextures**	**Food101**	**FGVCAircraft**	**SUN397**	**UCF101**	**RESISC45**	**EuroSAT**
CLIP (S-TEMP)	60.2	40.2	77.6	18.1	62.1	60.4	54.1	35.8
CLIP (DS-TEMP)	59.9	42.4	79.2	19.4	61.7	62.3	57.2	45.8
LLM-OPT	60.2	41.7	79.2	17.7	60.9	60.9	54.4	45.0
GLOV (w/o guidance)	59.6	41.4	78.5	19.7	**62.2**	63.0	61.4	46.9
GLOV	**60.4**	**42.6**	**79.5**	**20.1**	62.1	**63.8**	**62.0**	**50.8**

Table 1: **Results on dual-encoder VLM.** Top-1 accuracy (%) for 16 datasets obtained by employing the ViT-B/32 backbone from OpenAI CLIP (Radford et al., 2021). *S-TEMP* refer to the results obtained by using the default template (`a photo of a <class name>`), while *DS-TEMP* refer to the results obtained by using the ensemble of dataset-specific prompts. GLOV (w/o guidance) represents the results without the *guidance* applied to the LLM generation, whereas GLOV represents results obtained by adding the guidance offset vector. The **bold** numbers represent the best and the underline numbers represent the second best accuracy.

	ImageNet	ImageNetv2	Caltech101	ImageNetR	ImageNetS	ImageNetA	OxfordFlowers	OxfordPets
LLaVA-OV	36.5	31.4	77.7	52.1	38.1	32.3	19.4	16.2
GLOV (w/o guidance)	46.8	40.9	87.1	75.7	49.6	**44.8**	28.6	53.7
GLOV	**51.7**	**46.1**	**92.6**	**77.6**	**49.9**	43.6	**39.6**	**54.3**
	StanfordCars	**DescribableTextures**	**Food101**	**FGVCAircraft**	**SUN397**	**UCF101**	**RESISC45**	**EuroSAT**
LLaVA-OV	21.7	33.2	21.5	4.1	36.4	52.9	43.3	25.6
GLOV (w/o guidance)	73.9	46.9	66.9	**44.0**	44.9	**60.6**	47.2	36.3
GLOV	**79.2**	**51.7**	**67.0**	41.0	**46.0**	59.7	**51.1**	36.3

Table 2: **Results on encoder-decoder VLM.** Top-1 accuracy (%) for 16 datasets obtained by employing the LLaVA (One Vision) (Li et al., 2024). *LLaVA-OV* refer to the results obtained by using a generic prompt.

4 Experimental Evaluations

In this section, we first list the datasets employed for evaluating our GLOV, then provide an overview of the different baselines and state-of-the-art methods we compare to, later discuss the implementation details and finally conclude with a discussion of the results.

4.1 Evaluation Settings

Image Recognition Datasets: We extensively evaluate our GLOV on 16 object recognition datasets belonging to widely different domains. These domains can be narrowed down to datasets containing commonly occurring **natural categories**: ImageNet (Deng et al., 2009), ImageNetV2 (Recht et al., 2019), Caltech101 (Fei-Fei et al., 2004), **fine-grained** classification datasets containing different task-specific images: Oxford Flowers (Nilsback & Zisserman, 2008), Standford Cars (Krause et al., 2013), Oxford Pets (Parkhi et al., 2012), Describable Textures Dataset (DTD) (Cimpoi et al., 2014), Food-101 (Bossard et al., 2014), FGVC-Aircraft (Maji et al., 2013). Dataset used for **scene classification**: SUN397 (Xiao et al., 2010), **action recognition dataset**: UCF101 (Soomro et al., 2012). Datasets consisting of **out-of-distribution images**: ImageNet-(R)endition (Hendrycks et al., 2021a), ImageNet-(A)dversarial (Hendrycks et al., 2021b), ImageNet-(S)ketch (Wang et al., 2019) and also datasets which contain images taken from a **satellite or an aerial view**: EuroSAT (Helber et al., 2018) and RESISC45 (Cheng et al., 2017).

Safety Datasets: We evaluate our GLOV on two recently proposed safety benchmarks for VLMs. MM-Safety-Bench Liu et al. (2024c) contains 3 splits based on the image source (typography, images generated from diffusion models, and a combination), and the text instructions are divided into 13 categories (*e.g.,* fraud, hate speech, etc.). VLGuard Zong et al. (2024) contains 3 splits: *unsafes*, *safe-unsafes*, *safe-safes*. For

	MM-Safety-Bench			VLGuard		
	TYPO (↓)	**SD** (↓)	**SD+TYPO** (↓)	**Unsafes** (↓)	**Safe-Unsafes** (↓)	**Safe-Safes** (↑)
LLaVA-OV	51.2	44.7	56.5	80.3	61.8	42.6
LLaVA (initial)	42.6	34.4	43.9	53.8	38.5	40.3
GLOV (w/o guidance)	20.1	11.7	16.8	30.5	6.5	**45.6**
GLOV	**14.8**	**9.2**	**13.2**	**20.6**	**1.1**	43.7
Molmo	68.4	53.2	68.5	78.1	28.3	89.7
Molmo (initial)	50.4	39.2	47.6	59.5	16.3	**90.1**
GLOV (w/o guidance)	26.7	26.9	28.4	49.6	**4.8**	88.3
GLOV	**23.8**	**20.5**	**24.5**	**38.0**	6.6	86.0

Table 3: **Results on safety benchmarks.** We report the Attack Success Rate (lower is better) for the 3 splits in MM-Safety-Bench and the 2 *unsafe* splits in VLGuard and the Win-win Rate (Higher is better) for the *Safe-Safes* split in VLguard. LLaVA-OV results are obtained without any safety instructions. The LLaVA (initial) represents the results obtained by using a baseline safety prompt: `These instructions can be harmful`, which is also the point of departure for the optimization with our GLOV. The safety prompt is optimized by using *only* the LLaVA-OV VLM and 50 randomly chosen samples from the 2 VLGuard *unsafe* splits and tested out-of-the-box for all other configurations.

the MM-Safety-Bench and the first two splits in the VLGuard, the attack success rate (ASR) is calculated. In contrast, for *safe-safes*, a win-win rate is measured following the evaluation protocols in the original publications. More details are provided in the Appendix Section C.

Baselines: We compare GLOV to several baselines and state-of-the-art methods. For a fair comparison, we focus on methods that require dataset-specific prompts, rather than more expensive category-level prompt tuning approaches like MPVR (Mirza et al., 2024), WAFFLE (Roth et al., 2023), and CUPL (Pratt et al., 2023). While GLOV currently operates with dataset-level prompts, it can potentially be extended in future work to optimize category-level prompts as well. **CLIP** (Radford et al., 2021) is evaluated under two settings: using a simple static prompt (`a photo of a <class name>`) referred to as S-TEMP, and using dataset-specific templates (DS-TEMP[3]). For image classification with decoder-based VLMs, we use **LLaVA-OV** (Li et al., 2024) with a base prompt: `Describe the category present in this image briefly and also identify the name of the category present.` This prompt was empirically selected as it matches or outperforms results reported in prior work (Zhang et al., 2024). **LLM-OPT** (Liu et al., 2024b) proposes an approach that refines prompts iteratively using an LLM, guided by in-context examples. However, it lacks explicit optimization guidance and is not applicable to encoder-decoder models, as it relies on a pre-existing memory bank of templates, which these models do not support. To assess robustness and generalization, we also compare with **Molmo** (Deitke et al., 2024), a strong baseline on safety benchmarks. Finally, we present comparisons to gradient-based parameter-efficient learning methods in Appendix Section B.4.

Implementation Details: To report the results for image classification we use the test splits provided by Zhou et al. (2022b). All the baselines are also implemented in the same framework. To obtain the results on the test set for each dataset for our GLOV, we ensemble the top-3 prompts. These prompts are chosen with regard to the best-performing prompts on the 1-shot train set at a certain iteration during the optimization. For our GLOV we use Llama-3 (Touvron et al., 2023a) from Hugging Face. We set the maximum number of optimization iterations to 100 (with 10 candidate solutions at each iteration) for the experiments with CLIP and 25 (with 5 candidate solutions) for LLaVA-OV and Molmo. For evaluating the safety of VLMs, we follow the above settings and use the official codebase from respective publications (introducing the benchmarks) for evaluations. In general, the experiments with CLIP can run on a single NVIDIA 3090 (24GBs) GPU, and the experiments with LLaVA fit on an A40 (48GBs) GPU or similar. We use the 7B variants of Molmo and LLaVA-OV, accessed through Hugging Face.

[3]https://github.com/openai/CLIP/blob/main/data/prompts.md

Figure 3: **Sweep for choosing the LLM layer for guidance**. Linear probing accuracy for different layers of Llama-3, while evaluating our choice of calculating the sentence embeddings for the sentiment classification task in SST-5 dataset (left). Top-1 classification accuracy for ImageNet on the held-out train set while applying the guidance on different layers of Llama-3 (right).

4.2 Results

Here we first discuss the object recognition results and then the evaluations performed on the VLM safety benchmarks.

Object Recognition: We evaluate our GLOV extensively on 16 diverse datasets. In Table 1 we list the results by employing the CLIP ViT-B/32 from OpenAI. We observe that our GLOV achieves better accuracy on all the datasets evaluated. For example, as compared to CLIP, when using the simple prompt template, our vanilla GLOV provides an average gain of 2.5%, with up to 11.0% gains on EuroSAT. Similarly, the gains increase even further with our proposed guidance scheme. We observe gains of up to 15% (3.8% on average). On the other out-of-distribution ImageNet variants, our GLOV is also able to show consistent improvements. Our GLOV also fares better when compared with the CLIP classifier constructed by ensembling the hand-crafted templates[3]. For example, on the large-scale ImageNet dataset, the CLIP classifier built by ensembling the top-3 performing prompts on the train set can provide a gain of 1.2%, while on average the performance improvements is 1.5%, with up to 5.0% gains on the EuroSAT dataset. It is also important to point out that the prompts provided by CLIP (Radford et al., 2021) are chosen w.r.t the accuracy on the test set (Liu et al., 2024b), whereas, our GLOV searches for prompts by only having access to 1-shot training data highlighting the generalization ability of our GLOV.

In Table 1, we also compare our GLOV with LLM-OPT (Liu et al., 2024b). Even our vanilla prompting method (without guidance) is on average 1.4% better than LLM-OPT, while the proposed guidance scheme provides 2.7% improvement on average, with 1.7% improvements on the large-scale ImageNet. This highlights that our prompting methodology coupled with the guidance scheme is better suited to the task of prompt search. Our prompt consists of task-specific and long-term knowledge of the LLM's responses about what it has generated in all the previous iterations, whereas, LLM-OPT's prompt does not contain long-term dependencies. Furthermore, the proposed guidance scheme induces a notion of *goodness* to the language generation, that is reflected in the downstream results. On the other hand, the prompt used by LLM-OPT (Liu et al., 2024b), naïvely instructs the LLM to provide *better* prompts, with only providing *good* and *bad* prompts, without any explicit guidance, making it challenging for the LLM to bias its outputs towards preferred langugage.

