

Aligning VLM Assistants with Personalized Situated Cognition



*Yongqi Li (/profile?id=~Yongqi_Li3), Shen Zhou (/profile?id=~Shen_Zhou2),
Xiaohu Li (/profile?id=~Xiaohu_Li2), Xin Miao (/profile?id=~Xin_Miao4),
Jintao Wen (/profile?id=~Jintao_Wen1), Mayi Xu (/profile?id=~Mayi_Xu1),
Jianhao Chen (/profile?id=~Jianhao_Chen3), Birong Pan (/profile?id=~Birong_Pan1),
Hankun Kang (/profile?id=~Hankun_Kang1), Yuanyuan Zhu (/profile?id=~Yuanyuan_Zhu1),
Ming Zhong (/profile?id=~Ming_Zhong8), Tieyun Qian (/profile?id=~Tieyun_Qian1)*

15 Feb 2025 (modified: 24 Apr 2025) ACL ARR 2025 February Submission February, Senior Area Chairs, Area Chairs, Reviewers, Authors, Ethics Reviewers, Ethics Chairs, Commitment Readers Revisions (/revisions?id=eXpIwiFZZQ) CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>)

Abstract:

Vision-language models (VLMs) aligned with general human objectives, such as being harmless and hallucination-free, have become valuable assistants of humans in managing visual tasks. However, people with diversified backgrounds have different cognition even in the same situation. Consequently, they may have personalized expectations for VLM assistants. This highlights the urgent need to align VLM assistants with personalized situated cognition for real-world assistance. To study this problem, we first simplify it by characterizing individuals based on the sociological concept of Role-Set. Then, we propose to evaluate the individuals' actions to examine whether the personalized alignment is achieved. Further, we construct a benchmark named PCogAlignBench, which includes 18k instances and 20 individuals with different Role-Sets. Finally, we present a framework called PCogAlign, which constructs a cognition-aware and action-based reward model for personalized alignment. Experimental results and human evaluations demonstrate the reliability of the PCogAlignBench and the effectiveness of our proposed PCogAlign. We will open-source the constructed benchmark and code after being accepted.

Paper Type: Long

Research Area: Multimodality and Language Grounding to Vision, Robotics and Beyond

Research Area Keywords: alignment, multimodality, personalization

Contribution Types: NLP engineering experiment, Data resources

Languages Studied: English

Reassignment Request Area Chair: This is not a resubmission

Reassignment Request Reviewers: This is not a resubmission

Software: zip (/attachment?id=eXpIwiFZZQ&name=software)

A1 Limitations Section: This paper has a limitations section.

A2 Potential Risks: N/A

B Use Or Create Scientific Artifacts: Yes

B1 Cite Creators Of Artifacts: Yes

B2 Discuss The License For Artifacts: Yes

B3 Artifact Use Consistent With Intended Use: Yes

B4 Data Contains Personally Identifying Info Or Offensive Content: Yes

B5 Documentation Of Artifacts: N/A

B6 Statistics For Data: Yes

B6 Elaboration: Section 4

C Computational Experiments: Yes

C1 Model Size And Budget: N/A

C2 Experimental Setup And Hyperparameters: Yes

C2 Elaboration: C.3

C3 Descriptive Statistics: Yes

C3 Elaboration: Section 6

C4 Parameters For Packages: Yes

C4 Elaboration: C

D Human Subjects Including Annotators: Yes

D1 Instructions Given To Participants: N/A

D1 Elaboration: The human annotation we use does not require specially designed guidelines or annotation software.

D2 Recruitment And Payment: Yes

D2 Elaboration: Ethics Statement

D3 Data Consent: Yes

D3 Elaboration: Section 4

D4 Ethics Review Board Approval: N/A

D5 Characteristics Of Annotators: N/A

E Ai Assistants In Research Or Writing: Yes

E1 Information About Use Of Ai Assistants: No

E1 Elaboration: We only use AI for basic format processing.

Author Submission Checklist: yes

Reviewing Volunteers: 👁 Yongqi Li (/profile?id=~Yongqi_Li3)

Reviewing No Volunteers Reason: 👁 N/A - At least one volunteer was provided in the previous question.

Reviewing Volunteers For Emergency Reviewing: 👁 The volunteers listed above are willing to serve either as regular reviewers or as emergency reviewers.

TLDR: 👁 A benchmark and a framework for exploring a new task of aligning VLM assistants with personalized situated cognition.

Preprint: 👁 no

Preprint Status: 👁 There is a non-anonymous preprint (URL specified in the next question).

Preferred Venue: 👁 ACL

Consent To Share Data: 👁 yes

Consent To Share Submission Details: 👁 On behalf of all authors, we agree to the terms above to share our submission details.

Association For Computational Linguistics - Blind Submission License Agreement: 👁 On behalf of all authors, I agree

Submission Number: 2717

Discussion (?id=eXpIwiFZZQ#discussion)

Filter by reply type...▼

Filter by author...▼

Search keywords...

