# Cell-Based Representation of Relational Binding in Language Models

**Qin Dai**[1] **Benjamin Heinzerling**[2,1] **Kentaro Inui**[3,1,2]
[1]Tohoku University [2]RIKEN AIP [3]MBZUAI
qin.dai.b8@tohoku.ac.jp benjamin.heinzerling@riken.jp
kentaro.inui@mbzuai.ac.ae

## Abstract

Understanding a discourse requires tracking entities and the relations that hold between them. While Large Language Models (LLMs) perform well on relational reasoning, the mechanism by which they bind entities, relations, and attributes remains unclear. We study discourse-level relational binding and show that LLMs encode it via a Cell-based Binding Representation (CBR): a low-dimensional linear subspace in which each "cell" corresponds to an entity–relation index pair, and bound attributes are retrieved from the corresponding cell during inference. Using controlled multi-sentence data annotated with entity and relation indices, we identify the CBR subspace by decoding these indices from attribute-token activations with Partial Least Squares regression. Across domains and two model families, the indices are linearly decodable and form a grid-like geometry in the projected space. Finally, activation patching shows that manipulating this subspace systematically changes relational predictions and that perturbing it disrupts performance, providing causal evidence that LLMs rely on CBR for relational binding. Code and data are available at https://github.com/cl-tohoku/IRS-Subsapce.

## 1 Introduction

A core requirement for language comprehension is to keep track of entities and the relations between them as a discourse unfolds (1; 2; 3). It is believed that comprehenders achieve this via a fundamental "binding" operation that, on some representational level, "binds together" the internal representations of entities among which a discourse relation holds (4). For example, a reader may bind their internal representation of the *table* in Figure 1 to that of *Australia* since the *manufactured in* relation holds between these two entities. Recent work has found evidence that Large Language Models (LLMs) are able to track entities across discourse and has started to uncover mechanisms supporting relational binding (5; 6; 7; 8; 9). However, this line of research has primarily focused on very short texts involving only a small number of entities, leaving discourse-level relational binding in LLMs largely unexplored. Here, we extend the scope of analysis towards discourse-level relational structures spanning multiple sentences and show that relational binding in LLMs can be understood in terms of what we call **Cell-based Binding Representation (CBR)**. A CBR consists of cells that are arranged in a more or less grid-like pattern in a linear subspace of activation space, with each cell corresponding to an entity-relation pair that the model decodes to its bound attribute. As we will show, a CBR abstracts discourse-level relational structure into discrete entity indices $ei$ and relation indices $ri$, enabling attributes to be represented as bound to specific $[ei, ri]$ pairs as shown in Figure 1 (a).

Furthermore, these indices are linearly decodable from model activations, revealing a low-dimensional and interpretable relational binding subspace organized along two dominant directions corresponding to entity indices $ei$ and relation indices $ri$. Finally, through causal interventions using activation patching, we demonstrate that manipulating activations within this subspace systematically changes relational predictions, providing evidence that LLMs actively use this cell-based representational mechanism to bind and retrieve relational information over discourse.

## 2 Identifying Cell-based Binding Representation (CBR) Subspace

**CBR Indices** We formalize entity and relation indices in CBR as Indexed Relational Scheme (IRS). Specifically,

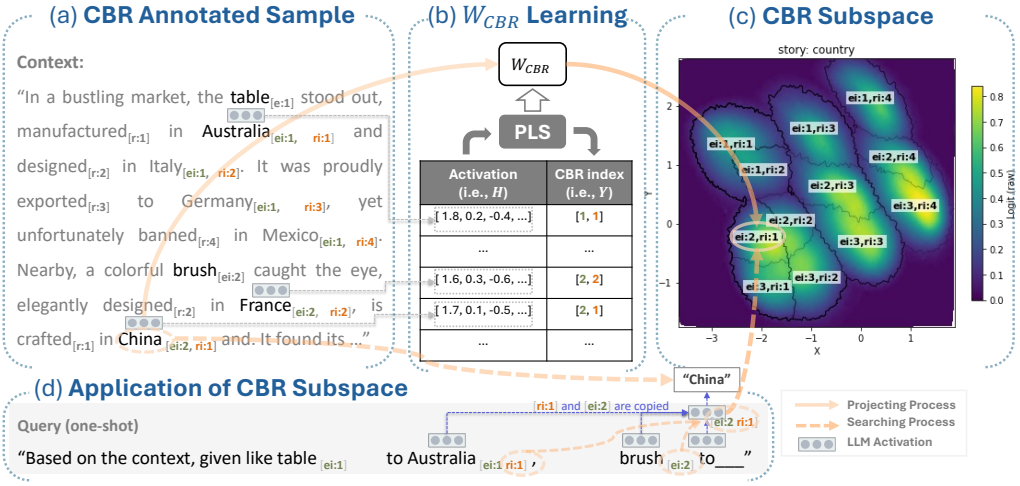



Figure 1: Overview of our Cell-based Binding Representation (CBR): (a) discourse annotated with entity and relation indices; (b) identifying the CBR subspace by predicting these indices from model activations via PLS; (c) visualization of the cells corresponding to each entity-relation index pair; and (d) cell–based retrieval, in which the model projects the query *brush* and hidden *manufactured in* onto the cell $[ei : 2, ri : 1]$ to retrieve the answer *China*.