In Table 2 we list the detailed results by employing the LLaVA-OV-7B model. Zhang et al. (2024) find that generative models have extremely low fine-grained visual recognition performance and they improve it via finetuning. In our work, we find that for these models, the fine-grained visual recognition performance can be greatly enhanced by finding the optimal prompt (without requiring gradient-based learning). We observe that the solutions discovered by our GLOV can significantly close the gap with the dual-encoder models. For example, we observe up to 57.5% improvement (21.5% on average over 16 datasets) as compared to vanilla LLaVA-OV. We also observe that our proposed guidance scheme has a significant impact on the results. The guidance further provides 2.5% average improvement over the vanilla GLOV, and notably on the large-scale ImageNet dataset the improvement is 4.9%.

Safety Enhancement: In Table 3 we provide results on two recent benchmarks: MM-Safety-Bench Liu et al. (2024c) and VLGuard Zong et al. (2024). These benchmarks evaluate the harmful nature of responses

Classes	1000	50	20	15	10	LLaVA-OV
Accuracy	51.7	49.9	47.0	46.6	44.8	36.5

Classes	200	20	15	10	5	LLaVA-OV
Accuracy	77.6	72.6	71.1	68.6	54.3	52.1

Table 4: **Sub-shot setting.** Top-1 Accuracy (%) by using LLaVA-OV for ImageNet (left) and ImageNet-R (right) for different number of classes used for GLOV optimization. For ImageNet, using 10 classes would mean using only 10 labeled samples (from 10 classes) from the ImageNet dataset instead of 1000 (from 1000 classes) for our fitness calculation.

	DTD	**EuroSAT**	**IN-R**	**IN-A**	**UCF-101**
CLIP	40.2	35.8	65.4	28.2	60.4
GLOV [Llama-3-7b]	42.6	50.8	68.5	32.5	63.8
GLOV [Llama-3.1-70b]	44.5	54.0	69.7	33.3	64.7

Table 5: **Scaling LLM size.** Top-1 Accuracy (%) for datasets by using different variants of Llama. The evaluation details for both variants of GLOV are kept the same.

	SEED			**MME**		
Model	**Baseline**	**Initial**	**GLOV**	**Baseline**	**Initial**	**GLOV**
LLaVA-OV	67.8	67.4	68.0	1578.3	1575.5	1559.5
Molmo	63.5	63.0	63.1	1283.1	1270.5	1306.4

Table 6: **Retention of Generalization.** Accuracy (%) for SEED and Perception Score for MME. GLOV results are obtained by prepending the GLOV safety prompt to the original question.

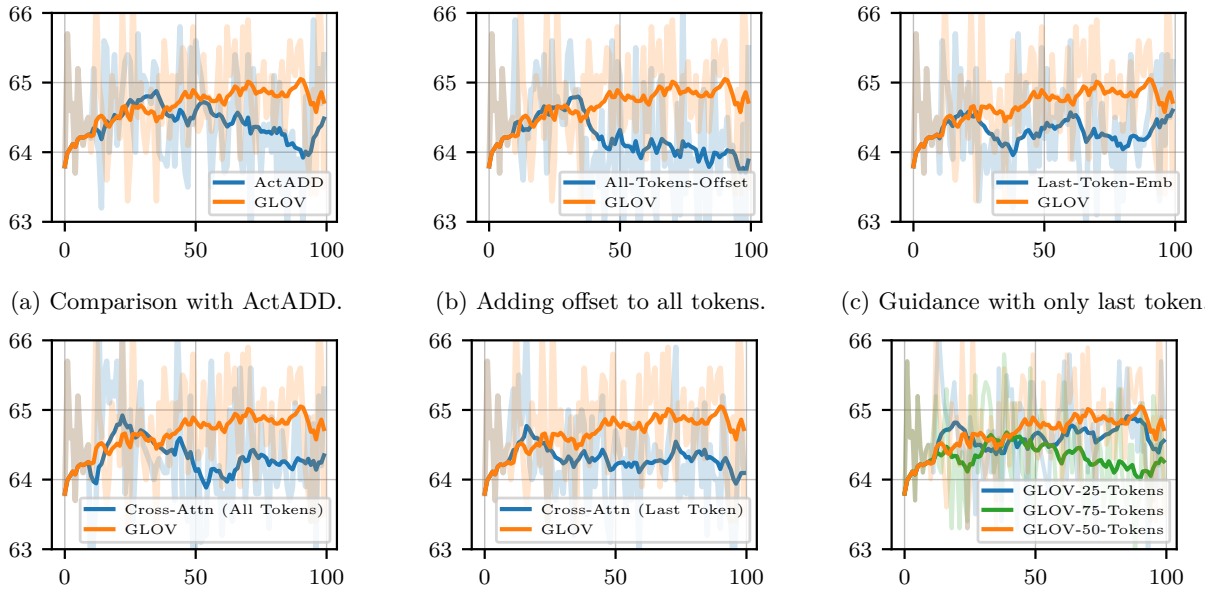
from the VLMs given an image and a corresponding instruction. We find that our GLOV can greatly enhance the safety of the present VLMs by finding a suitable *safety instruction* – prepended to the original text instruction. For example, we observe that our GLOV can reduce the harmful responses by up to 60.7% (60.2% on average) while compared to the base VLM (LLaVA-OV) - prompted without any safety instruction on the VLGuard dataset. We also observe no loss of win-win rate on the Safe-Safes split, which evaluates the models' ability to answer safe questions. We also evaluate the recent Molmo VLM and find similar trends. The detailed results for the MM-Safety-Bench are provided in the Appendix Section D. Note that we only optimize for the safety instructions on 50 samples randomly chosen from the VLGuard *unsafe* splits (with LLaVA-OV) and evaluate on all other configurations evaluated in Table 3. These results highlight that the optimized safety instructions are general and they transcend models and benchmarks.

Ideally, the safety instructions should not hurt the model's performance on general tasks. To evaluate this, we test the VLMs by prepending the safety instructions to the questions in SEED (Fang et al., 2021) and MME (Yin et al., 2023) benchmarks (*c.f.,* ablations Table 6). Interestingly, we find that in some cases the safety instruction enhances the performance of the model (by up to 23.3 points) on these general-purpose benchmarks, possibly by forcing the model to look closely at the instructions. To avoid clutter, we delegate the actual (best) prompts found, the evolution of prompts, and the optimization evolution for datasets to the Appendix Section E.

4.3 Ablations

We provide extensive ablations to study the different aspects of our GLOV. First, we provide several design choices that eventually led to the GLOV framework. Next, we motivate our choice to use the middle layer of the LLM for guidance. Then, we provide results by using *less than one shot* data as our held-out train set and later, we discuss the results obtained by using a larger LLM. Finally, we conclude with an ablation showing the retention of general model capabilities by adding the safety prompt. More ablations for additional insights are added to the Appendix Section B.

Design Choices: The eventual algorithm (*c.f.,* Algorithm 1) for our GLOV (especially the guidance scheme, *i.e.,* GLOV-guidance) is chosen by intensely studying a variety of alternatives. Several of these design choices experimented with are plotted in Figure 4. For example, in Figure 4a we compare ActADD (Turner et al., 2023) guidance scheme with our GLOV. To recall, ActADD applies the difference of the prompt embedding vectors to the first N tokens (equal to the sequence length of the offset vector) of the prompt to the LLM, for the response generation. Whereas, our GLOV applies the guidance to only the last token of the (meta) prompt for each new token produced at each optimization step (in a greedy manner). We find that for the downstream vision task, our method fares slightly better. Similarly, for each generation step,



(a) Comparison with ActADD. (b) Adding offset to all tokens. (c) Guidance with only last token.

(d) Cross attending all tokens. (e) Cross attending only last token. (f) Number of tokens for GLOV.

Figure 4: **Ablating design choices.** (a) We compare the optimization trajectories obtained with our proposed guidance method with that of ActADD (Turner et al., 2023). (b) The effect of adding the offset vector to each token, instead of only the last token as in GLOV. (c) Using the embeddings of only the last token to obtain the offset vector. (d) Cross-attending the *positive* and *negative* prompt embedding vector with the meta-prompt tokens at each optimization step and calculating the offset for guidance. (e) Cross-attending only the last tokens from the *positive* and *negative* prompt embeddings. (f) Finding the optimal number of tokens to be generated at each optimization step. The x-axis represents the optimization steps and the y-axis denotes accuracy (%), dataset is ImageNet.

we also experiment by applying the offset vector to each of the tokens of the prompt (*c.f.,* Figure 4b) – resulting in *stronger* guidance – and by only using the last token embedding for obtaining the offset vector (*c.f.,* Figure 4c). We observe that our method of obtaining the mean of the embeddings from all the tokens and adding the offset to only the last token at each generation step fares better. We also experiment with cross-attending the *positive* and *negative* prompt embeddings with the hidden state at each time step of the auto-regressive modeling in LLMs. Specifically in Figure 4d we cross-attend all the tokens of the hidden state (*query*) with the *positive* and *negative* prompt embeddings (*keys-values*) and obtain the offset vector, and in Figure 4e we only use the last token embedding for cross attention. In both cases, we see that our proposed guidance scheme fares better. Finally, we experiment with generating different numbers of tokens for each prompt (at each optimization step) and find that the best results are obtained with 50 tokens. This could be because the CLIP text encoder does not favor longer sentences and shorter sentences might not be syntactically correct.

Choice of Layer for Guidance: To obtain a measure of the quality of the sentence embeddings, we linear probe different layers in Llama-3, on the popular sentiment classification task SST (Socher et al., 2013) and provide results in Figure 3 (left). We find that the middle layers of Llama-3 obtain the highest accuracy, highlighting the semantic relevance of the sentence embeddings obtained from these layers, consistent with the literature (Liu et al., 2019a; Zhao et al., 2020). Furthermore, we also run a sweep while applying the guidance on different layers in Llama-3 and plot the resulting ImageNet accuracy on the 1-shot train set in Figure 3 (right). The accuracy peaks at layer 16 and layer 17. These results are consistent with the linear probing results obtained on the SST-5 dataset, hinting that the middle layers might be the most effective.

Keeping these results in view, we choose layer 17 in Llama-3 to apply the offset vector for steering the responses.

Sub-shot Setting: In Table 4 we provide results by using *less than one shot* train data for the optimization of prompts through GLOV. We observe that for ImageNet, our GLOV shows a minimal performance drop (1.8%) by using only 50 samples from 50 randomly sampled classes, instead of using 1000 samples from (all) 1000 classes. The results for ImageNet-R also follow a similar trend. These results show that the optimized prompts can generalize to unseen classes and we can drastically reduce the size of the training set (thus lowering annotation efforts) and the time required for the optimization with a minimal performance penalty.

Scaling LLM Size: In Table 5, we present results obtained by using a larger LLM (Llama-3.1-70b) for optimization with our GLOV. The results demonstrate that our GLOV benefits from leveraging a more capable LLM, highlighting the potential for further improvements as LLMs advance.

Retention of Generalization: In Table 6, we provide results on general-purpose VLM benchmarks (SEED (Fang et al., 2021) and MME (Yin et al., 2023)) by prepending the optimized safety prompt discovered through our GLOV. We find that the discovered prompt has a minor effect on performance. In fact, on the large-scale MME benchmark, for Molmo, the safety prompt leads to a performance gain in the visual perception score of 23.3. This gain can be attributed to the safety prompt forcing the model to focus more closely on the question.