Sort: Newest First

≡

≡

≡

-

=

≡

🔗

Official Comment



Issues: 1.1 Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing, 1.2 Avoid harm, 1.6 Respect privacy, 2.5 Give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks, 3.3 Manage personnel and resources to enhance the quality of working life

Explanation: This paper seeks to personalize visual language model (VLM) assistants by providing responses adapted to individuals' expectations and "situated cognition". As part of doing so, it gathers visual scenes from search engines (via LLM-generated scene descriptions) and LLM-generated queries. The paper's Ethics Statement section states: "The personalized alignment problem proposed in our work considers visual scenes that align with general human values, which will promote better collaboration between humans and AI, without posing potential ethical risk." However, even work aimed at socially good outcomes carries ethical risks -- see this statement from the program chairs of the 2024 FAccT conference for an elaboration (<https://medium.com/@alexandra.olteanu/responsible-ai-research-needs-impact-statements-too-7b7141031faf>) -- and I strongly urge the authors of this paper to expand their ethical considerations section by meaningfully engaging with the possible risks of this work. For example, as one reviewer notes, the dependence of the method on search engine and LLM results to obtain visual scenes and queries may have introduced stereotypes or other kinds of biases and harms into collected data, as work in NLP and elsewhere has amply demonstrated -- might this have happened, and was the quality control process equipped to investigate or mitigate this? I note that even in Figure 1 the individual with the mother role is also simply a community member, but the individual with the father role is a repairman in the community; of course this is just an example, but if there is a preponderance of such examples in the dataset (vs. e.g., individuals with mother roles who are also repairwomen) then stereotypes might be reproduced. Similarly, the paper seeks a diverse dataset, but it's not clear what/whose conception of diversity is meant. And personalization of this type -- adapting model responses to role-sets -- necessarily requires making assumptions about what responses are useful for individuals with different role-sets; of course the underlying assumption is that those assumptions enable more useful assistants, but those assumptions unavoidably carry risks as they encode beliefs about what people might be like and what they might want (or not want) from systems. Not engaging with these types of concerns risks violating Sections 1.1, 1.2, and 2.5 of the ACL Code of Ethics ("Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing", "Do no harm", "Give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks"). While the paper states in the Ethics Statement section that annotators were appropriately paid, the paper might also benefit from statements about whether informed consent was required and/or obtained from annotators not on the author team (Sections 1.6 and 3.3, "Respect privacy" and "Manage personnel and resources to enhance the quality of working life").

Add:

Author-Editor Confidential Comment

Official Comment



Response to insightful comments of Ethics Reviewer g3pP

Edit ▾



Official Comment

by Authors (Birong Pan (/profile?id=~Birong_Pan1), Yongqi Li (/profile?id=~Yongqi_Li3), Mayi Xu (/profile?id=~Mayi_Xu1), Jintao Wen (/profile?id=~Jintao_Wen1), +8 more (/group/info?id=aclweb.org/ACL/ARR/2025/February/Submission2717/Authors))

03 Apr 2025, 11:45 (modified: 24 Apr 2025, 23:28)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers, Ethics Chairs, Ethics Reviewers, Authors, Commitment Readers

Revisions (/revisions?id=uKCdMDexHw)

Comment:

Dear Ethics Reviewer,

We greatly appreciate your valuable suggestions, which will significantly help us avoid potential ethical issues in our study. We will expand the ethical considerations section to address potential ethical risks in this research. Below is a further explanation of the sections we intend to expand:

1. During the data collection process, we made necessary designs in the role definition phase and quality control process to mitigate potential ethical risks. Specifically, in the role definition phase, we thoroughly discussed and defined roles that meet ethical standards to avoid factors like gender bias. For instance, to avoid gender bias, we defined that the responsibilities of "mother@home" and "father@home" are fairly distributed, including shared household chores, childcare, and handling family emergencies. Besides, in our human-led quality control process, we aimed to "avoid potential ethical risks, including gender and racial biases, etc." For example, if an annotator repeatedly encounters images showing "women doing housework" and relatively few "men doing housework" images, they are required to replace some of the "women doing housework" images with "men doing housework". (Thanks to the carefully designed role definitions which considers avoiding potential bias, the automatically generated visual scene descriptions have largely avoided such biases, and we encountered less than 1% of such cases during the quality control process.)
2. Considering the vast number of potential role combinations (theoretically 6300 combinations in our setup), which significantly reduces the feasibility of conducting academic experiments, we selected a subset of role combinations (20 Role-Sets) to form the Role-Sets in our dataset. This might raise concerns of certain bias. Although we believe this is acceptable in a simulated research environment, in actual industry development, we encourage companies/developers to consider various user backgrounds to form comprehensive, unbiased Role-Sets for data collection and personalized alignment training.
3. The "diversity" of the dataset encompasses two aspects: increasing the diversity of scenarios, and ensuring diversity to avoid potential biases, such as racial and gender biases.
4. We obtained informed consent from all annotators about the statement of "all human annotators being paid by the laboratory following local wage requirements" before the manuscript submission.

Additionally, we fully agree with you about the risks related to AI personalization you mentioned, i.e., "... of course, the underlying assumption is that those assumptions enable more useful assistants, but those assumptions unavoidably carry risks as they encode beliefs about what people might be like and what they might want (or not want) from systems". We also think that this is an important issue arising from technological advancements that cannot be overlooked. However, since this issue may pertain more to the human-computer interaction (HCI) research field, it might be beyond the scope of our current work. We hope to see further research and discussion on the personalization challenges within the HCI field in the future.

Again, thank you for your detailed feedback, which greatly helps to enhance our work.

Best regards,
Paper 2717 Authors

Add: Author-Editor Confidential Comment Official Comment



Official Comment by Ethics Reviewer g3pP

Official Comment by Ethics Reviewer g3pP 📅 04 Apr 2025, 04:20 (modified: 24 Apr 2025, 23:28)
👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers, Ethics Chairs, Ethics Reviewers, Authors, Commitment Readers
📄 Revisions (/revisions?id=AdM9YTuijc)

Comment:
Thank you for your response; the explicit engagement in the annotation process with potential biases, and the description of the planned additional detail, mitigates to a significant extent these concerns.

