Multimodal Task Vectors Enable Many-Shot Multimodal In-Context Learning

Supplementary Material

Here, we provide additional information about our experimental results, qualitative examples, implementation details, and datasets. Specifically, Section A provides more experiment results, Section B provides additional method details, Section C provides additional implementation details, and Section D provides qualitative visualizations to illustrate our approach.

A Additional Experiment Results

We present several additional experiments that further demonstrate the benefits of our MTV approach.

A.1 Additional Experiments

Here we provide additional ablations that further illustrate different characteristics of MTV.

Attention head generalization on object classification tasks Table 1a. We also test generalization for object classification tasks identical to the formulation described in Section **??**. For clarity, MTV shows another kind of generalization when it is leveraged alongside additional explicit ICL samples. This capability is described in Section **??**. To summarize our experiment, we calculate MTV using the Flowers dataset using 1-shot ICL example for 100 iterations for both the mean activations μ_j^{MTV} and the attention head locations λ_j^{MTV}. Then, we apply MTV to the CUB task *using the same set of attention head locations from Flowers*. We just calculate the mean activations for the CUB dataset using a 1-shot for 100 iterations (halving our data requirement for this specific scenario). Once again, we find that the heads of MTV can indeed generalize between similar classes.

Table 1: **Generalization & Direct ICL Comparison** (Left) MTV-Flowers evaluated on OK-VQA. (Right) Direct comparison of MTV extracted from 4-shots, 1-iteration (MTV_4shot_1it) compared to 4-shot ICL

(a) Attention Head Generalization

Model	Flowers	CUB
ViLA-1.5-8B		
+ 1-shot-ICL	87.4	88.4
+ **MTV-Flowers**+1-shot-ICL	89.3	**89.9**

(b) Comparison to Other Methods

Model	VizWiz	OK-VQA
ViLA-1.5-8B	28.0	32.8
+ 4-shot-ICL	39.3	35.6
+ **MTV**_4shot_1it	57.4	40.0

MTV one-to-one comparison with ICL Table 1b. Although not directly comparable, we consider an extreme case of MTV where we encode only 4-shots of ICL examples for 1 iteration. This matches the exact setting used in standard 4-shot ICL. Interestingly, MTV applied to both VizWiz and OK-VQA exceeds performance on the 4-shot-ICL case and even MTV formulated on 4-shots per 100 iterations for calculating the mean activations. This result suggests that there may be scope for MTV to be effective in both high and low-data regimens. More research needs to be done to explore this idea.

Effect of permutation order of examples. We consider applying five random seeds to both 4-shot-ICL and MTV extracted on 4-shots per 100 iterations on VizWiz. We find the 4-shot-ICL average and standard deviation to be 41.3 % (± .8%) and the MTV average and standard deviation to be 45.2 % (± .7 %). This suggests that MTV is stable across different permutations of the given ICL examples.

Scaling on Flowers Dataset. We provide additional results on the scaling property of MTV on the Flowers dataset. We again note that the examples are *2-way*, one-shot examples with 2 examples (one positive and one negative) for each sample. As in the main paper, we fix 1 shot per iteration to calculate the mean activations, scaling up to 500 total examples used. Our results show that there is a saturation of MTV at 100 examples (i.e., 1 example per 100 iterations). While this still indicates some scaling as the result is an improvement over 20 examples, the results show that the task vector

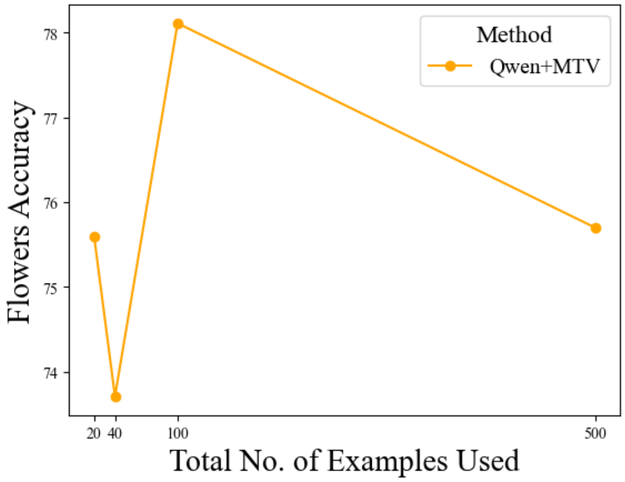

Figure 1: **Efficiency.** We show that for Flowers, MTV does scale to but only up to 100 examples in our experiments.

can reach its best accuracy with fewer shots depending on the complexity of the task. Future work to probe more deeply into the scaling nature of MTV across different tasks would be valuable.