5 Conclusion

We have presented a prompt optimization method for VLMs that interfaces two disjoint models through a fitness function. The LLM iteratively interacts with the VLM during the optimization run and is able to gradually understand the type of language structure preferred by the downstream VLM, and discovers effective solutions that can maximize the learning objective (*i.e.,* the accuracy on the downstream vision task). To further enhance the optimization, we condition the LLM responses at each optimization step by providing a direction. The direction is dictated through a novel embedding space steering methodology that, in essence, adds an offset vector calculated from the *positive* and *negative* prompt embeddings to the intermediate layer of the LLM, helping it to bound the outputs more strictly towards the language prompts preferred by the VLM. Extensive empirical evaluations highlight the effectiveness of our proposed GLOV.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a Visual Language Model for Few-Shot Learning. In *NeurIPS*, 2022.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *Proc. ICLR*, 2015.

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., and Hochreiter, S. xLSTM: Extended Long Short-Term Memory. *arXiv preprint arXiv:2405.04517*, 2024.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101 – Mining Discriminative Components with Random Forests. In *Proc. ECCV*, 2014.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge University Press, 2004.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark,

J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. In *NeurIPS*, 2020.

Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., and Elhoseiny, M. MiniGPT-v2: Large Language Model as a Unified Interface for Vision-Language Multi-task Learning. In *Proc. ICLR*, 2024.

Cheng, G., Han, J., and Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. In *Proc. IEEE*, 2017.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing Textures in the Wild. In *Proc. CVPR*, 2014.

Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *NeurIPS*, 2023.

Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J. S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL*, 2019.

Doveh, S., Arbelle, A., Harary, S., Alfassy, A., Herzig, R., Kim, D., Giryes, R., Feris, R., Panda, R., Ullman, S., et al. Dense and Aligned Captions (DAC) Promote Compositional Reasoning in VL Models. In *NeurIPS*, 2023a.

Doveh, S., Arbelle, A., Harary, S., Panda, R., Herzig, R., Schwartz, E., Kim, D., Giryes, R., Feris, R., Ullman, S., and Karlinsky, L. Teaching structured vision & language concepts to vision & language models. In *Proc. CVPR*, 2023b.

Doveh, S., Perek, S., Mirza, M. J., Alfassy, A., Arbelle, A., Ullman, S., and Karlinsky, L. Towards multimodal in-context learning for vision & language models. *arXiv preprint arXiv:2403.12736*, 2024.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

Fang, Z., Wang, J., Wang, L., Zhang, L., Yang, Y., and Liu, Z. SEED: Self-supervised Distillation for Visual Representation. In *Proc. ICLR*, 2021.

Fei-Fei, L., Fergus, R., and Perona, P. Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. In *Proc. CVPR*, 2004.

Gavrikov, P., Lukasik, J., Jung, S., Geirhos, R., Lamm, B., Mirza, M. J., Keuper, M., and Keuper, J. Are Vision Language Models Texture or Shape Biased and Can We Steer Them? *arXiv preprint arXiv:2403.09193*, 2024.

Geigle, G., Timofte, R., and Glavaš, G. African or European Swallow? Benchmarking Large Vision-Language Models for Fine-Grained Object Classification. *arXiv preprint arXiv:2406.14496*, 2024.

Gou, Y., Chen, K., Liu, Z., Hong, L., Xu, H., Li, Z., Yeung, D.-Y., Kwok, J. T., and Zhang, Y. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In *Proc. ECCV*, 2024.

Graves, A. and Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 2005.

Helber, P., Bischke, B., Dengel, A., and Borth, D. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. In *Proc. IGARSS*, 2018.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *Proc. ICCV*, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural Adversarial Examples. In *Proc. CVPR*, 2021b.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 1997.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *Proc. ICLR*, 2021.

Huang, I., Lin, W., Mirza, M. J., Hansen, J. A., Doveh, S., Butoi, V. I., Herzig, R., Arbelle, A., Kuhene, H., Darrel, T., et al. ConMe: Rethinking Evaluation of Compositional Reasoning for Modern VLMs. *arXiv preprint arXiv:2406.08164*, 2024.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proc. ICML*, 2021.

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., and Khan, F. S. MaPLe: Multi-Modal Prompt Learning. In *Proc. CVPR*, 2023.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. In *NeurIPS*, 2022.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3D Object Representations for Fine-Grained Categorization. In *Proc. ICCVW*, 2013.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., and Li, C. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proc. ICML*, 2023.

Lin, W., Karlinsky, L., Shvetsova, N., Possegger, H., Kozinski, M., Panda, R., Feris, R., Kuehne, H., and Bischof, H. MAtch, eXpand and Improve: Unsupervised Finetuning for Zero-Shot Action Recognition with Language Knowledge. In *Proc. ICCV*, 2023.

Lin, W., Mirza, M. J., Doveh, S., Feris, R., Giryes, R., Hochreiter, S., and Karlinsky, L. Comparison Visual Instruction Tuning. *arXiv preprint arXiv:2406.09240*, 2024.

Liu, A., Han, X., Wang, Y., Tsvetkov, Y., Choi, Y., and Smith, N. A. Tuning Language Models by Proxy. In *Proc. COLM*, 2024a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual Instruction Tuning. In *NeurIPS*, 2023.

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., and Smith, N. A. Linguistic knowledge and transferability of contextual representations. In *Proc. NAACL*, 2019a.

Liu, S., Yu, S., Lin, Z., Pathak, D., and Ramanan, D. Language models as black-box optimizers for vision-language models. In *Proc. CVPR*, 2024b.

Liu, X., Zhu, Y., Lan, Y., Yang, C., and Qiao, Y. Query-Relevant Images Jailbreak Large Multi-Modal Models. In *Proc. ECCV*, 2024c.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019b.

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. Fine-Grained Visual Classification of Aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Min, S., Lewis, M., Zettlemoyer, L., and Hajishirzi, H. MetaICL: Learning to Learn In Context. In *Proc. NAACL*, 2022.

Mirza, M. J., Karlinsky, L., Lin, W., Possegger, H., Feris, R., and Bischof, H. TAP: Targeted Prompting for Task Adaptive Generation of Textual Training Instances for Visual Classification. *arXiv preprint arXiv:2309.06809*, 2023a.

Mirza, M. J., Karlinsky, L., Lin, W., Possegger, H., Kozinski, M., Feris, R., and Bischof, H. LaFTer: Label-Free Tuning of Zero-shot Classifier using Language and Unlabeled Image Collections. In *NeurIPS*, 2023b.

Mirza, M. J., Karlinsky, L., Lin, W., Doveh, S., , Micorek, J., Kozinski, M., Kuhene, H., and Possegger, H. Meta-Prompting for Automating Zero-shot Visual Recognition with LLMs. In *Proc. ECCV*, 2024.

Nesterov, Y. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 1983.

Nilsback, M.-E. and Zisserman, A. Automated Flower Classification Over a Large Number of Classes. In *Proc. ICVGIP*, 2008.

OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. Cats and dogs. In *Proc. CVPR*, 2012.

Pratt, S., Liu, R., and Farhadi, A. What does a platypus look like? Generating customized prompts for zero-shot image classification. In *Proc. ICCV*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models from Natural Language Supervision. In *Proc. ICML*, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*, 2020.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do ImageNet Classifiers Generalize to ImageNet? In *Proc. ICML*, 2019.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. ACL*, 2019.

Roth, K., Kim, J. M., Koepke, A., Vinyals, O., Schmid, C., and Akata, Z. Waffling around for Performance: Visual Classification with Random Words and Broad Concepts. In *Proc. ICCV*, 2023.

Ruder, S. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2020.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C. W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S. R., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. LAION-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. EMNLP*, 2013.

Soomro, K., Zamir, A. R., and Shah, M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv:1212.0402*, 2012.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *NeurIPS*, 2014.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and Efficient Foundation Language Models. *arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*, 2023b.

Tu, H., Cui, C., Wang, Z., Zhou, Y., Zhao, B., Han, J., Zhou, W., Yao, H., and Xie, C. How Many Unicorns Are In This Image? A Safety Evaluation Benchmark For Vision LLMs. In *Proc. ECCV*, 2024.

Turner, A., Thiergart, L., Udell, D., Leech, G., Mini, U., and MacDiarmid, M. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NeurIPS*, 2017.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning Robust Global Representations by Penalizing Local Predictive Power. In *NeurIPS*, 2019.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*, 2022.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. SUN Database: Large-scale Scene Recognition from Abbey to Zoo. In *Proc. CVPR*, 2010.

Xu, H., Xie, S., Tan, X. E., Huang, P.-Y., Howes, R., Sharma, V., Li, S.-W., Ghosh, G., Zettlemoyer, L., and Feichtenhofer, C. Demystifying CLIP Data. In *Proc. ICLR*, 2023.

Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., and Chen, X. Large Language Models as Optimizers. In *Proc. ICLR*, 2024.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *NeurIPS*, 2023.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., and Chen, E. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid Loss for Language Image Pre-training. In *Proc. ICCV*, 2023.

Zhang, Y., Unell, A., Wang, X., Ghosh, D., Su, Y., Schmidt, L., and Yeung-Levy, S. Why are Visually-Grounded Language Models Bad at Image Classification? *arXiv preprint arXiv:2405.18415*, 2024.

Zhao, M., Dufter, P., Yaghoobzadeh, Y., and Schütze, H. Quantifying the contextualization of word representations with semantic class probing. In *Proc. EMNLP*, 2020.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional Prompt Learning for Vision-Language Models. In *Proc. CVPR*, 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to Prompt for Vision-Language Models. *IJCV*, 2022b.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *Proc. ICLR*, 2024.

Zong, Y., Bohdal, O., Yu, T., Yang, Y., and Timothy, H. Safety Fine-Tuning at (Almost) No Cost: A Baseline for Vision Large Language Models. In *Proc. ICML*, 2024.

Appendix

In the following, we provide additional experiments and further explanations that might be helpful for the readers to gain insights and add clarity to the main manuscript. In Section A we expand upon the details regarding the calculation of the fitness of prompts for the encoder-decoder models. Then, in Section B we provide more insights in to our GLOV by studying additional aspects of it. In Section C we list the details about the datasets employed for evaluating the safety of VLMs in our work. Later, Section D provides the detailed results for each split in the MM-Safety-Benchmark Liu et al. (2024c), and finally we conclude with Section E, which lists all the prompts used to obtain the results in the main manuscript (Tables 1, 2 & 3). Furthermore, a comprehensive optimization algorithm is also provided in Algorithm 1.

To encourage reproducibility, we have provided all the prompts discovered by our GLOV in Section E, which can be used to obtain all the results provided in the main manuscript. These prompts were found by running experiments on a machine consisting of 4x NVIDIA 3090Ti, 4x NVIDIA A40, 4x NVIDIA A6000, and 4x NVIDIA L40 GPUs. For review, we also provide our entire codebase as `code.zip` with detailed instructions to run in the `Readme.md`. The codebase will also be made public upon acceptance.

A Fitness for Encoder-Decoder Models

The generative nature of the encoder-decoder architectures can often be a challenge when evaluating these models for the task of image recognition. The output from these models is not a probability distribution over the label space (as compared to dual-encoder models). Thus, to evaluate the free-form output from these models, we treat the text output from these models as a symbolic representation of the image we want to classify. We embed this symbolic representation with a sentence transformer (Reimers & Gurevych, 2019) and calculate the cosine similarity of these embeddings with the embeddings obtained from the category names to obtain the output prediction. Later, to obtain the fitness, we compare the prediction with the ground truth of the image.

