

## A Appendix / supplemental material

### A.1 Different class order

Since the constructed sub-graph is influenced by the order in which classes are encountered, different class orders result in different sub-graphs. To evaluate this effect, we conducted multiple experiments with varying class sequences, and the results are presented in Tab. 5 and Tab. 6. As shown, while our model’s performance fluctuates due to the order variations, it consistently maintains a high level of accuracy, demonstrating the robustness of our method.

Table 5: Comparison results between our method and other methods on Tiny-ImageNet and ImageNet-R under different class orders.

Type	Method	Exemplar	B100-5 tasks		Tiny-ImageNet B100-10 tasks		B100-20 tasks		ImageNet-R B0-10 tasks
			Avg	Last	Avg	Last	Avg	Last	Last
Conventional	EWC [24]	✗	19.01	6.00	15.82	3.79	12.35	4.73	35.00
	LwF [25]	✗	22.31	7.34	17.34	4.73	12.48	4.26	38.50
	iCaRL [40]	✓	45.95	34.60	43.22	33.22	37.85	27.54	-
	EEIL [6]	✓	47.17	35.12	45.03	34.64	40.41	29.72	-
	UCIR [15]	✓	50.30	39.42	48.58	37.29	42.84	30.85	-
	PASS [69]	✗	49.54	41.64	47.19	39.27	42.01	32.93	-
	DyTox [11]	✓	55.58	47.23	52.26	42.79	46.18	36.21	-
Discriminative PT models	Continual-CLIP[49]	✗	70.49	66.43	70.55	66.43	70.51	66.43	72.00
	L2P [56]	✗	83.53	78.32	76.37	65.78	68.04	52.40	72.92
	DualPrompt [55]	✗	85.15	81.01	81.38	73.73	73.45	60.16	68.82
	CODA-Prompt [44]	✗	85.91	81.36	82.80	75.28	77.43	66.32	73.88
	MoE-CLIP [59]	✗	81.12	76.81	80.23	76.35	79.96	75.77	80.87
	RAPF [17]	✗	78.64	74.67	77.42	73.57	76.29	72.65	80.28
	Linear Probe	✗	74.38	65.40	69.73	58.31	60.14	49.72	45.17
Generative PT models	Zero-shot	✗	58.16	53.72	58.10	53.72	58.13	53.72	67.38
	GMM [5]	✗	83.42	76.98	82.49	76.51	81.70	76.03	80.72
	KG-GMM (Ours) order1	✗	85.88	80.93	83.91	77.35	82.98	77.95	84.30
	KG-GMM (Ours) order2	✗	<b>86.35</b>	82.02	84.45	78.47	<b>83.43</b>	<b>78.66</b>	<b>84.62</b>
	KG-GMM (Ours) order3	✗	86.27	<b>82.62</b>	<b>84.75</b>	<b>78.67</b>	83.13	78.36	83.95
	KG-GMM (Ours)	✗	86.17 ± 0.21	81.86 ± 0.70	84.37 ± 0.35	78.16 ± 0.58	83.18 ± 0.19	78.32 ± 0.29	84.29 ± 0.27

Table 6: Comparison results of our method with other conventional baselines and methods on the mini-ImageNet dataset for few-shot class incremental learning under different class orders. The table includes one base task and eight incremental tasks. **PD** is the performance drop between the first and last session. \* indicates our re-implementation based on PILOT [46].