B Additional Method Details

Here we provide some additional method details about MTV, Visual Task Vectors (VTV) [1], and Function Vectors [3] (FV).

B.1 MTV-EXTRACT

We describe the particulars of our MTV-EXTRACT algorithm for finding the set of attention head locations that best align with the downstream task as follows (Q_s and R_s are formatted identically to the downstream task):

Algorithm 1 MTV-EXTRACT for finding task vector locations

```
Require: $F$ (LMM), $S$ (examples), $\mu_j$ (mean activations), $Q_s, R_s$ (queries and responses)
Ensure: $\lambda_j^{\text{MTV}}$ (optimized attention head locations)
Initialize $\theta$ randomly
for $s \leftarrow 1$ to $S$ do
    for $i \leftarrow 1$ to 32 do                                  ▷ Sampling heads 32 times
        Sample $\lambda_i \sim \text{Bernoulli}(\sigma(\theta))$
        Replace activations for $\lambda_i$ in $F$ with $\mu_{l,j}$
        Compute output logits $O_s \leftarrow F(Q_s)$              ▷ Pass $Q_s$ to LMM $F$
        $L_i \leftarrow$ Negative Cross-Entropy($O_s$, $R_s$)
    end for
    $\theta \leftarrow \text{Adam}(\theta, \nabla_\theta \frac{1}{32} \sum_{i=1}^{32} L_i)$    ▷ Update rule
end for
Sample final $\lambda_j^{\text{MTV}} \sim \text{Bernoulli}(\sigma(\theta))$    ▷ Final set of head locations
return $\lambda_j^{\text{MTV}}$
```

We point out a few important factors. It is important to note that none of the parameters of F are being finetuned through any gradient update. We take the negative cross-entropy (negative as MTV_EXTRACT draws inspiration from REINFORCE [5], which is a policy optimization algorithm) between the output logits O_s and the first token of the target response R_s for a simple update scheme. This along with the choice of 32 samples of the Bernoulli distribution are ones we encourage more experimentation with in future work.

B.2 Visual TaskVectors (VTV) Adaptation for Multimodal ICL

Visual Task Vectors (VTV) [1] were originally designed to be applied to large vision-transformer-based models. We make as few changes as possible to apply this method for multimodal tasks. We preserve VTVs distinct factors like a the usage of 1-shot examples for both calculation of the mean activations and attention head locations regardless of the format of the downstream task. Furthermore, we fix the number of iterations for both mean activation and attention head calculation at 10. Finally, we replace the proposed MSE loss with a cross-entropy loss that is more suited for an LMM task.

B.3 Function Vectors (FV)

Because Function Vectors describe text-only task vectors, we follow the implementation of Function Vectors [3] almost exactly as LLMs and LMMs are similar. The only major change made is the use of many-shot multimodal ICL examples for mean activation calculation. We preserve the lack of an optimization method for the layer used to replace the mean activations. Rather than performing a standard grid search over the set of layers, we set the layer number to 20 as recommended for LLaMA and LLaMA-based models by the paper. The only other difference is the encoding of multimodal ICL examples. Again due the the similarity between LMMs and text-only LLMs, these tests can be used as needed as long as the multimodal inputs are properly processed by the LMM.

C Additional Implementation Details

To run all of our experiments, we use 1 NVIDIA RTX 6000 GPU. Importantly, this includes the runtime and efficiency ablations, which were evaluated on the same GPU for consistency. Please refer to the respective model's paper for their specific implementation details of the architecture. Besides the output token generation length, which varies depending on the standard setting for each task, we use the default generation parameters (e.g. temperature and no. of beams in beam search) recommended for each model. In the following sections, we describe some of the finer nuances of our MTV-EXTRACT process as well as our implementations of the Visual Task Vectors (VTV) and Function Vectors (FV) implementations.

C.1 VizWiz

Dataset. The VizWiz dataset is designed to challenge and evaluate the capabilities of Large Multimodal Models (LMMs) in understanding and responding to real-world visual questions. This dataset is comprised of images accompanied by spoken questions, which have been transcribed and paired with answers. Each image in this dataset is sourced from visually impaired individuals seeking assistance, thereby incorporating a wide array of everyday challenges they face. This setup is inherently diverse and often requires high-level visual understanding combined with contextual reasoning, making them a robust benchmark for assessing the practical utility of LMMs in assistive technologies. The format of the dataset samples is an image paired with a text question. The LMM is required to provide a short response limited to 10 tokens or respond with "unanswerable" if the question is not answerable give the image.