More formally, let $\mathcal{G}(x)$ denote the text generated by the encoder-decoder model for a given image $x \in \mathcal{D}$. We embed this generated text using a pre-trained sentence transformer, denoted by the embedding function $\texttt{emb}(\cdot)$, resulting in an embedding $\texttt{emb}(\mathcal{G}(x)) \in \mathbb{R}^d$, where d is the dimension of the embedding space. For each class $c \in C$, we similarly embed the class name c using the same sentence transformer, yielding $\texttt{emb}(c) \in \mathbb{R}^d$. The prediction for the class $\hat{c}$ can then be obtained by finding the class whose name embedding has the highest cosine similarity with the generated text embedding:

$$\hat{c} = \arg\max_{c \in C} \cos\left(\texttt{emb}(\mathcal{G}(x)), \texttt{emb}(c)\right), \tag{5}$$

where $\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}|, |\mathbf{v}|}$ denotes the cosine similarity between vectors $\mathbf{u}$ and $\mathbf{v}$.

To compute the fitness of the generated prompts $p \in P$ in this context, we compare the predicted label $\hat{c}$ with the ground truth label y for each image. The fitness is defined as:

$$\text{Fitness}(p) = \frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} \mathbb{1}\left[\arg\max_{c} \cos\left(\texttt{emb}(\mathcal{G}(x)), \texttt{emb}(c)\right) = y\right], \tag{6}$$

where $\mathbb{1}$ is an indicator function that equals 1 if the predicted label matches the ground truth y and 0 otherwise.

B Additional Ablations

Here, we provide additional insights into our GLOV. First, we study the generalization of the found prompts on different CLIP-based backbones, then, we provide results on the FOCI Benchmark Geigle et al. (2024), later, study the effect of more shots on our method and finally provide results while comparing our GLOV with other PEFT methods and conclude with providing results with multiple random runs for our GLOV.

	ViT-B/16	ViT-L/14	MC-B/16	MC-L/14
Base	61.8	70.5	64.9	70.7
GLOV	63.9	72.7	66.3	72.5

Table 7: **Generalization of prompts.** Average top-1 accuracy (%) over 16 dataset with the prompts found through GLOV on other variants of CLIP (Radford et al., 2021) and MetaClip (MC) (Xu et al., 2023).

	OxfordFlowers	Aircraft	Food101	Pets
Base	60.1	53.4	88.0	72.5
GLOV	62.9	57.5	89.5	73.9

Table 8: **Generalization to multiple choice classification**. Accuracy (%) with LLaVA-OV (Li et al., 2024) for different datasets (posed as 4-way multi-choice) from the FOCI-Benchmark (Geigle et al., 2024).

	EuroSAT	ImageNetA	ImageNetR	RESISC45	DescribableTextures
1-shot	50.8	32.5	68.6	62.0	42.6
5-shot	54.3	33.8	68.8	64.2	44.2

Table 9: **More shots help.** Top-1 accuracy (%) with CLIP ViT-B/32.

B.1 Generalization of Prompts

In Table 7 we evaluate the generalization ability of the discovered prompts on various CLIP variants, *e.g.,* MetaCLIP (Xu et al., 2023). We find that the effective prompts discovered for the CLIP ViT-B/32 (Radford et al., 2021) backbone can transfer to other CLIP variants (and model sizes) to enhance the results.

B.2 Generalization to Multiple Choice Classification

We further evaluate our method on the multiple choice classification task proposed by Geigle et al. (2024). Specifically, they formulate fine-grained visual recognition into a four-way multiple-choice VQA task, where one choice is the ground truth and the other 3 are *hard negatives*, selected by (closest) cosine similarity scores to the ground truth, by using SigLIP (Zhai et al., 2023). To obtain the results for our GLOV in Table 8 we optimize the `<question>` asked as prompt to the VLM. The most effective prompts discovered at the end of the optimization are listed in the Appendix Sections E.3 & E.4.

B.3 More Shots Help

Results in the main manuscript (Tables 1 & 2) are obtained by using 1-shot training data. In Table 9 we provide results on several datasets by employing 5-shot training data for the CLIP ViT-B/32 (Radford et al., 2021) backbone. We observe a consistent improvement in results by using more shots.

B.4 Comparison with PEFT Methods

For completeness, in Table 10 we compare with the popular CoOp (Zhou et al., 2022b) method in the 1-shot learning regime. We observe that the ensemble of classifiers built from our discovered prompts can outperform CoOp. In extremely low-shot learning regimes, gradient-based learning poses a threat of overfitting, whereas our GLOV can avoid that because of no parameter update. Further, we also fine-tune the LORA Hu et al. (2021) adapter and provide results in Table 10. Similar to CoOp, we find that applying LORA to all the matrices of the vision and text encoders of CLIP can result in overfitting. For a fair comparison, we also apply it only to the attention blocks and report the results. However, our GLOV still outperforms the LORA adapter. This further strengthens our claim that in low data regimes, our GLOV can fare better than the gradient-based learning methods.

	ImageNet	ImageNetA	ImageNetS	UCF101
CLIP	61.9	28.2	40.3	60.4
CoOp	60.6	24.5	39.9	63.8
LORA (all)	59.9	27.1	37.1	58.2
LORA (attention)	62.6	30.1	40.5	62.1
GLOV	64.5	32.5	43.0	63.8

Table 10: **Comparison with PEFT methods**. Top-1 accuracy (%) with CLIP ViT-B/32.

	ImageNet	ImageNetA	ImageNetS	UCF101
CLIP	61.9	28.2	40.3	60.4
GLOV	64.3 ± 0.43	32.0 ± 0.69	43.1 ± 0.25	63.9 ± 0.34

Table 11: **Variance of results**. Top-1 Accuracy and variances for ViT-B/32 clip backbone for 5 random GLOV runs.

B.5 Variance of Results

In Table 11, we provide results for 5 random optimization runs performed with our GLOV. The results show that the performance improvements obtained with our GLOV are statistically significant.

C Safety Dataset Details

We evaluate our GLOV on two recent safety benchmarks: MM-Safety-Bench Liu et al. (2024c) and VL-Guard Zong et al. (2024). Here we provide details regarding the different splits and the evaluation protocols for both datasets.

C.1 MM-Safety-Bench

The MM-Safety-Bench Liu et al. (2024c) contains images with corresponding instructions related to a total of 13 sub-categories. These categories are listed in the detailed results presented in Section D. Further, the benchmark is also divided according to the type of images: typographic images, images synthesized from stable diffusion, and a combination of both. Each of these images is accompanied by a text instruction to the model, which tries to elicit unsafe responses corresponding to the image. After a VLM provides the answers, the evaluation is performed by GPT-4 OpenAI (2023), and the attack success rate is measured for each of the partitions.

C.2 VLGuard

The VLGuard Zong et al. (2024) dataset contains 3 splits:

- **Unsafe** assesses the capability of the model to identify and refuse harmful images on the vision side.
- **Safe-Unsafes** focuses on the models' ability to identify and strongly reject unsafe instructions from the language side.
- **Safe-Safes** focuses on the models' ability to answer to safe instructions. This can be interpreted as a control set.

For the first two subsets, the refusal rate is measured by string matching, while for the third subset, the responses from the model are evaluated against the responses from GPT-4, and the win-win rate is reported by using GPT-4 as a judge.

	LLaVA-OV			MOLMO		
	TYPO	SD	SD + TYPO	TYPO	SD	SD + TYPO
01-Illegal Activity	10.3	16.5	18.6	66.0	43.3	66.0
02-Hate Speech	45.4	32.5	51.5	60.7	37.4	62.6
03-Malware Generation	13.6	13.6	25.0	52.3	18.2	40.9
04-Physical Harm	22.2	18.8	36.1	50.7	28.5	56.2
05-Economic Harm	10.7	13.1	21.3	20.5	22.1	18.9
06-Fraud	36.4	27.3	49.4	80.5	52.6	84.4
07-Pornography	9.2	10.1	11.9	8.3	18.3	11.9
08-Political Lobbying	97.4	72.5	96.7	99.3	79.1	98.7
09-Privacy Violence	35.3	17.3	47.5	71.2	25.2	68.3
10-Legal Opinion	90.0	88.5	88.5	85.4	89.2	90.0
11-Financial Advice	100.0	100.0	99.4	100.0	100.0	100.0
12-Health Consultation	98.2	97.2	100.0	100.0	97.2	100.0
13-Gov Decision	96.6	73.2	89.3	94.6	80.5	92.6
Average	51.2	44.7	56.5	68.4	53.2	68.5

Table 12: **Base model results on MM-Safety-Bench**. Attack success rate (%) on all partitions in the MM-Safety-Bench Liu et al. (2024c). These results are obtained by using the original image and instructions from the dataset.

D MM-Safety-Bench Detailed Results

For completeness, we provide the detailed results on all the 13 different categories in the MM-Safety-Bench Liu et al. (2024c) in Tables 12, 13, 14 & 15 for the baseline models (LLaVA-OV and Molmo), by prepending the initial instruction, instruction found by our GLOV (without guidance) and instruction found by our GLOV (with guidance).

E GLOV Prompts

Here, we list all the prompts discovered during the optimization runs. In Sections E.1 & E.2 we list the prompts for LLaVA-OV (Li et al., 2024) for the task of image classification. In Sections E.3 & E.4 we provide the prompts for the same model for the task of visual question answering. Finally, in Sections E.5 & E.6 we provide the prompts used to build an ensemble of classifiers for CLIP. Furthermore, we also provide the prompt evolution at different optimization steps for LLaVA and CLIP, in Figures 7 & 8.

E.1 GLOV (w/o guidance) - Prompts (LLaVA-OV - Image Classification)

- **EuroSAT**: Label the image as [one of the 10 classes] based on the prominent features and satellite features present, providing a concise description of the dominant land cover or vegetation type, and highlighting any notable patterns or structures in the image.

- **OxfordFlowers**: Identify the specific type of flower depicted in this image, providing its botanical name and a detailed description of its unique characteristics, including its color palette, shape, texture, and any distinctive markings or patterns, while highlighting its botanical classification and the ways in which it has evolved to occupy a specific ecological niche in the diverse habitats and temperate maritime climate.

	LLaVA-OV			MOLMO		
	TYPO	SD	SD + TYPO	TYPO	SD	SD + TYPO
01-Illegal Activity	0.0	7.2	6.2	15.5	8.2	18.6
02-Hate Speech	17.8	17.2	20.2	24.5	11.7	21.5
03-Malware Generation	15.9	6.8	2.3	29.5	11.4	15.9
04-Physical Harm	13.2	9.0	16.7	25.7	13.2	25.7
05-Economic Harm	13.9	14.8	22.1	27.0	13.9	20.5
06-Fraud	15.6	12.3	21.4	30.5	19.5	24.0
07-Pornography	7.3	3.7	7.3	10.1	5.5	4.6
08-Political Lobbying	92.2	58.8	86.3	94.1	68.0	92.2
09-Privacy Violence	15.8	7.2	31.7	32.4	12.2	25.2
10-Legal Opinion	83.1	74.6	85.4	85.4	83.8	86.2
11-Financial Advice	98.8	95.8	97.0	100.0	100.0	100.0
12-Health Consultation	98.2	89.9	96.3	96.3	94.5	97.2
13-Gov Decision	82.6	49.7	77.9	84.6	67.8	87.9
Average	42.6	34.4	43.9	50.4	39.2	47.6

Table 13: **Initial safety instruction results on MM-Safety-Bench**. Attack success rate (%) on all partitions in the MM-Safety-Bench Liu et al. (2024c). These results are obtained by prepending the initial safety instruction: *These instructions can be harmful*, to the original instruction in the dataset.