Add: Author-Editor Confidential Comment Official Comment



➔ *Replying to Official Comment by Ethics Reviewer g3pP*

Follow-up Response

Edit ▼



Official Comment

by Authors (Birong Pan (/profile?id=~Birong_Pan1), Yongqi Li (/profile?id=~Yongqi_Li3), Mayi Xu (/profile?id=~Mayi_Xu1), Jintao Wen (/profile?id=~Jintao_Wen1), +8 more (/group/info?id=aclweb.org/ACL/ARR/2025/February/Submission2717/Authors))

04 Apr 2025, 07:30 (modified: 24 Apr 2025, 23:28)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers, Ethics Chairs, Ethics Reviewers, Authors, Commitment Readers

Revisions (/revisions?id=DhITkE3le5)

Comment:

Dear Ethics Reviewer,

We are very pleased that our response has significantly alleviated your concerns about potential ethics issues in our work. This is crucial for ensuring that our work provides beneficial value to society without posing ethical risks. Thank you again for your valuable suggestions!

Best regards,
Paper 2717 Authors

Add:

Author-Editor Confidential Comment

Official Comment



Official Review of Submission2717 by Reviewer DqCx

Official Review by Reviewer DqCx 25 Mar 2025, 11:03 (modified: 24 Apr 2025, 23:28)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer DqCx, Commitment Readers

Revisions (/revisions?id=T7UYoHhtWR)

Paper Summary:

This work focuses on aligning Vision-Language Model (VLM) assistants with personalized situated cognition for real-world assistance. The authors introduce a new benchmark, PCogAlignBench, to explore this novel task. They also present a framework called PCogAlign, which constructs a cognition-aware and action-based reward model for personalized alignment.

Summary Of Strengths:

1. Aligning VLM assistants with personalized situated cognition is an interesting and valuable task for both VLM and embodied AI.
2. The proposed dataset is very helpful for future research.
3. The writing is clear and easy to follow.

Summary Of Weaknesses:

1. Could you provide samples from PCogAlignBench? What does one sample include?
2. The performance improvement of PCogAlign does not appear to be very promising, as shown in Table 1.

Comments Suggestions And Typos:

N/A

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

Soundness: 4 = Strong: This study provides sufficient support for all of its claims. Some extra experiments could be nice, but not essential.

Excitement: 4.0 = Exciting: I would mention this paper to others and/or make an effort to attend its presentation in a conference.

Overall Assessment: 3.5 = Borderline Conference

Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No

Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

Datasets: 5 = Enabling: The newly released datasets should affect other people's choice of research or development projects to undertake.

Software: 1 = No usable software released.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Add:

Author-Editor Confidential Comment

Official Comment



Response to insightful comments of Reviewer DqCx

Edit



Official Comment

by Authors (Birong Pan (/profile?id=~Birong_Pan1), Yongqi Li (/profile?id=~Yongqi_Li3), Mayi Xu (/profile?id=~Mayi_Xu1), Jintao Wen (/profile?id=~Jintao_Wen1), +8 more (/group/info?id=aclweb.org/ACL/ARR/2025/February/Submission2717/Authors))

02 Apr 2025, 08:54 (modified: 24 Apr 2025, 23:28)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer DqCx, Commitment Readers

Revisions (/revisions?id=EbL4dr1xgN)

Comment:

Deer Reviewer,

We are delighted that our work has received your recognition, and we deeply appreciate your dedicated efforts. Below are our responses to address any remaining concerns caused by unclear presentations.

Concern 1. Could you provide samples from PCogAlignBench? What does one sample include?

A: To facilitate your understanding of the collected samples in the PCogAlignBench, we present a training sample and a test sample in JSON format.

An example from the training split:

```
{
  "individual_RoleSet": {
    "Home": "Grandma", "Community": "Member", "Museum": "Visitor", "Airport": "Passenger", "Store": "Shelf Stocker"
  },
  "image": {
    "file_path": "images/HCMAS/train/HCMAS-train-I10-Community-22-1.png"
  },
  "query": "What kind of party is this?"
}
```

An example from the test split:

```
{
  "individual_RoleSet": {
    "Home": "Father", "Community": "Policeman", "Museum": "Visitor", "Airport": "Passenger", "Store": "Customer"
  },
  "image": {
    "file_path": "images/HCMAS/test/HCMAS-test-I2-Community-13-0.png"
  },
  "query": "What safety tips should I share with kids during events?",
  "eval_info": {
    "Image Description": "A young boy in a blue jacket stands confidently outdoors, surrounded by colorful buildings and a bustling crowd behind him.",
    "Oracle Guidance ": "As \"A Policeman at Community (A law enforcement officer who maintains public order, ensures safety, and enforces local laws);\n\" (Primary Role) and \"A Father at Home; A Visitor at Museum; A Passenger at Airport; A Customer at Store; \n\" (Secondary Roles), I want to convey important safety measures to children in an engaging manner. \n- Body Behavior: I want to be approachable and friendly, demonstrating a reassuring presence that encourages kids to listen and participate. \n- Mind Feelings: I want to feel confident and responsible, ensuring that I communicate vital safety tips in a way that is easy for kids to understand and remember. \nI appreciate AI assistance that provides clear, practical safety tips relevant to the context of the event, along with interactive ideas to keep the children engaged, such as incorporating games or activities that reinforce those safety messages."
  }
}
```

Each training sample consists of "individual_RoleSet", "image", and "query". Each test sample consists of "individual_RoleSet", "image", "query", and "eval_info" (including image description and oracle guidance).