For this research paper, we specifically utilize the VizWiz dataset to benchmark the performance of our proposed task vectors in multimodal in-context learning (MM-ICL) on a dataset that challenges visual scene understanding of LMMs. We extract MTV on the training set and evaluate on the evaluation set containing 4,319 validation image/question pairs.

Inference details. We use the standard VQA question-answer response format that is outlined in the QwenVL repository `https://github.com/QwenLM/Qwen-VL`. Put simply, the LMM is presented with an image and a corresponding text question. The response is then expected in a short text format of no more than 10 tokens (set as the "max_tokens" parameter in the LMM). One nuance is the special answer "unanswerable". We handle this by providing MTV and all baselines with the following prompt for every question: "First carefully understand the given examples. Then use the given image and answer the question in the same way as the examples. If the question can not be answered, respond unanswerable. " The official dataset can be downloaded at `https://vizwiz.org/tasks-and-datasets/vqa/`.

C.2 OK-VQA

Dataset. The OK-VQA dataset, differs from traditional VQA datasets in its focus on necessitating knowledge beyond what is presented in the given images. This dataset encompasses over 14,000 questions that are not merely reliant on visual cues but require associative reasoning with external data sources, making it a unique tool for evaluating AI's capability in handling complex, knowledge-driven queries. Thus, we evaluate on this dataset to test whether MTV can be beneficial for this type of reasoning.

We once again extract MTV on the train set and evaluate on the validation set. OK-VQA is formatted as an image with a corresponding text question. However, it is important to note that the text question heavily relies on external knowledge to answer. Examples of questions can be found in Section D.

Inference details. We use the standard VQA question-answer response format that is outlined in the QwenVL repository `https://github.com/QwenLM/Qwen-VL`. Put simply, the LMM is presented with an image and a corresponding text question. The response is then expected in a short text format of no more than 10 tokens (set as the "max_tokens" parameter in the LMM). We do not add any additional prompts or special tokens apart from prompt format or image tokens required by the model being evaluated. The official dataset can be downloaded at `https://okvqa.allenai.org/`.

C.3 Flowers

Dataset. Flowers [2] is an object classification dataset that requires fine-grained classification of 102 different flower species. The Flowers dataset is formulated as a 2-way, 1-shot task where one example is the positive sample and the other is the negative sample. In this way, the data poses a unique challenge for MTV having to store examples with two associated images. Thus, given the 2-way examples and the query image, the LMM is tasked with selecting the correct class from the given two options. Examples can be found in Section D

Implementation Details. We use the official data released by the authors which is available at `https://www.robots.ox.ac.uk/~vgg/data/flowers/`. We provide a Python code snippet below showing the Flowers data format:

```
def format_flower(cur_data):
    pos = cur_data["pos"]
    neg = cur_data["neg"]
    pos_label = cur_data["pos_label"]
    neg_label = cur_data["neg_label"]
    query = cur_data["query"]
    rand_num = random.randint(0,1)
    if rand_num == 0:
        pos_example = f"<img>{pos}</img>What is the type of flower in the image? A.{
            pos_label} B.{neg_label}\nAnswer with the option's letter from the given
             choice directly. Answer: A\n"

        neg_example = f"<img>{neg}</img>What is the type of flower in the image? A.{
            pos_label} B.{neg_label}\nAnswer with the option's letter from the given
             choice directly. Answer: B\n"

        cur_query = f"<img>{query}</img>What is the type of flower in the image? A.{
            pos_label} B.{neg_label}\nAnswer with the option's letter from the given
             choice directly. Answer:"
        query_label = "A"
        return pos_example + neg_example + cur_query, query_label, -1

    else:
        pos_example = f"<img>{pos}</img>What is the type of flower in the image? A.{
            neg_label} B.{pos_label}\nAnswer with the option's letter from the given
             choice directly. Answer: B\n"

        neg_example = f"<img>{neg}</img>What is the type of flower in the image? A.{
            neg_label} B.{pos_label}\nAnswer with the option's letter from the given
             choice directly. Answer: A\n"

        cur_query = f"<img>{query}</img>What is the type of flower in the image? A.{
            neg_label} B.{pos_label}\nAnswer with the option’s letter from the given
             choice directly. Answer:"
        query_label = "B"
        return neg_example + pos_example + cur_query, query_label, -1
```

C.4 CUB

Dataset. CUB [4] or CUB-200-2011 is an object classification dataset that tests the fine-grained classification of 200 classes of birds. Similar to the Flowers dataset, CUB is formulated as a 2-way, 1-shot task where one example is the positive sample and the other is the negative sample. In this way, the data poses a unique challenge for MTV having to store examples with two associated images. Thus, given the 2-way examples and the query image, the LMM is tasked with selecting the correct class from the given two options.