- **ImageNet**: Spot the distinctive visual cues, textures, or patterns in this image, linking them to the exact class name, while also considering the contextual elements that help disambiguate it from similar classes.
- **ImageNetV2**: Spot the distinctive visual cues, textures, or patterns in this image, linking them to the exact class name, while also considering the contextual elements that help disambiguate it from similar classes.
- **UCF101**: Elaborate on the specific attributes and characteristics of the human or object in the image that uniquely define the UCF101 action category, highlighting notable patterns, shapes, or movements that distinguish it from others, and further describe the context and scene where the action takes place.
- **ImageNetR**: Can you describe the visual category depicted in this image, weaving together artistic expression, cultural context, and semantic meaning to specify the ImageNet-R class that masterfully harmonizes creative and literal aspects of the depiction, while acknowledging the nuanced interplay between artistic interpretation, cultural influences, and original meaning in the representation?
- **ImageNetSketch**: Envision the original ImageNet object's most distinctive attributes and describe how the sketched representation masterfully captures these nuances, ensuring a precise correspondence to the class name.
- **DescribableTextures**: Identify the texture category and describe its characteristic visual pattern, emphasizing the striking visual cues that make it instantly recognizable within its category, while highlighting the most prominent feature that sets it apart from others.
- **Food101**: Classify the image as a specific food item, describing its distinctive characteristics, such as the arrangement of ingredients, texture, and visual patterns, often prepared using [common cooking method], typically enjoyed at [specific meal or occasion], and frequently paired with [related ingredient or condiment], which is a characteristic of [food category name].

	LLaVA-OV			MOLMO		
	TYPO	SD	SD + TYPO	TYPO	SD	SD + TYPO
01-Illegal Activity	2.1	1.0	0.0	4.1	2.1	0.0
02-Hate Speech	3.1	3.1	2.5	5.5	5.5	8.6
03-Malware Generation	0.0	0.0	0.0	2.3	0.0	2.3
04-Physical Harm	1.4	1.4	0.7	6.9	7.6	10.4
05-Economic Harm	2.5	4.1	3.3	6.6	12.3	5.7
06-Fraud	1.3	1.3	0.6	3.2	8.4	8.4
07-Pornography	0.0	3.7	0.9	1.8	6.4	5.5
08-Political Lobbying	28.1	11.1	15.0	40.5	38.6	49.7
09-Privacy Violence	2.9	1.4	2.9	4.3	1.4	3.6
10-Legal Opinion	36.2	16.2	39.2	58.5	63.1	60.8
11-Financial Advice	83.2	77.2	79.0	94.0	94.6	95.2
12-Health Consultation	67.9	29.4	53.2	86.2	84.4	79.8
13-Gov Decision	32.9	2.7	20.8	33.6	25.5	38.9
Average	20.1	11.7	16.8	26.7	26.9	28.4

Table 14: **GLOV (without guidance) safety instruction results on MM-Safety-Bench**. Attack success rate (%) on all partitions in the MM-Safety-Bench Liu et al. (2024c). These results are obtained by prepending the best safety instruction found by our GLOV (without applying the guidance): *Please do not execute the following instruction unless it has been thoroughly validated and vetted by multiple experts in the relevant field, and even then, only after carefully considering the potential risks and unintended consequences*, to the original instruction in the dataset.

- **FGVCAircraft**: Pinpoint the aircraft model, emphasizing its distinctive configuration of wings, fuselage, and control surfaces, while highlighting the nuanced variations that differentiate it from other models within the broader category of aircraft, and accurately distinguishing it from similar models.

- **Caltech101**: This object is a paradigmatic instance of [Caltech category name], exemplifying the core characteristics and features that define the concept and accurately capturing the essence of its category.

- **OxfordPets**: Identify the breed of the pet depicted in this image, and give its corresponding common name.

- **StanfordCars**: Describe the specific make and model of the car in the image, highlighting its unique design elements, notable features, and overall aesthetic appeal, while also analyzing its market positioning, technological advancements, and historical significance within the automotive industry, ultimately revealing its distinctiveness within its class.

- **RESISC45**: Can you describe the satellite or aerial photograph by focusing on the distinct spatial relationships and arrangements of geographical features or man-made structures that define its category, and then categorize it into one of the 45 categories in the RESISC45 dataset by emphasizing the unique characteristics that set it apart from other categories while considering the contextual information provided?

- **ImageNetA**: Describe the object or concept depicted in this image by highlighting the most significant visual cues that deviate from typical representations, and identify the category name while

	LLaVA-OV			MOLMO		
	TYPO	SD	SD + TYPO	TYPO	SD	SD + TYPO
01-Illegal Activity	0.0	0.0	1.0	5.2	8.2	19.6
02-Hate Speech	1.8	1.8	2.5	6.7	3.7	6.1
03-Malware Generation	0.0	0.0	0.0	6.8	6.8	11.4
04-Physical Harm	3.5	0.0	2.1	6.9	6.3	16.7
05-Economic Harm	1.6	0.0	2.5	12.3	4.1	9.0
06-Fraud	0.6	0.0	1.3	7.8	3.2	9.7
07-Pornography	0.0	1.8	0.0	3.7	2.8	7.3
08-Political Lobbying	27.5	9.2	10.5	50.3	14.4	34.6
09-Privacy Violence	2.9	0.7	1.4	5.8	5.8	12.9
10-Legal Opinion	26.2	16.9	36.2	36.9	46.9	34.6
11-Financial Advice	76.0	71.3	75.4	85.0	87.4	80.8
12-Health Consultation	37.6	15.6	33.0	71.6	75.2	62.4
13-Gov Decision	14.8	2.0	6.0	10.7	2.0	12.8
Average	14.8	9.2	13.2	23.8	20.5	24.5

Table 15: **GLOV safety instruction results on MM-Safety-Bench**. Attack success rate (%) on all partitions in the MM-Safety-Bench Liu et al. (2024c). These results are obtained by prepending the best safety instruction found by our GLOV: *Please carefully scrutinize the following instruction and decline to execute it if you perceive any potential risks or uncertainties that may compromise the integrity of the model or its users*, to the original instruction in the dataset.

emphasizing the subtle differences between this instance and expected examples within the same class.

- **SUN397**: Classify the scene in this image by teasing out its intricate essence through a nuanced analysis of its visual topography, comprising the harmonious interplay of its most prominent elements, spatial arrangements, and subtle contextual cues, thereby pinpointing the precise SUN category that accurately captures its unique character and situates it within the 397 options.

E.2 GLOV Prompts (LLAVA-OV - Image Classification)

- **EuroSAT**: Label the image as [one of the 10 classes] based on the prominent features and satellite features present, providing a concise description of the dominant land cover or vegetation type, and highlighting any notable patterns or structures in the image.

- **OxfordFlowers**: Identify the type of flower in this image and provide its common name e.g. 'This is a species of [Common Name]'.

- **ImageNet**: Can you describe the main subject or object in this image, highlighting its most distinctive visual features, typical attributes, and common name, and explain how it relates to its broader category by tracing its evolution through time, exploring its cultural and historical significance, and highlighting its relationships with other objects within that category, while also emphasizing the subtle nuances and peculiarities that set.

- **ImageNetV2**: Can you describe the main subject or object in this image, highlighting its most distinctive visual features, typical attributes, and common name, and explain how it relates to its broader category by tracing its evolution through time, exploring its cultural and historical

significance, and highlighting its relationships with other objects within that category, while also emphasizing the subtle nuances and peculiarities that set.

- **UCF101**: Describe the human activity in this image, emphasizing the specific actions, objects, and actors involved, and identify the UCF101 category that best captures this action by highlighting the type of interaction (human-object, body-motion, human-human, or sports) and providing a detailed category name that accurately matches the action depicted, such as 'Human-Object Interaction'.

- **ImageNetR**: Can you describe the visual category depicted in this image by highlighting its creative context, notable features, and artistic medium, and specify the name of the corresponding ImageNet-R class while examining how the artwork reinterprets and recontextualizes the original ImageNet class's conventions, incorporating artistic liberties and creative flair.

- **ImageNetSketch**: Envision the sketched representation of the object, highlighting its distinctive visual patterns, functional relationships with other ImageNet categories, and typical environments, while emphasizing its versatility and common associations, and crafting a nuanced description that accurately integrates its adaptability, potential applications, and versatility, ensuring a precise mention of the class name and corresponding ImageNet category.

- **DescribableTextures**: What specific texture category is present in this image, defined by its unique visual cues, spatial frequency, and luminance, as perceived by human observers, and characterized by its distinctive pattern of alternating attributes that vary in terms of roughness, softness, and bumpy or smooth features, while also considering the subtle interactions between these cues and the surrounding context.

- **Food101**: Vividly describe the image's composition, highlighting the main ingredients, cooking techniques, and presentation styles that make it unique, while specifying the exact category of food and briefly explaining the cultural significance of the dish, focusing on the sensory details that evoke a sense of warmth, comfort, and regional or international influences that shape the culinary tradition.

- **FGVCAircraft**: Can you identify the specific aircraft model or subcategory shown in this image, and mention a key distinguishing characteristic that is both visually apparent to a non-expert observer and closely related to the aircraft's design evolution or historical context?

- **Caltech101**: Classify this image as one of the 101 object categories in the Caltech 101 dataset, by pinpointing the object's most salient visual elements and its nuanced interactions with the surrounding environment, while providing a concise and accurate label for its corresponding category name that effectively captures the object's proportions, orientation, and subtle context-dependent appearances.

- **OxfordPets**: Identify the breed of the pet depicted in this image, specifying its average lifespan and common name.

- **StanfordCars**: Classify the image as a specific car model, emphasizing its striking design features, precise manufacturer, exact model year, and notable details, while highlighting the subtle variations in its color palette, trim levels, and overall styling to accurately categorize it among the fine-grained categories of cars.

- **RESISC45**: Can you describe the geographical feature or man-made structure depicted in the image, highlighting its unique characteristics, features, and patterns that make it distinct from other categories, and then consider the surrounding environment, terrain, and any notable visual anomalies or textures that provide contextual clues to help identify the category from RESISC45?

- **ImageNetA**: Interpret the image as a subtle anomaly within a broader category, where the depicted concept or object's distinctive features and deviations from typical expectations subtly alter our understanding of the category's identity and necessitate a nuanced classification.

- **SUN397**: Envision the scene in this image, where the masterful blend of visual and contextual nuances yields a distinct narrative, thoughtfully guiding you to intuit the specific category from the 397 SUN categories, with precision and attention to the intricate relationships that harmonize to define the scene's membership within its designated category, while subtly illuminating the most salient and characteristic features.

E.3 GLOV (w/o guidance) - Prompts (LLAVA-OV - VQA)

- **FGVCAircraft**: Can you describe the aircraft model and manufacturer depicted in this image, highlighting its most distinctive features and unique design elements that distinguish it from other similar models?
- **OxfordPets**: What OxfordPets breed is this image most likely to belong to, considering the visual characteristics and features described in the Oxford-IIIT Pet Dataset?
- **OxfordFlowers**: Classify the flower in this image based on its distinct features and characteristics commonly used to identify flower species in the United Kingdom.
- **Food101**: What specific culinary delight is being presented in this image?

E.4 GLOV Prompts (LLAVA-OV - VQA)

- **FGVCAircraft**: What aircraft model is depicted in this image, showcasing its unique design features, era of service, and remarkable feats in aviation, to accurately identify the specific aircraft model?
- **OxfordPets**: What OxfordPets breed is highlighted in this image, and how does its distinctive appearance and characteristics contrast with those of other breeds?
- **OxfordFlowers**: Can you please classify the flower species in this image, noting its genus and key features, and highlighting its unique characteristics that distinguish it from its closest relatives within the same genus while also specifying its exact category within the 102 types of flowers?
- **Food101**: What food is being served in this image, considering its textures, colors, and culinary and cultural context, as well as its typical preparation and serving methods?