There are two special considerations about the collected samples:

1. Unlike previous QA datasets, we do not include the “expected response” in the training sample because we hope the alignment methods evaluated on our benchmark contain a sampling strategy to generate responses from the target VLM in an on-policy optimization manner.
2. To enable accurate and automatic evaluation of the test split, we adopt a collaborative approach involving humans and GPT-4 to collect “oracle guidance” for each test sample. This “oracle guidance” is then used as auxiliary information to score the evaluated response via the llm-as-a-judge evaluation (described in Section 4.2).

Concern 2. The performance improvement of PCogAlign is not very promising

A: Since there is no previous method designed for this new task, we have made some adaptations to previous methods as baselines. It is important to note that **the best baseline** Self-Refine (S) (and other SFT/DPO baselines) **is already enhanced by** the “Cognition and Action Estimation” and “Key Points Generation” strategies described in Sections 5.1 and 5.2, which are **important modules of our proposed PCogAlign**. Even with these enhanced baselines, the results in Table 1 demonstrate that PCogAlign surpasses the best baseline by an average of **2.4%** in Win Rate, which clearly validates the effectiveness of our proposed method.

Thank you once again for your hard work during the review process! We hope our clarification above helps mitigate misunderstandings and strengthens your confidence in our contributions.

Best regards,
Paper 2717 Authors.

Add:

Author-Editor Confidential Comment

Official Comment



Official Review of Submission2717 by Reviewer h5u4

Official Review by Reviewer h5u4 📅 24 Mar 2025, 20:47 (modified: 24 Apr 2025, 23:28)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer h5u4, Commitment Readers

📄 Revisions (/revisions?id=qYxVJtemz0)

Paper Summary:

This paper addresses how individuals with different social roles and backgrounds can hold varying expectations for the same visual context. Existing alignment methods typically produce one-size-fits-all solutions, overlooking these individualized needs. By contrast, the authors propose aligning VLMs with "personalized situated cognition", ensuring that model outputs reflect each user's specific situation and intended course of action. Key contributions:

1. Role-Set Characterization: The authors introduce the idea of a "Role-Set" to account for individual differences Each set of roles captures the user's social positions, motivations, and expectations - even for the same visual stimulus.
2. PCogAlignBench: They create a benchmark dataset of 18k instances, each containing: (1) a Role-Set, (2) an image, and (3) a query from the user. For the test split, they also provide "oracle guidance", which describes the expected characteristics of an ideal personalized response.
3. PCogAlign: Their proposed framework first estimates the user's cognitive and emotional state in a given scene, along with a desirable next action. Two subsystems then work together to generate multiple candidate responses tailored to the user's background. A specialized reward model compares these candidates, selects the best one, and uses it to further refine the main model's alignment via supervised fine-tuning.

Experimental results suggest that this personalized approach outperforms general-purpose alignment models, thereby illustrating the benefit of modeling individual differences.

Summary Of Strengths:

- Problem Significance: The paper introduces a benchmark specifically for personalized VLM alignment and lays out a clear evaluation protocol. This angle is sound and addresses a notable research gap.
- Fine-grained Evaluation: The authors conduct a comprehensive assessment on the testing data and adopt a breakdown of scoring dimensions and demonstrate strong consistency between GPT-based and human-based assessments.

Summary Of Weaknesses:

- Approach Novelty: Although the paper introduces a new benchmark, the lack of actual visual content in the sample undercuts its multimodal claim. In practice, the dataset appears more akin to standard visual QA plus personalized large language modeling rather than a vision-language setup.

- Data Quality Concerns: Human quality checks are applied only to the test set. Given the authors' mention of low-quality images at line 313 (around 1/6 needing replacement), there is some concern about the overall quality of the training data.
- Marginal Gains: While the results show improvement, the performance edge over baselines is often under one score, calling into question the practical significance. Although author mentioned their method is relatively simple in the limitations section, its overall effectiveness remains somewhat unconvincing.

Comments Suggestions And Typos:

A question here: Have you tried repeating the three-step optimization process multiple times? If so, does it yield additional performance improvements beyond what is reported?

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Soundness: 3.5

Excitement: 2.5

Overall Assessment: 3 = Findings: I think this paper could be accepted to the Findings of the ACL.

Ethical Concerns:

The authors included the role sets used to make situated environments. However, such environments can stimulate the biases in LLM. For example, as shown in the figure in the introduction. A mother may be instructed to care for their baby but a father does not, which creates bias and potential discrimination on certain groups. The author needs to state to what extent such bias may occur in their dataset and how they control the biases in the dataset/framework.

Needs Ethics Review: Yes

Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

Datasets: 3 = Potentially useful: Someone might find the new datasets useful for their work.

Software: 1 = No usable software released.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Add:

Author-Editor Confidential Comment

Official Comment



**Response to insightful
comments of Reviewer h5u4 (2
/ 2)**

Edit ▾



Official Comment

by Authors (Birong Pan (/profile?id=~Birong_Pan1), Yongqi Li (/profile?id=~Yongqi_Li3), Mayi Xu (/profile?id=~Mayi_Xu1), Jintao Wen (/profile?id=~Jintao_Wen1), +8 more (/group/info?id=aclweb.org/ACL/ARR/2025/February/Submission2717/Authors))

02 Apr 2025, 08:56 (modified: 24 Apr 2025, 23:28)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer h5u4, Commitment Readers

Revisions (/revisions?id=UvmDYZHWqU)

Comment:

Question. Have you tried repeating the three-step optimization process multiple times?