	0	1	2	3	4	5	6	7	8	PD↓	HAcc↑
iCaRL [40]	61.31	46.32	42.94	37.63	30.49	24.00	20.89	18.80	17.21	44.10	32.45
EEIL [6]	61.31	46.58	44.00	37.29	33.14	27.12	24.10	21.57	19.58	41.73	28.43
LUCIR [15]	61.31	47.80	39.31	31.91	25.68	21.35	18.67	17.24	14.17	47.14	35.65
TOPIC [48]	61.31	50.09	45.17	41.16	37.48	35.52	32.19	29.46	24.42	36.89	32.98
CEC [61]	72.00	66.83	62.97	59.43	56.70	53.73	51.19	49.24	47.63	24.37	15.96
F2M [43]	72.05	67.47	63.16	59.70	56.71	53.77	51.11	49.21	47.84	24.21	19.23
MetaFSCIL [8]	72.04	67.94	63.77	60.29	57.58	55.16	52.90	50.79	49.19	22.85	14.35
Entropy-reg [28]	71.84	67.12	63.21	59.77	57.01	53.95	51.55	49.52	48.21	23.63	19.29
L2P* [56]	94.12	87.20	80.99	75.67	70.94	66.76	63.11	59.81	56.83	37.29	0.0
DualPrompt* [55]	93.97	86.85	80.67	75.31	70.61	66.44	62.77	59.58	56.80	37.17	0.1
CODA-Prompt* [44]	<b>95.37</b>	88.86	82.69	77.87	74.47	70.16	66.46	63.73	61.14	34.23	0.0
Zero-shot	58.08	58.95	57.76	57.89	58.19	57.42	56.26	54.82	54.95	<b>3.13</b>	52.47
GMM	89.35	88.40	86.11	85.07	83.61	81.35	78.97	77.34	75.18	14.17	71.45
KG-GMM order 1	91.32	89.87	87.92	87.67	85.24	83.12	80.97	79.43	77.82	13.50	74.59
KG-GMM order 2	91.54	<b>90.14</b>	88.13	<b>87.83</b>	85.42	83.69	<b>81.45</b>	<b>80.23</b>	<b>78.43</b>	13.11	74.87
KG-GMM order 3	90.12	88.40	<b>88.28</b>	86.37	<b>86.23</b>	<b>84.28</b>	80.95	79.23	77.95	12.17	<b>74.96</b>
KG-GMM avg	90.99 ± 0.39	89.50 ± 0.61	88.11 ± 0.02	87.29 ± 0.43	85.63 ± 0.19	83.70 ± 0.22	81.12 ± 0.05	79.63 ± 0.19	78.07 ± 0.07	12.93 ± 0.31	74.81 ± 0.02

### A.2 Experiments on fine-grained and medical datasets

We conducted exploratory experiments on several fine-grained datasets and medical imaging datasets, as detailed below:

#### A.2.1 Datasets and settings

CUB-200 [52] is a fine-grained bird species dataset with detailed part-level attributes. We construct a knowledge graph using each class’s most certain attributes (confidence score 4), such as *has\_bill\_shape::hooked* or *has\_bill\_shape::cone*, resulting in approximately 17 relations per class after our graph construction method. All 200 classes are equally separated into 10 tasks.

HAM-10000 [50] is a medical image dataset of 10,000 pigmented skin lesions of 7 classes for melanoma classification. Since we could not find a suitable knowledge graph for medical data, we instead used DeepSeek-O1 to generate 10 representative attributes for each of the 7 classes (such as *color\_dominance::light brown*, *background\_skin::No pigmentation*), serving as their semantic descriptions or attribute sets. We follow [2] to separate 7 classes in 3 tasks, containing [2,2,3] classes each.

FVGC-Aircraft [31] is a dataset comprising 100 different aircraft classes. As no suitable external knowledge graph was available, we utilized the structured metadata provided with the dataset. Specifically, each aircraft class is described using three hierarchical levels: 41 manufacturers, 70 families, and 100 variants. These are represented as three relations—hasManufacturer, hasFamily, and hasVariant—with each class associated with exactly one relation at each level, resulting in three relations per class.

Herbarium [9] is a dataset containing 46,000 herbarium specimens across more than 680 species within the flowering plant family Melastomataceae. Due to the limited rebuttal time, it was difficult to process the entire dataset, so we randomly selected 10 species and used DeepSeek-O1 to generate six attribute-based relations: hasLeafShape, hasLeafMargin, hasLeafVeins, hasFlowersColor, hasStem, and hasFruit, each with corresponding class-specific properties. The selected classes were evenly divided into 5 tasks, with 2 species per task.

Table 7: Performance comparison across fine-grained and medical datasets.

Datasets	CUB-200		HAM-10000		FGVC-Aircraft		Herbarium19	
	Avg	Last	Avg	Last	Avg	Last	Avg	Last
GMM	49.34	40.39	79.22	63.81	51.63	47.25	86.34	79.88
KG-GMM	52.65	43.84	81.45	64.34	51.88	47.88	88.59	81.14

The experimental results in Tab. 7 demonstrate that, with sufficiently rich KG information (CUB-200, HAM-10000), our method consistently improves upon GMM. When relevant knowledge is limited (as in the case of FVGC-Aircraft), the performance gains from our method are relatively modest.

### A.3 Analysis on training overhead

In Tab. 8 we present more detailed comparisons on memory footprint (additional memory occupied by graph-based labels), training overhead (training time per batch), KG time (time used for constructing sub-graph  $G'_t$ ), and the final accuracy. The experimental results demonstrate that the graph-based labels incur only a marginal overhead compared to word-based labels, requiring at most 169KB of additional memory and 0.07 seconds of extra processing time. Notably, when converting this 169KB capacity into exemplars for memory replay, the performance improvement proved statistically insignificant (from 80.72 to 80.91). This observation substantiates that our proposed method achieves substantial performance enhancements while maintaining minimal computational and storage overhead.