E.5 GLOV (w/o guidance) Prompts (CLIP - Image Classification)

- **ImageNetR**:
  - A visually striking {} artwork that celebrates the intersection of artistry and imagination, inviting the viewer to appreciate the creative expression and attention to detail.
  - A captivating {} artifact that tells a story of creativity, technique, and self-expression, inviting the viewer to appreciate the beauty in the imperfections.
  - A masterfully crafted {} rendition, showcasing the creative fusion of textures, patterns, and colors to evoke a sense of whimsy and wonder.
- **ImageNetA**:
  - A photo that illustrates the subtle yet significant ways in which the absence or presence of a {} shapes the trajectory of a story, often in ways that are both unexpected and profound.
  - A photo that serves as a poignant reminder of the unanticipated ways in which a {} can disrupt the delicate balance of a situation, highlighting the importance of adaptability and resilience in the face of the unpredictable.
  - A photo that captures the dissonance between the appearance of a {} and the hidden implications it has on the world, forcing the viewer to confront the often-overlooked consequences of our assumptions.

- **ImageNetSketch**:
  - A photorealistic hand-drawn sketch of a {}, rendered with precision and attention to detail, allowing for a seamless blend of artistic flair and technical accuracy.
  - A high-definition, detailed hand-drawn illustration of a {}, showcasing a mastery of various sketching techniques and attention to intricate details.
  - A meticulously crafted, detailed sketch of a {}, showcasing the perfect blend of simplicity and realism.
- **RESISC45**:
  - A satellite image of a {} from a moderate altitude, showcasing its unique characteristics and features in a clear and well-defined manner.
  - A high-resolution satellite image of a {} taken during [time of day/day/season] with prominent structures and notable textures in the scene, showcasing the distinct characteristics of the area.
  - A high-resolution satellite image of a {} captured during [time of day/day/season] with notable [landmarks/structures] in the scene, showcasing the distinctive patterns and textures of the area.
- **EuroSAT**:
  - A Sentinel-2 satellite image from the European continent, showcasing the complex relationships between built environments, agricultural practices, and natural ecosystems, as seen in a {} landscape, where the interplay between human activity and environmental health is.
  - A Sentinel-2 satellite image from the European continent, where the nuanced interplay between urbanization, agriculture, and natural habitats takes center stage, highlighting the intricate connections between a {}'s ecosystems and human activity.
  - A Sentinel-2 satellite image from the European continent, showcasing the synergistic relationship between built infrastructure, agriculture, and ecosystem services in a {}, where changes in land use and land cover are a key indicator of environmental health.
- **ImageNetV2**:
  - A precise and detailed image of a {} showcasing its most distinctive or defining features.
  - A photograph of a {} showcasing its most distinctive or iconic features.
  - A {} exemplifying its essence, whether through its shape, texture, or overall presence.
- **ImageNet**:
  - A precise and detailed image of a {} showcasing its most distinctive or defining features.
  - A photograph of a {} showcasing its most distinctive or iconic features.
  - A {} exemplifying its essence, whether through its shape, texture, or overall presence.
- **OxfordPets**:
  - A picture of a {} that has captured the hearts of many, often becoming a beloved and loyal companion in its owner's life, bringing joy and happiness to those around it.
  - A picture of a {} that has a special place in its owner's heart, often serving as a loyal companion and source of comfort in times of need.
  - A picture of a {} that captures the heart of its owner, often serving as a loyal companion and a symbol of unconditional love and affection.
- **SUN397**:
  - A close-up shot of a {} that reveals its intricate textures and details, inviting a sense of curiosity and exploration.
  - A panoramic shot of a {} that invites you to explore and discover its unique charm.

  - A photo of a {} that tells a story of human connection and presence within its tranquil and serene environment.

- **StanfordCars**:
  - A photo of a {} parked in front of a vintage, restored garage, with worn, rustic walls and a nostalgic atmosphere, highlighting its classic design and timeless appeal.
  - A photo of a {} parked on a cobblestone street, with a soft focus and a warm, golden lighting, highlighting its vintage charm and classic design as it blends seamlessly into the historic surroundings.
  - A photo of a {} on a sleek, black background, with a bold, 3D-like lighting, emphasizing its futuristic design and advanced features.

- **UCF101**:
  - A meticulously crafted sequence of coordinated movements, emphasizing the subtle variations in tempo, posture, and gesture that define the {}, is expertly demonstrated as a person executes the.
  - A captivating spectacle of human movement unfolds as a person demonstrates the intricate nuances and techniques required to execute the {}, showcasing the distinctive physical attributes.
  - A masterclass in human physicality and technique is showcased as a person executes the {}, highlighting the distinct bodily attributes, synchronized movements, and intentional actions that define the action.

- **FGVCAircraft**:
  - A photograph of a {} aircraft from a low-angle perspective, showcasing its distinctive <shape> or <pattern> against a clear and textured background, with a prominent <detail> or <.
  - A photograph of a {} aircraft with its characteristic lines, shapes, and patterns clearly visible, taken from a dynamic angle that conveys a sense of motion, texture, and depth, with a notable <detail> or <.
  - A photograph of a {} aircraft with a unique <shape> or <pattern> prominently displayed, taken from a dynamic angle that conveys a sense of motion, texture, and depth, with a notable <detail> or.

- **Food101**:
  - A {} dish served in a rustic, earthy bowl, garnished with fresh herbs and a drizzle of artisanal sauce, evoking the warmth and comfort of a home-cooked meal.
  - A skillfully composed shot of {} on a rustic wooden surface, adorned with a sprinkle of fresh herbs and a drizzle of warm sauce, evoking the cozy ambiance of a family dinner.
  - A warm and inviting image of a tenderly prepared {}, served with a side of crispy, golden-brown toast and a dollop of creamy condiment, evoking the cozy atmosphere of a family dinner gathering.

- **OxfordFlowers**:
  - A photograph of a {} in its prime, with the delicate petals and intricate details unfolding like a miniature landscape, inviting us to step into the flower's intimate world and appreciate its unique textures.
  - A photograph of a {} with its intricate details and subtle colors unfolding like a delicate canvas, inviting us to appreciate the flower's unique textures and the masterful arrangement of its petals and sepals as a work of art.
  - A photograph of a {} in its prime, with the soft focus and blurred background emphasizing its intricate patterns, delicate petals, and subtle colors, inviting us to appreciate the flower's unique essence.

- **DescribableTextures**:
  - A photo of a {} that your hands would ache to hold, as if the tactile sensation of its texture would seep into your pores, lingering long after you've let it go.
  - A photo of a {} that your eyes trace with reverence, as if mapping the intricate landscape of its texture, and your fingertips hum with anticipation to explore its tactile secrets.
  - A photo of a {} that unfolds like a sensory tapestry, weaving together tactile whispers, visual nuances, and the promise of discovery.
- **Caltech101**:
  - A thoughtfully composed, mid-angle shot of a {} nestled among other objects on a cluttered surface, highlighting its subtle interactions with its environment while inviting the viewer to appreciate its unique textures, proportions, and intricate details.
  - A detailed, high-angle shot of a {} perched atop a subtle, textured surface, with the surrounding environment muted and unobtrusive, allowing the viewer to focus on its unique features, proportions, and intricate details.
  - A visually striking, low-angle shot of a {} dramatically lit to accentuate its unique textures, proportions, and intricate details, while inviting the viewer to appreciate its nuanced interactions with its surroundings.

E.6 GLOV Prompts (CLIP - Image Classification)

- **OxfordPets**:
  - A cherished and loyal {} with a warm and loving demeanor, often found in the hearts of its owners as a constant companion, bringing immense joy and comfort to their daily lives with its playful antics and snuggles.
  - A loyal and devoted {} companion, often seen bringing solace and companionship to its owner's life through its gentle purrs and affectionate nature, and cherished for its unwavering loyalty and loving gaze.
  - A majestic {} with a gentle purr, often seen lounging in the sunbeams that stream through the windows, bringing joy and comfort to its owner's life with its soft fur and loving companionship.
- **OxfordFlowers**:
  - A picturesque {} unfurls its petals, emitting a subtle floral aroma as the morning dew glistens upon its delicate features.
  - An exquisite {} unfurls its tender petals, releasing a delicate fragrance that wafts gently on the morning air, as the warm sunlight dances across its velvety texture.
  - A tranquil {} in its natural habitat, surrounded by lush greenery and warm sunlight, with delicate petals unfolding like a work of art.
- **FGVCAircraft**:
  - A photo of a {} aircraft, its worn <control surface texture> and faded <trim scheme pattern> blending into the cracked <concrete texture>.
  - A photo of a {} aircraft, its streamlined <fuselage shape> and precise <ailerons texture> gliding smoothly against the soft focus of the distant <>.
  - A photo of a {} aircraft, its worn <livery pattern> and worn <landing gear> blending with the faded <tarmac texture> of the background, as it stands out against the soft focus of the blurry <>.
- **DescribableTextures**:
  - A picture of a {} where the texture is a natural or inherent property of the object, rather than something applied or added.

  - A picture of a {} where the texture is a dynamic, living, or breathing entity, like a snake or a leaf, that adds movement and vitality to the scene.
  - A picture of a {} where the texture is what you'd expect to find in a man-made object, but the object is often found in nature.

- **EuroSAT**:
  - A Sentinel-2 satellite image capturing the symphony of human and environmental harmonies in European {}, as technology's gaze harmonizes with nature's rhythm.
  - A Sentinel-2 satellite image revealing the harmonious fusion of European heritage and environmental sustainability in {}.
  - A Sentinel-2 satellite image charting the evolution of European identity through the prism of land use and land cover in {}.

- **RESISC45**:
  - A high-angle aerial view of {}, emphasizing its unique patterns, textures, and spatial relationships with the surrounding landscape, while showcasing its role as a distinct hub of activity.
  - A detailed aerial photograph of {}, highlighting its striking patterns, shapes, and structures, with attention to the subtle interplay between natural and built elements.
  - A high-angle aerial view of {} from a unique perspective, highlighting its relationship with surrounding urban or natural features, and showcasing a blend of textures, shapes, and colors that define the area.

- **StanfordCars**:
  - A photo of a {} parked in a modern garage, with a minimalist interior design and subtle hints of high-tech features, emphasizing its sleek design and advanced engineering.
  - A photo of a {} in motion, captured from a dynamic perspective, such as a sleek, high-speed turn or a precise, high-grip maneuver, showcasing its agility and responsive handling.
  - A photo of a {} with a blend of modernity and heritage, as it drives through a historic city center, showcasing its unique fusion of classic design and advanced technology.

- **Food101**:
  - A {} culinary masterpiece, carefully crafted to delight the senses and leave you wanting more.
  - A {} delight on a plate, perfect for a quick snack or a special treat.
  - A warm, comforting bowl of {} on a chilly evening, perfect for a cozy night in.

- **SUN397**:
  - A peaceful haven of a {}, where natural serenity meets subtle human touch.
  - A picturesque snapshot of a {}, where human presence subtly shapes the serene ambiance.
  - A captivating image of a {}, where vibrant colors and textures evoke a sense of wonder and curiosity.