A: Yes, that is a really good question! In fact, the proposed framework can essentially be seen as an "on-policy" alignment framework, which means that the trained model is also used for sampling responses.

Following your suggestion, we have tried to repeat the three-step optimization process more rounds. The results are shown in the Table 1.

Method	Rounds	LS1→LS1				LS1→LS2			
		P. Score	Win↑	Tie	Lose↓	P. Score	Win↑	Tie	Lose↓
Base	-	3.781	0.0%	100.0%	0.0%	3.715	0.0%	100.0%	0.0%
PCogAlign	1	4.161	53.3%	29.0%	17.7%	4.156	56.6%	27.5%	15.9%
PCogAlign	2	4.170	54.7%	28.8%	16.5%	4.172	56.6%	28.9%	14.5%

Table 1: Experimental results (P. Score and Win Rate) under four settings. The P. Score ranges from 1 to 5. The Win/Tie/Lose Rate is obtained by comparing the P. Score of the response being evaluated with that of the Base response.

From the results in Table 1, we can see that **further round of alignment training leads to additional performance improvements over the reported ones**, even though we have not thoroughly designed the iterative strategy. We believe that it is interesting to explore how to further enhance such "on-policy" personalized alignment training in the future. We will include relevant experiments and discussions in the revised version.

Additionally, such experiments demonstrate that our proposed **PCogAlign has the potential to achieve better performance with further rounds of alignment training**. This may also help alleviate your concern about the "marginal gains of the proposed PCogAlign" (Concern 3).

Ethical Concern

A: We sincerely apologize for any misunderstanding or inconvenience caused by the example in Figure 1. In fact, the toy example in Figure 1 aims to emphasize the different expectations between “a mother and a repairman” but not between “a mother and a father”. In other words, in **real** samples in our benchmark, the response provided to the individual with Role-Set “Father@Home, Repairman@Community” is expected to include **both professional advice on repair work and guidance on taking care of the child**. We will make modifications to Figure 1 to avoid similar misunderstandings.

Thank you once again for your efforts during the review process! We hope our clarification above helps eliminate misunderstandings and provides a clearer understanding of our contributions.

Best regards,
Paper 2717 Authors.

Add:

Author-Editor Confidential Comment


Official Comment





Response to insightful
comments of Reviewer h5u4 (1 /
2)

Edit ▾

Official Comment

by Authors ( Birong Pan (/profile?id=~Birong_Pan1), Yongqi Li (/profile?id=~Yongqi_Li3), Mayi Xu (/profile?id=~Mayi_Xu1), Jintao Wen (/profile?id=~Jintao_Wen1), +8 more (/group/info?id=aclweb.org/ACL/ARR/2025/February/Submission2717/Authors))

 02 Apr 2025, 08:56 (modified: 24 Apr 2025, 23:28)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer h5u4, Commitment Readers

 Revisions (/revisions?id=sHSVOfbBji)

Comment:

Deer Reviewer,

We deeply appreciate your dedicated work and valuable comments. We hope the following clarifications can alleviate any misunderstanding caused by unclear presentations.

Concern 1. Novelty of the vision-language setup

A: Beyond simply combining “standard visual QA” and “personalized large language modeling” to create a personalized visual QA setup, one of the most important novelties/contributions/claims of our work is that **we are the first to consider personalized situated cognition in visual QA**. The focused personalized situated cognition inspired by cognitive science is a crucial factor in enhancing the communication efficiency of human-AI interaction.

The existence of personalized situated cognition requires the VLM assistant to understand **both the vision information** (image) **and the language information** (user query and Role-Set) to respond **in a personalized manner**. For example, in Figure 1, without the visual information of “a broken swing” in the image, there would be **various** possible personalized situated cognitions, and it would be impossible for the assistant to accurately analyze the user’s personalized situated cognition and provide a well-personalized response.

Therefore, we believe that our proposed task inherently involves a vision-language setup.

Concern 2. The quality of the training data without human checks

A: We do not apply time-consuming human checks on training data mainly because **we want to simulate the challenge of training data noise in real-world applications**. Such challenge will encourage future exploration on developing **robust** personalized alignment methods.

The primary reason for such a simulation is that in some scenarios, **the system is unable to collect training data from the user side due to reasons like privacy protection**. For example, a user might consider interactions in settings like homes or hospital rooms to potentially contain private images or user queries, and may not agree to the system using such data for model training.

Therefore, in such scenarios, **we need to gather relevant data from public large-scale image databases**, such as automatically retrieving role-related images from the internet. This automatic collection process certainly contains many noises and is unable to employ human checks.

Concern 3. Marginal gains of the proposed PCogAlign

A: Since there is no previous method designed for this new task, we have made some adaptations to previous methods as baselines. It is important to note that **the best baseline** Self-Refine (S) (and other SFT/DPO baselines) **is already enhanced by** the “Cognition and Action Estimation” and “Key Points Generation” strategies described in Sections 5.1 and 5.2, which are **important modules of our proposed PCogAlign**. Even with these enhanced baselines, the results in Table 1 demonstrate that PCogAlign surpasses the best baseline by an average of **2.4%** in Win Rate, which clearly validates the effectiveness of our proposed method.