Table 8: Training overhead analysis on ImageNet-R.

Method	Memory Usage	Training Time	KG Time	Acc
GMM	244K	0.36	0	80.72
KG-GMM (task 1)	332K	0.40	2.46	—
KG-GMM (task 6)	368K	0.41	2.57	—
KG-GMM (task 10)	413K	0.43	2.73	84.29
GMM + 3 exemplar	484K	0.36	0	80.91

### A.4 More visualization results

In Fig. 6, we provide additional visual examples of our methods against the original GMM. It can be observed that KG-GMM utilizes attributes such as the bird’s color and location to refine its









Image	Ground-Truth label	GMM predicted text	GMM prediction	KG-GMM predicted text	KG-GMM Prediction
	Granny Smith	This is a photo of a apple.	Pineapple	Apple IsA Fruit Apple HasColor Green Apple AtLocation Store	Granny Smith
	Great white shark	This is a photo of a shark.	Great white shark	Shark IsA Animal Shark RelatedTo Great Shark AtLocation Ocean	Great white shark
	Hammerhead	This is a photo of a shark.	Great white shark	Shark RelatedTo Head Shark RelatedTo Hammer Shark AtLocation Sea	Hammerhead
	Goldfinch	This is a photo of a bird.	Hummingbird	Bird RelatedTo Gold Bird AtLocation Backyard Bird IsA Finch	Goldfinch
	Tree frog	This is a photo of a toad.	Newt	Toad RelatedTo Frog Toad AtLocation Tree Toad IsA Amphibian	Tree frog
	Tarantula	This is a photo of a spider.	Spider web	Spider IsA Arachnid Spider AtLocation USA Spider RelatedTo Tarantula	Tarantula
	Cobra	This is a photo of a snake.	Iguana	Snake IsA reptile Snake AtLocation zoo Snake RelatedTo python	Iguana
	Basset Hound	This is a photo of a hound.	Afghan Hound	Hound IsA hound Hound CapableOf bark Hound IsA dog	Beagle

Figure 6: More text examples of our methods against the original GMM. The first four examples illustrate how our method corrects the model drift in continual learning. The last two examples with red boxes, showcase where both our method and GMM made incorrect predictions.

classification to the finer-grained category of “goldfinch.” In contrast, the original GMM, although retaining knowledge of the broader category, loses the ability to distinguish finer details after learning additional concepts. However, for certain categories that belong to the same broader class (e.g., hound) and have nearly identical characteristics (i.e., similar relations), our method may still make errors, though it typically misclassifies them into closely related categories (e.g., misclassifying a Basset Hound as a Beagle rather than an Afghan Hound based on text similarity). We believe that if similar classes were more distinctly separated with different relations or finer-grained graphs with detailed attributes were incorporated, it would further enhance our method’s performance.

### A.5 Limitations and Border Impact

There are still several limitations in our work. As our work represents the first attempt to integrate knowledge graphs with continual learning, our experiments were primarily conducted on established continual learning benchmarks. We acknowledge that collecting KGs for some specialized datasets (e.g., fine-grained classification datasets and medical datasets) is tricky and remains unexplored. Furthermore, given that our primary baseline comparison focuses on GMM [5], our experiments were limited to consistent backbone architectures (EVA-CLIP and Vicuna-7B). Employing more advanced visual-language models could yield performance improvements. With the rapid development of LLMs, continual learning methods will become more important across various applications. Knowledge graphs emerge as a valuable yet previously underutilized tool for mitigating generalization loss during continual fine-tuning of LLMs. Exploring optimal methodologies for KG integration presents a promising research direction worthy of in-depth investigation.

## NeurIPS Paper Checklist

### 1. Claims

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

Answer: [\[Yes\]](#)

Justification: The abstract and introduction in Section 1 accurately reflect our contributions.

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 discuss the limitations of our work in Appendix A.5.

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: We do not include theoretical results.

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: All information are detailed in Section 4

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: [No]

Justification: All datasets can be obtained on the internet. We will soon open source our code.

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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide specific details of all experiments in Sec 4

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: [Yes]

Justification: We provide experiment results on three different class orders in the Appendix.

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: We provide sufficient information on the computer resources in Appendix A.3.

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 have read the NeurIPS Code of Ethics, and conduct research with it.

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: We discuss the broader impacts in Appendix A.5.

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: [NA]

Justification: The paper poses no such risks.

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: [Yes]

Justification: All existing assets we use are cited in section 4

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](https://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: The paper does not release new assets

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: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

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: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

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.

### 16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We describe the usage of the Vicuna 7B in Section 3

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.