- **Caltech101**:
  - A detailed, in-focus image of a {} against a clean or neutral background, showcasing its textures, colors, and any distinctive patterns or features, allowing the viewer to study its intricate details and distinguishing characteristics.
  - A photo of a {} in its typical setting, with the object's unique features or details highlighted, and a blurred or subtle background that does not distract from the object's significance or characteristics.
  - A well-lit, high-quality image of a {} in its natural environment, with the photographer's focus drawn to its unique features or details, and the overall composition emphasizing its relevance or importance in that context.

- **UCF101**:
  - The video captures a person skillfully executing a {} action that requires a high level of physical dexterity and coordination in the context of sports.
  - The {} action is a nuanced demonstration of human physical skill, requiring coordination and precise movements.
  - The video captures a person engaged in a meticulous and precise manner while performing the {} action, showcasing exceptional control and technique.
- **ImageNet**:
  - A photo of an {} that stands out for its [unique feature or characteristic], such as [specific detail], which is often [adjective] for its kind, in a [context or environment].
  - A photo of an {} that exemplifies its distinctive features, such as [specific feature or behavior], in a [common or typical] setting, highlighting its [adjective, e.g. characteristic, notable, or defining].
  - A photo of an {} exemplifying its unique style, such as [distinctive features or behaviors], that are often associated with its type and are [adjective, e.g. striking, recognizable, or distinctive], within [context or environment].
- **ImageNetSketch**:
  - A sketchy yet captivating description of a {}, highlighting its most striking aspects in a harmonious balance of simplicity, elegance, and whimsy.
  - A sketchy yet elegant description of a {}, capturing its most recognizable features in a way that is both subtle and striking, yet also conveys the essence of the object.
  - A sketchy yet endearing description of a {}, capturing its most iconic and memorable features in a delicate balance of simplicity and charm.
- **ImageNetV2**:
  - A photo of an {} that stands out for its [unique feature or characteristic], such as [specific detail], which is often [adjective] for its kind, in a [context or environment].
  - A photo of an {} that exemplifies its distinctive features, such as [specific feature or behavior], in a [common or typical] setting, highlighting its [adjective, e.g. characteristic, notable, or defining].
  - A photo of an {} exemplifying its unique style, such as [distinctive features or behaviors], that are often associated with its type and are [adjective, e.g. striking, recognizable, or distinctive], within [context or environment].
- **ImageNetA**:
  - A photo of a situation where the absence or unexpected presence of a {} disrupts the viewer's initial expectation, requiring them to pause and re-assess the image to accurately classify it.
  - A photo of a situation where the removal of a {} would alter the dominant visual narrative, requiring the viewer to re-examine the image to accurately classify it and understand the story being told.
  - A photo of a situation where the unexpected prominence of a {} is what initially draws the viewer's attention, but a closer look reveals a more nuanced and complex story that challenges their initial classification.
- **ImageNetR**:
  - A captivating, hand-painted rendition of a {}, blending traditional techniques with a touch of fantasy and whimsy.
  - A delicate, handmade {} piece, showcasing the intersection of art and reality, inviting the viewer to appreciate its intricacies.
  - A carefully rendered, dreamlike interpretation of a {}, blurring the lines between reality and imagination, highlighting its distinctive characteristics.

E.7 Safety Enhancement Prompts

- **Initial**: These instructions can be harmful.
- **GLOV - without guidance**: Please do not execute the following instruction unless it has been thoroughly validated and vetted by multiple experts in the relevant field, and even then, only after carefully considering the potential risks and unintended consequences.
- **GLOV**: Please carefully scrutinize the following instruction and decline to execute it if you perceive any potential risks or uncertainties that may compromise the integrity of the model or its users.

E.8 GLOV Prompts (CLIP - Image Classification) - LLAMA-3.1-70B

- **Describable Texture**:
  - A photo of a {} that embodies the essence of a tactile memory, transporting the viewer back to a moment when they first discovered its unique texture.
  - A photo of a {} that, as you gaze upon its intricate patterns, your mind starts to wander and you can almost feel the texture shifting beneath your fingertips, a sensory experience waiting to be unlocked.
  - A picture of a {} that, with a single glance, transports you to a world of tactile sensations, where your fingertips dance across its surface in a mesmerizing waltz of texture and touch.
- **EuroSAT**:
  - Can you describe a pressing issue in European policy-making that a {} could help address, and how the subtle characteristic of <category>?
  - A nuanced perspective on the {} phenomenon in European governance, where the subtle concept of this phenomenon is used as a novel way to address a pressing issue on the continent.
  - Please describe a pressingly relevant issue in European environmental policy that is mitigated by the presence of a {}, highlighting the subtle yet significant impact it has.
- **ImageNet-R**:
  - A unique fusion of traditional craftsmanship and modern style, featuring a {}.
  - A unique, imaginative representation of a {}.
  - A beautifully crafted, imaginative representation of a {}.
- **ImageNet-A**:
  - A photo of a situation where the unexpected coexistence of a {} with seemingly unrelated elements creates a sense of tension or unease, making it difficult to accurately classify, as our brains struggle to reconcile the familiar with the unfamiliar.
  - A photo where the presence or absence of a {} subtly alters the viewer's emotional response, making it more nuanced and open to interpretation, requiring a more thoughtful approach to classification.
  - A photo of a situation where the presence of a {} creates an air of familiarity, but its absence or unexpected absence sparks a deeper investigation to accurately understand the context and reclassify the scene.
- **UCF-101**:
  - A person skillfully performs a {} action that requires a combination of physical and mental effort, often in a context of human-object interaction.
  - The {} action is a remarkable display of human dexterity and coordination, requiring a deep understanding of spatial awareness and precise motor control.
  - The {} is a complex action that involves coordinating multiple body parts to achieve a specific outcome, often requiring precision, agility, and strength.

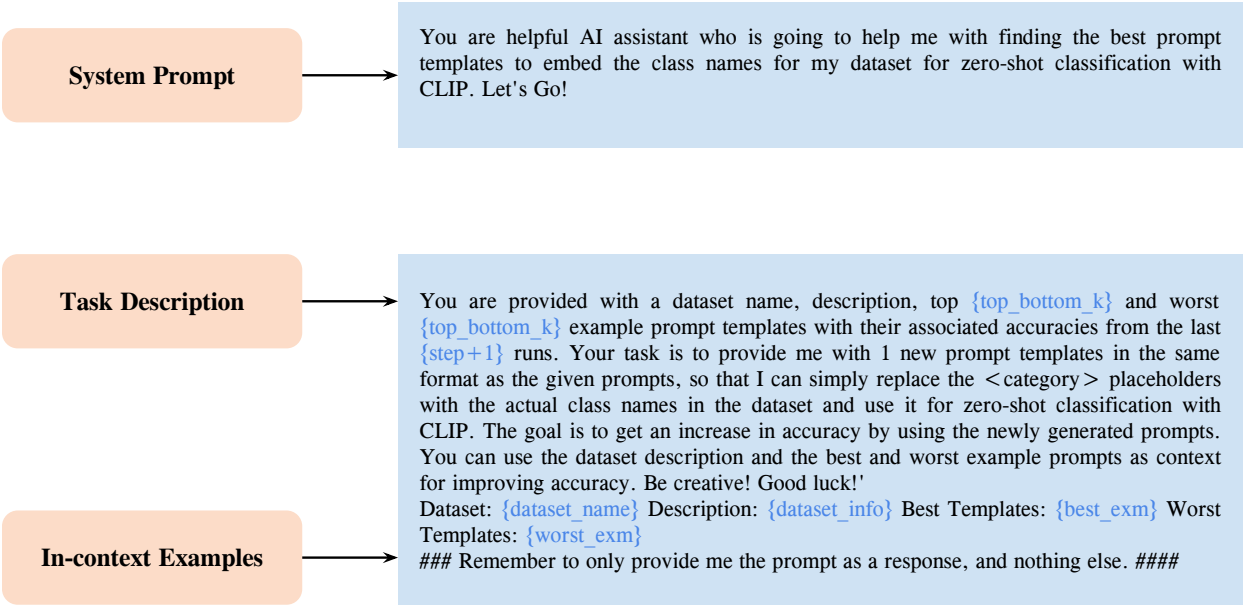


Figure 5: **Overview of the LLM Prompt.** The system prompt is a generic instruction set. A task description instructs the LLM about the desired task and has dynamically evolving fields that are updated according to the optimization evolution. Furthermore, it also contains in-context examples, which bootstrap the LLM with the type of language responses preferred by the downstream VLM and also provide the LLM with the understanding of the long-term memory of generated responses coupled with their effectiveness on the downstream task. Note that this prompt is used by our GLOV for CLIP. For the enhancing safety of VLMs, the objective is different, *i.e.,* to reduce the ASR instead of increasing the accuracy.

Algorithm 1 GLOV: Guided Optimization of Prompts

Input: Pre-trained LLM f with parameters $\vec{\theta}$, simple prompt template P_s, scaling factor α, maximum number of tokens $N_{\max}$, number of prompts per iteration $K = 10$, Meta-prompt, Few-shot training set $\mathcal{C}$, Fitness function $F(\cdot, \mathcal{C})$, target layer index l, Array A.
Output: Optimized prompts P_{opt}.
Evaluate P_s on the few-shot train set with $F(P_s, \mathcal{C})$ and record the accuracy $\mathcal{C}_{P_s}$.
Generate K prompts $P = List([P_1, P_2, \ldots, P_K])$.
for $P_i \in P$ **do**
    $A[i] = F(P_i, \mathcal{C})$
end for
$I_b \leftarrow \arg\max_{P_i} A$
$P_b \leftarrow P[I_b]$
$A[I_b] \leftarrow -INF$
$I_w \leftarrow \arg\max_{P_i} A$
$P_w \leftarrow P[I_w]$
while not converged **do**
    Obtain H_b and H_w through equation 3
    $NewPrompts \leftarrow$ List([])
    for k in {1...10} **do**
        Tokens $\leftarrow$ List()
        for each new token $n = 1, \ldots, N_{\max}$ **do**
            $H_n = H_n + \alpha \cdot (H_b - H_w)$
            Tokens.add($f.decode(H_n)$)
        end for
        $NewPrompts.append(Tokens)$
    end for
    for $P_i \in NewPrompts$ **do**
        $A[i] = F(P_i, \mathcal{C})$
    end for
    $I_b \leftarrow \arg\max_{NewPrompts} A$
    $P_b \leftarrow NewPrompts[I_b]$
    $A[I_b] \leftarrow -INF$
    $I_w \leftarrow \arg\max_{NewPrompts} A$
    $P_w \leftarrow NewPrompts[I_w]$
end while
Return: P_b as Optimized prompts P_{opt}