Add:

Author-Editor Confidential Comment

Official Comment



Official Review of Submission2717 by Reviewer LPh7

Official Review by Reviewer LPh7 📅 24 Mar 2025, 12:16 (modified: 24 Apr 2025, 23:28)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer LPh7, Commitment Readers

📄 Revisions (/revisions?id=Yw7OcQRj8L)

Paper Summary:

The paper works on the problem of aligning VLM assistants with personalized situated cognition. The idea is that different individuals interpret the same visual scene differently depending on their roles (e.g., child vs. repairman). They introduce PConAlignBench, which is an evaluation corpus with 18k instances. They also introduce the PCogAlign framework to align VLMs with these expectations. This framework outperforms standard baselines across five dimensions on their benchmark.

Summary Of Strengths:

- The paper tackles an impactful problem of personalized situated cognition. The notations are clear.
- The benchmark could be a valuable resource to evaluate state-of-the-art VLMs on this task, however, evaluations in this paper are only limited to variations of Qwen2-VL. The collection process of PCogAlignBench is intuitive and sound.
- They propose PCogAlign, a two-stage framework that aligns VLMs with personalized situated cognition by first estimating an individual's likely cognitive state and optimal action based on their Role-Sets and then generating tailored responses using two agents, guided by a cognition-aware reward model to select the most contextually appropriate output. The results on all settings demonstrate that PCogAlign outperforms the baselines.

Summary Of Weaknesses:

- The paper lacks important statistics on the quality of the collection process, such as "what percentage of data was discarded during human annotation." This would help to know how much noise is present in the training set (which hasn't been annotated by humans) and whether the benchmark can be expanded automatically.
- The instructions provided to human annotators are not presented and it's unclear how they recruited the annotators. L. 315 uses the term "trained annotator" without explaining what training means in this case.
- The collection process refers to the terms "Visual Scene Type", "Visual Scene Phrase" and "Visual Scene Description" without explaining what they are and how they differ. I think the authors can include an example of the dataset in the paper (maybe in Figure 2).
- The reason behind training and evaluating LS1 and LS2 in 4 different settings is unclear. Is this supposed to show one subset is better than the other? Isn't it that they were both collected through the same process with the only difference being the role locations?
- The paper misses an important RAG-based baseline. The training samples can be retrieved based on Role-Sets and be used as in-context examples during test split evaluation. This baseline should be compared with PCogAlign to understand the significance of PCogAlign.
- Additionally, the paper evaluates using only one model (Qwen2-VL). I think PCogAlignBench has the potential to be incorporated to evaluate more VLMs, this should also help to understand the significance of PCogAlign.

Comments Suggestions And Typos:

Footnote superscript should go after punctuation. (L. 043, L 336, 487, etc.)

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

Soundness: 3.5

Excitement: 3.5

Overall Assessment: 4.0 = Conference: I think this paper could be accepted to an *ACL conference.

Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No

Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

Datasets: 4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.

Software: 3 = Potentially useful: Someone might find the new software useful for their work.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.


Add:


Author-Editor Confidential Comment


Official Comment



Official Comment by Reviewer LPh7

Official Comment by Reviewer LPh7  04 Apr 2025, 03:49 (modified: 24 Apr 2025, 23:28)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer LPh7, Commitment Readers

 Revisions (/revisions?id=5esLx6cUTU)

Comment:

Thanks for the careful explanations. My major concerns are resolved, and based on the RAG-baselines provided (which seems to be a better baseline, and I hope the authors incorporate it in their final version) and the new models added to the tables, I will increase my rating from 3 to 4.

Add:

Author-Editor Confidential Comment

Official Comment





Follow-up Response


Edit ▾




Official Comment

by Authors ( Birong Pan (/profile?id=~Birong_Pan1), Yongqi Li (/profile?id=~Yongqi_Li3), Mayi Xu (/profile?id=~Mayi_Xu1), Jintao Wen (/profile?id=~Jintao_Wen1), +8 more (/group/info?id=aclweb.org/ACL/ARR/2025/February/Submission2717/Authors))

 04 Apr 2025, 07:26 (modified: 24 Apr 2025, 23:28)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer LPh7, Commitment Readers

 Revisions (/revisions?id=ORJkdVNj6i)

Comment:

Dear Reviewer,

Thank you so much for taking the time to read our responses and raise the score. We are pleased to know that we have addressed your concerns. We will enhance the clarity of the paper and include experimental results in the next version based on your suggestions. Such a discussion is helpful to improve our work. Thanks again.

Best regards,
Paper 2717 Authors.

Add:

Author-Editor Confidential Comment

Official Comment



Response to insightful comments of Reviewer LPh7 (3 / 3)

Edit



Official Comment

by Authors (Birong Pan (/profile?id=~Birong_Pan1), Yongqi Li (/profile?id=~Yongqi_Li3), Mayi Xu (/profile?id=~Mayi_Xu1), Jintao Wen (/profile?id=~Jintao_Wen1), +8 more (/group/info?id=aclweb.org/ACL/ARR/2025/February/Submission2717/Authors))

02 Apr 2025, 09:00 (modified: 24 Apr 2025, 23:28)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer LPh7, Commitment Readers

Revisions (/revisions?id=PB2F0Kvb4M)

Comment:

Concern 6. Evaluate more VLMs

A: In our original paper, we have considered two VLMs from two series to compare different personalized alignment methods, including:

1. Qwen2-VL-7B-Instruct from the Qwen2-VL series. The results are shown in Table 1 in the main text.
2. Qwen2.5-VL-3B-Instruct from the Qwen2.5-VL series. The results are shown in Table 8 in the appendix.