Implementation Details. We use the official data released by the authors which is available at `https://www.vision.caltech.edu/datasets/cub_200_2011/`. We provide a Python code snippet below showing the Flowers data format:

```
def format_cub(cur_data):
    pos = cur_data["pos"]
    neg = cur_data["neg"]
    pos_label = cur_data["pos_label"]
    neg_label = cur_data["neg_label"]
    query = cur_data["query"]
    rand_num = random.randint(0,1)
    if rand_num == 0:
        pos_example = f"<img>{pos}</img>What is the type of bird in the image? A.{
            pos_label} B.{neg_label}\nAnswer with the option’s letter from the given
             choice directly. Answer: A\n"

        neg_example = f"<img>{neg}</img>What is the type of bird in the image? A.{
            pos_label} B.{neg_label}\nAnswer with the option’s letter from the given
             choice directly. Answer: B\n"

        cur_query = f"<img>{query}</img>What is the type of bird in the image? A.{
            pos_label} B.{neg_label}\nAnswer with the option’s letter from the given
             choice directly. Answer:"
        query_label = "A"
        return pos_example + neg_example + cur_query, query_label, -1

    else:
        pos_example = f"<img>{pos}</img>What is the type of bird in the image? A.{
            neg_label} B.{pos_label}\nAnswer with the option’s letter from the given
             choice directly. Answer: B\n"

        neg_example = f"<img>{neg}</img>What is the type of bird in the image? A.{
            neg_label} B.{pos_label}\nAnswer with the option’s letter from the given
             choice directly. Answer: A\n"

        cur_query = f"<img>{query}</img>What is the type of bird in the image? A.{
            neg_label} B.{pos_label}\nAnswer with the option’s letter from the given
             choice directly. Answer:"
        query_label = "B"
        return neg_example + pos_example + cur_query, query_label, -1
```

D Qualitative Visualizations

We present further qualitative success and failure cases of **QwenVL-MTV** in Figure 2 on OK-VQA and Flowers.