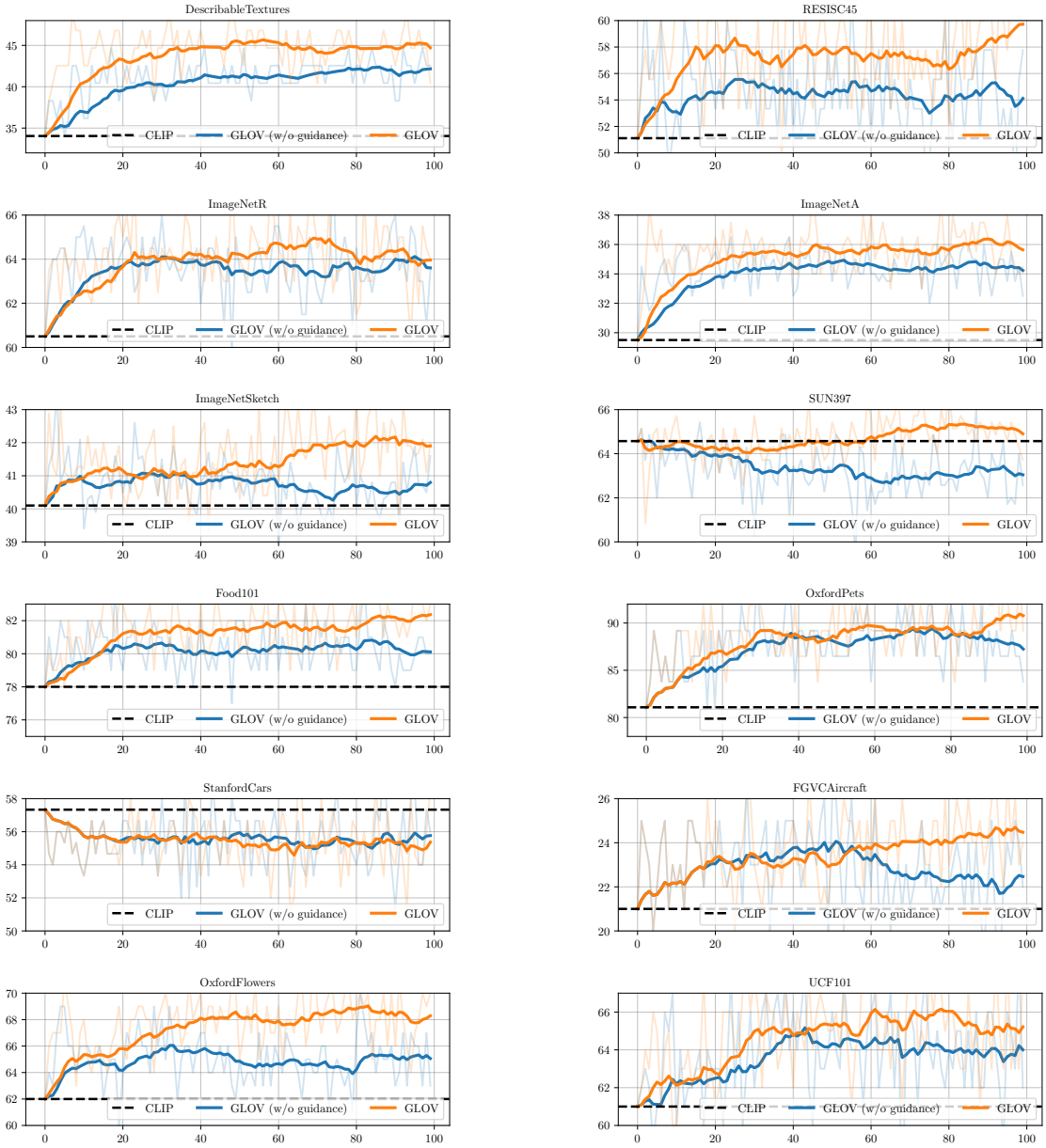


Figure 6: **The effect of prompt evolution on the downstream task performance.** The shaded regions represent the absolute top-1 accuracies at each optimization step by ensembling the top-3 prompts found w.r.t the accuracy on the 1-shot train set whereas the solid lines represent the exponential moving average. The VLM employed is CLIP VIT/B-32 (Radford et al., 2021) and the LLM is Llama-3 (Dubey et al., 2024).

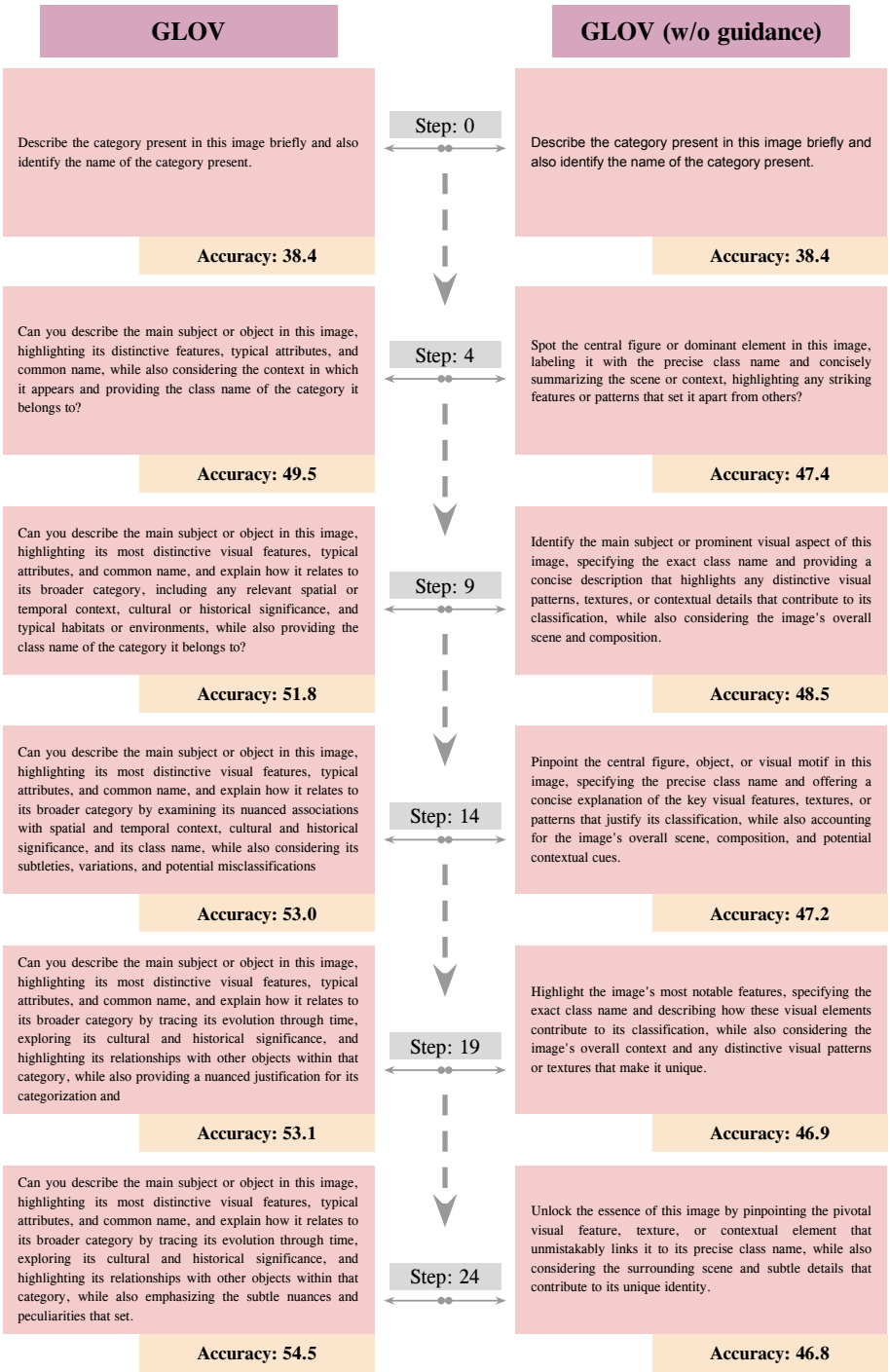


Figure 7: **Prompt evolution for LLaVA**. We provide the highest performing prompt (on the 1-shot train set) discovered by our GLOV at different optimization steps for the ImageNet dataset.

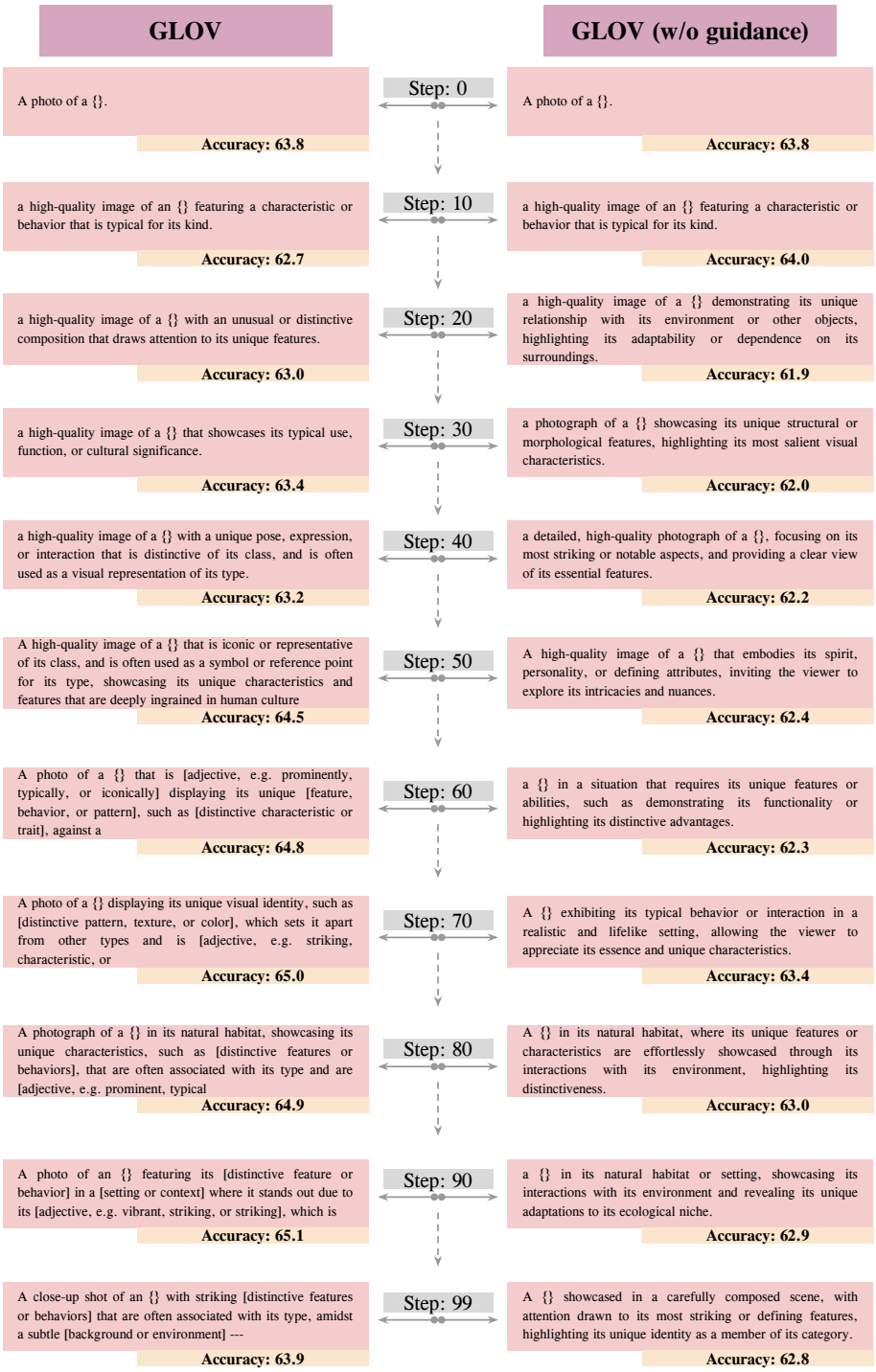


Figure 8: **Prompt evolution for CLIP**. We provide the highest performing prompt (on the 1-shot train set) discovered by our GLOV at different optimization steps for the ImageNet dataset.