Nonetheless, we agree with you that **our proposed PCogAlignBench can also serve as a benchmark for evaluating the personalization adaptation ability of different VLMs.**

To this end, we've included several VLMs from additional series to benchmark their personalization adaption abilities. Due to the limited time of the rebuttal period, we consider the prompt-based baselines and the prompt variant of our PCogAlign for this experiment. The results are also shown in the following Table 2.

VLM	Method	LS1→LS1		LS1→LS2		LS2→LS1		LS2→LS2		Average	
		<i>P. Score</i>	<i>Win Rate</i>	<i>P. Score</i>	<i>Win Rate</i>	<i>P. Score</i>	<i>Win Rate</i>	<i>P. Score</i>	<i>Win Rate</i>	<i>P. Score</i>	<i>Win Rate</i>
Qwen2-VL-7B-Instruct	Base	3.781	0.0%	3.715	0.0%	3.781	0.0%	3.715	0.0%	3.748	0.0%
	RS Prompt	3.989	44.1%	4.008	47.1%	3.989	44.1%	4.008	47.1%	3.999	45.6%
	RAG	4.022	43.9%	3.949	44.3%	4.046	46.4%	3.952	45.8%	3.992	45.1%
	PCogAlign (P)	4.070	47.5%	4.056	50.3%	4.070	47.5%	4.056	50.3%	4.063	48.9%
Qwen2.5-VL-7B-Instruct	Base	4.126	0.0%	4.079	0.0%	4.126	0.0%	4.079	0.0%	4.102	0.0%
	RS Prompt	4.163	31.0%	4.122	33.0%	4.163	31.0%	4.122	33.0%	4.143	32.0%
	RAG	4.242	38.1%	4.184	37.5%	4.220	36.4%	4.200	37.3%	4.212	37.3%
	PCogAlign (P)	4.275	40.8%	4.277	44.2%	4.275	40.8%	4.277	44.2%	4.276	42.5%
Phi-3.5-vision-instruct	Base	3.268	0.0%	3.235	0.0%	3.268	0.0%	3.235	0.0%	3.251	0.0%

VLM	Method	LS1→LS1		LS1→LS2		LS2→LS1		LS2→LS2		Average	
MiniCPM-V-2_6	RS Prompt	3.419	40.3%	3.379	38.1%	3.419	40.3%	3.379	38.1%	3.399	39.2%
	RAG	3.730	65.4%	3.613	58.8%	3.772	67.0%	3.684	62.1%	3.700	63.3%
	PCogAlign (P)	3.797	67.3%	3.745	64.2%	3.797	67.3%	3.745	64.2%	3.771	65.7%
	Base	3.743	0.0%	3.725	0.0%	3.743	0.0%	3.725	0.0%	3.734	0.0%
	RS Prompt	4.055	60.7%	4.037	60.7%	4.055	60.7%	4.037	60.7%	4.046	60.7%
	RAG	4.261	79.8%	4.232	78.4%	4.262	79.3%	4.235	78.8%	4.248	79.1%
	PCogAlign (P)	4.303	80.1%	4.321	81.0%	4.303	80.1%	4.321	81.0%	4.312	80.5%

Table 2: Experimental results (P. Score and Win Rate) under four settings. The best results on each VLM are in bold. The P. Score ranges from 1 to 5. The Win Rate is obtained by comparing the P. Score of the response being evaluated with that of the Base response.

As depicted in Table 2, the MiniCPM-V-2_6 shows the strongest personalization adaptation ability (average P. Score). Still, our PCogAlign (P) method consistently outperforms the baselines across all VLMs. We will include more relevant experiments and discussions in the revised version.

Typos

Thanks for your careful reading, we will fix it per your suggestion.

Thank you once again for your efforts during the review process! We hope our clarifications above help resolve any misunderstanding and offer a clearer understanding of our contributions.

Best regards,
Paper 2717 Authors.

Add:


Author-Editor Confidential Comment


Official Comment





Response to insightful comments of Reviewer LPh7 (2 / 3)

Official Comment

by Authors ( Birong Pan (/profile?id=~Birong_Pan1), Yongqi Li (/profile?id=~Yongqi_Li3), Mayi Xu (/profile?id=~Mayi_Xu1), Jintao Wen (/profile?id=~Jintao_Wen1), +8 more (/group/info?id=aclweb.org/ACL/ARR/2025/February/Submission2717/Authors))

 02 Apr 2025, 09:00 (modified: 24 Apr 2025, 23:28)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer LPh7, Commitment Readers

 Revisions (/revisions?id=TACdDxRRmj)

Comment:

Edit ▾



Concern 5. RAG-based baseline

A: Many thanks for your insightful suggestion! Following your advice, we include the RAG-based baseline. Specifically, during the evaluation on the test split, we retrieve 3 in-context examples from the training samples based on Role-Sets. We use the Levenshtein distance as the retrieval metric because it can accurately reflect the similarity between Role-Sets.

The comparison results on the Qwen2-VL-7B-Instruct (the same model used in the main experiments in the paper) are presented in Table 1.