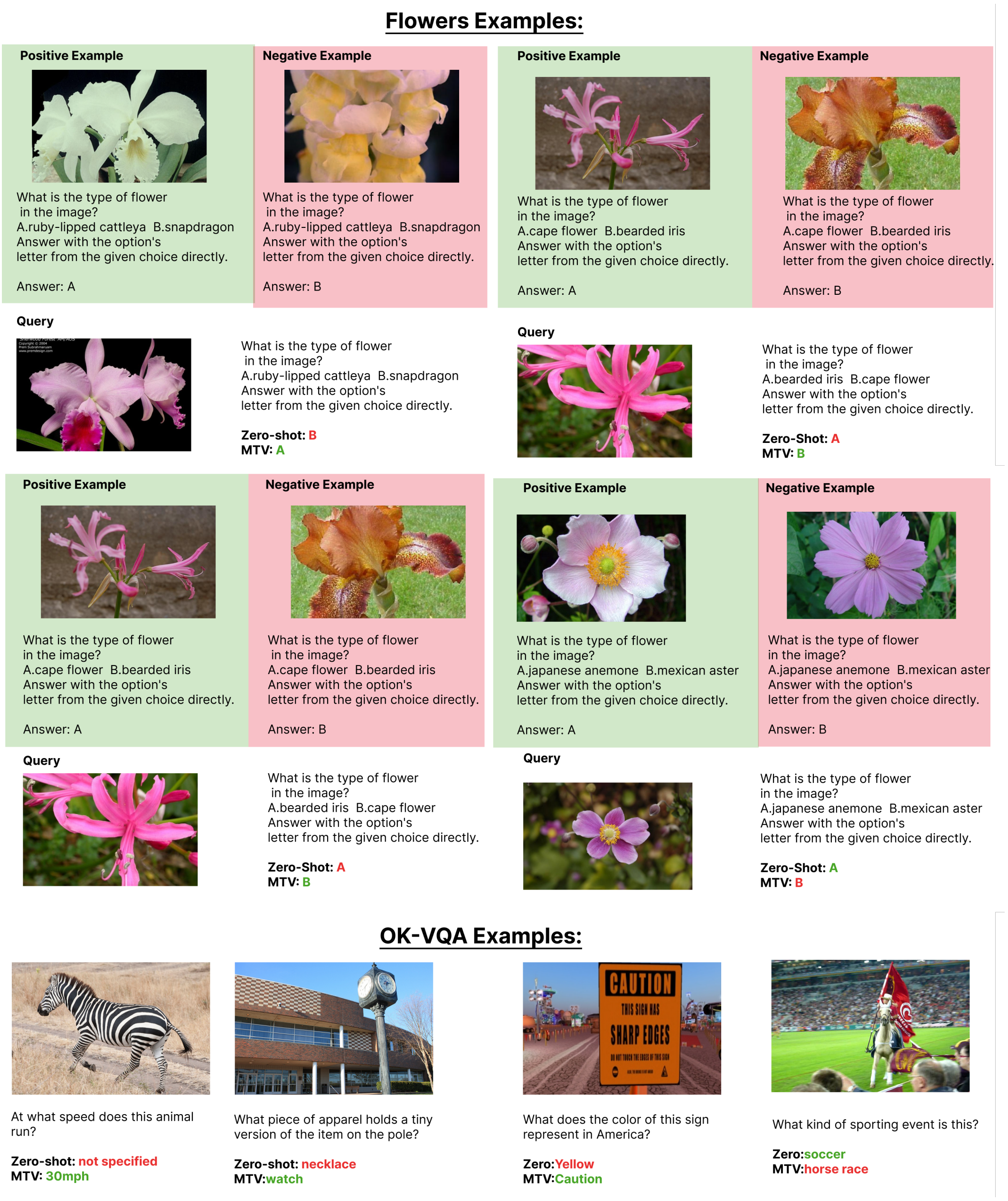
Flowers Examples:
Positive Example
What is the type of flower in the image? A.ruby-lipped cattleya B.snapdragon Answer with the option's letter from the given choice directly.
Answer: A
Negative Example
What is the type of flower in the image? A.ruby-lipped cattleya B.snapdragon Answer with the option's letter from the given choice directly.
Answer: B
Positive Example
What is the type of flower in the image? A.cape flower B.bearded iris Answer with the option's letter from the given choice directly.
Answer: A
Negative Example
What is the type of flower in the image? A.cape flower B.bearded iris Answer with the option's letter from the given choice directly.
Answer: B
Query
What is the type of flower in the image? A.ruby-lipped cattleya B.snapdragon Answer with the option's letter from the given choice directly.
Zero-shot: B
MTV: A
Query
What is the type of flower in the image? A.bearded iris B.cape flower Answer with the option's letter from the given choice directly.
Zero-Shot: A
MTV: B
Positive Example
What is the type of flower in the image? A.cape flower B.bearded iris Answer with the option's letter from the given choice directly.
Answer: A
Negative Example
What is the type of flower in the image? A.cape flower B.bearded iris Answer with the option's letter from the given choice directly.
Answer: B
Positive Example
What is the type of flower in the image? A.japanese anemone B.mexican aster Answer with the option's letter from the given choice directly.
Answer: A
Negative Example
What is the type of flower in the image? A.japanese anemone B.mexican aster Answer with the option's letter from the given choice directly.
Answer: B
Query
What is the type of flower in the image? A.bearded iris B.cape flower Answer with the option's letter from the given choice directly.
Zero-Shot: A
MTV: B
Query
What is the type of flower in the image? A.japanese anemone B.mexican aster Answer with the option's letter from the given choice directly.
Zero-Shot: A
MTV: B
OK-VQA Examples:
At what speed does this animal run?
Zero-shot: not specified
MTV: 30mph
What piece of apparel holds a tiny version of the item on the pole?
Zero-shot: necklace
MTV:watch
What does the color of this sign represent in America?
Zero:Yellow
MTV:Caution
What kind of sporting event is this?
Zero:soccer
MTV:horse race

Figure 2: **Efficiency.** We show that for Flowers, MTV does scale to but only up to 100 examples in our experiments.

E Licenses and Privacy

The license, PII, and consent details of each dataset are in the respective papers. In addition, we wish to emphasize that the datasets we use do not contain any harmful or offensive content, as many other papers in the field also use them. Thus, we do not anticipate a specific negative impact, but, as with any machine learning method, we recommend exercising caution.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, it does.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed about limitations.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Not relevant; it is not a theory paper.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Everything is reproducible.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Yes, code is provided.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, all is included.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Our paper does not require error bars or statistical significance, only accuracy.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we describe required compute for our method

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader impact is discussed in Supp.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: We don't have any safeguards to discuss here.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: All data and code are credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: Not-relevant

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: Not relevant

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: Not relevant

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

References

[1] Alberto Hojel, Yutong Bai, Trevor Darrell, Amir Globerson, and Amir Bar. Finding visual task vectors. 2024.

[2] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008.

[3] Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. *ArXiv*, abs/2310.15213, 2023.

[4] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[5] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 2004.

;