VLM	Method	LS1→LS1		LS1→LS2		LS2→LS1		LS2→LS2		Average	
		<i>P. Score</i>	<i>Win Rate</i>	<i>P. Score</i>	<i>Win Rate</i>	<i>P. Score</i>	<i>Win Rate</i>	<i>P. Score</i>	<i>Win Rate</i>	<i>P. Score</i>	<i>Win Rate</i>
Qwen2-VL-7B-Instruct	Base	3.781	0.0%	3.715	0.0%	3.781	0.0%	3.715	0.0%	3.748	0.0%
	RS Prompt	3.989	44.1%	4.008	47.1%	3.989	44.1%	4.008	47.1%	3.999	45.6%
	RAG	4.022	43.9%	3.949	44.3%	4.046	46.4%	3.952	45.8%	3.992	45.1%
	PCogAlign (P)	4.070	47.5%	4.056	50.3%	4.070	47.5%	4.056	50.3%	4.063	48.9%
	PCogAlign	4.161	53.3%	4.156	56.6%	4.150	51.4%	4.151	53.8%	4.154	53.8%

Table 1: Experimental results (P. Score and Win Rate) under four settings. The best results are in bold. The P. Score ranges from 1 to 5. The Win Rate is obtained by comparing the P. Score of the response being evaluated with that of the Base response.

From Table 1, we can observe that:

1. Our **PCogAlign (P)**, i.e., the tuning-free version of our PCogAlign, **outperforms the RAG baseline**. It’s important to note that **PCogAlign (P) is tuning-free and does not require access to training data**, whereas the RAG baseline requires access to the training data.
2. Our proposed PCogAlign, which utilizes a specifically designed reward model for personalized alignment training, shows improvements over our PCogAlign (P) variant, and surpasses the RAG baseline with remarkable scores (8.7% in Win Rate).

Add:

Author-Editor Confidential Comment

Official Comment



Response to insightful
comments of Reviewer LPh7 (1 / 3)

Edit ▾

Official Comment

by Authors (Birong Pan (/profile?id=~Birong_Pan1), Yongqi Li (/profile?id=~Yongqi_Li3), Mayi Xu (/profile?id=~Mayi_Xu1), Jintao Wen (/profile?id=~Jintao_Wen1), +8 more (/group/info?id=aclweb.org/ACL/ARR/2025/February/Submission2717/Authors))

02 Apr 2025, 08:58 (modified: 24 Apr 2025, 23:28)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer LPh7, Commitment Readers

Revisions (/revisions?id=pdnXjluNND)

Comment:

Deer Reviewer,

We are delighted that our work has received your recognition, and we deeply appreciate your detailed comments and suggestions. To further alleviate your remaining concerns, we make the following responses to your questions.

Concern 1. Statistics on the quality of the collection process

A: We will provide more details about the collection process in the appendix, especially statistics related to human checks. Here are some brief statistics from the human quality control on the test split:

- 1. For images, approximately 1,000 images in the test samples (17%) were replaced by human-collected ones. During this process, annotators not only need to ensure that the images show proper scenarios in the expected locations, but also need to consider the diversity of the images. In other words, even if an image is high-quality, if it appears multiple times (>2 times) in the dataset, it is replaced with a new one.
- 2. Thanks to the well-designed query collection strategy and high-quality human-checked images, only about 20 queries (0.3%) were found to be unanswerable and revised by annotators.

We will include these collection details in the open-sourced benchmark to assure readers of the dataset's quality.

Concern 2. Meaning of “trained annotator”

A: We will present detailed instructions provided to the human annotator in the appendix. Besides, the term "trained annotator" in L. 315 means that:

Human annotators (master/Ph.D. students) who have a background of natural language processing or computer vision, are told about the specific requirements of the expected high-quality samples in a group meeting. Then, these annotators check a small set of the automatically samples and replace 20 low-quality samples based on their understandings about the quality control requirements. After that, the benchmark construction manager checks their replaced samples with the help of two other annotators. Finally, the benchmark construction manager provides feedback to the annotators to help them better understand the quality control requirements.

In this way, the annotators can be trained as trained ones, who can conduct reliable checks for the automatically collected data.

Concern 3. Explaining terms in the collection process and how they differ

A: Due to the limited space in the main text, we have presented prompt templates in **Table 10 in appendix** used for collecting visual scene types, visual scene phrases, and visual scene descriptions, where we use demonstrations (examples) in each prompt template to explain what are "Visual Scene Type", "Visual Scene Phrase", and "Visual Scene Description" and how they differ. We will include an example in Figure 2 to make these terms clearer to readers.

Here, for your convenience, we have copied examples from Table 10 in appendix to explain these terms.

"Visual Scene Type": "Household Labour"
"Visual Scene Phrase" (expanded by the type "Household Labour"): ["Wall cleaning", "Window washing", "Garden care", "Dishwashing", "Tidying"]
"Visual Scene Description" (expanded by the phrase "Wall cleaning"): "a smudged wall in a hallway, with a bucket of soapy water and a sponge nearby, ready for cleaning"

Concern 4. The reason behind training and evaluating LS1 and LS2 in 4 different settings

A: The reason for considering four different settings, i.e., “LS1→LS1”, “LS1→LS2”, “LS2→LS1”, and “LS2→LS2”, is **not** to show that one subset is better than the other.

The reasons are outlined in **Lines 246-255** in the paper. To further clarify and address your concerns, here is additional explanation: We recognize that in real-world applications, there might be totally new user's Role-Sets that have not been encountered in training. **Inspired by studies on out-of-distribution (OOD) generalization**, we intend to evaluate whether each method remains effective when Role-Sets in the test split are unseen during training. To this end, **we divide the original data into two subsets with non-overlapping Role-Sets**, "LS1" and "LS2". We then conduct experiments on two "OOD" settings, i.e., "LS1→LS2" and "LS2→LS1", and two normal settings, i.e., "LS1→LS1" and "LS2→LS2".

Add:

Author-Editor Confidential Comment

Official Comment