an IRS abstracts discourse-level bindings into **entity indices** $ei$ and **relation indices** $ri$, corresponding to the order in which entities and relation types are introduced, and associates each attribute token with a specific index pair $[ei, ri]$. For example, in Figure 1 (a), *China* is represented as bound to $[ei{:}2,\ ri{:}1]$ and *Italy* to $[ei{:}1,\ ri{:}2]$.

**Data and models.** To test robustness across semantic domains and surface forms, we construct five discourse contexts of 1,000 samples each varying in entity, relation, and attribute inventories (e.g., countries, cities and occupations). Examples of CBR (or IRS) indices annotated samples are shown in Sample 4 and Sample 5 in Appendix (§A). Overall, our experiments cover five domains (i.e., country, city, job, relation and object) and two model families including Llama3-8B-Instruct and Qwen3-8B.

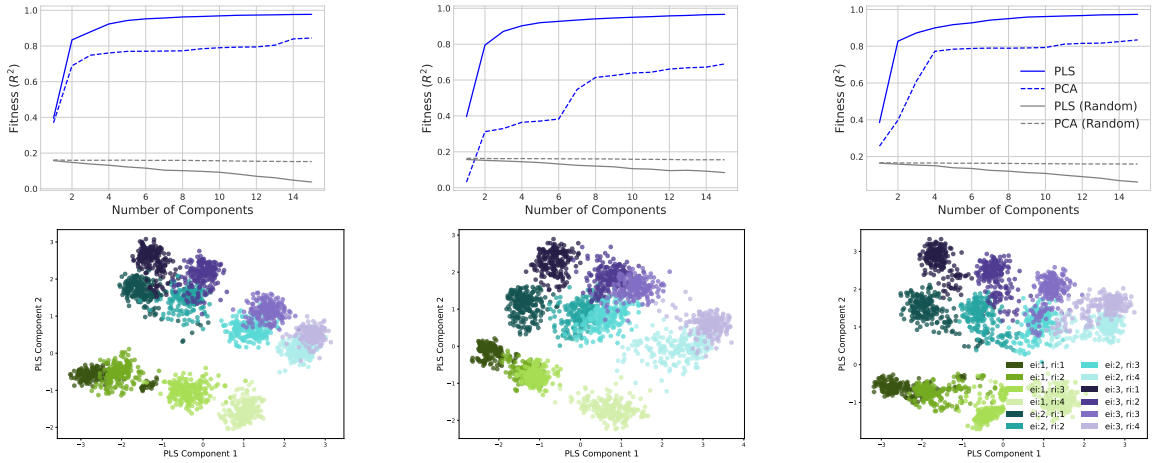
**Method.** To identify the CBR subspace, we fit a Partial Least Squares (PLS; 10) regression model to learn projection matrix $W_{\text{CBR}}$ that maps activations of attributes (e.g., "France") onto entity and relation indices $ei$ and $ri$ (e.g., $[2, 2]$), as shown in Figure 1 (b).

**Results.** We vary the number of components of PLS models and evaluate using goodness of fit. As shown for Llama3-8B-Instruct across three domains in Figure 2 (top), PLS models achieve near-perfect fits with a small number of components, i.e., both entity and relation indices can be linearly decoded from low-dimensional subspace of activation space. Projecting the activations of attribute tokens (e.g., "Australia" in Figure 1 (a)) onto the top two PLS components, we obtain the visualization shown in Figure 2 (bottom) for Llama3-8B-Instruct. Qwen3-8B exhibits similar patterns. The plots reveal two dominant and interpretable directions: one that separates the points by $ei$ and another that separates them by $ri$. Attributes associated with the same entity cluster along the $ei$ direction, while those participating in the same relation align along the $ri$ direction, supporting our hypothesis of a CBR subspace that jointly represents entity and relation indices.

**Stability of CBR Subspace** To examine the stability of the CBR subspace, we evaluate how it behaves under controlled perturbations to the discourse while keeping the underlying index information unchanged. As shown in Figure 6 (Appendix A), the geometric structure of the distribution remains essentially unchanged, and high predictive performance ($R^2$ is around 0.8) of $W_{CBR}$, indicating that the CBR subspace is robust to the perturbations.

**Generality of $W_{\text{CBR}}$** To test whether a $W_{\text{CBR}}$ learned from one context can effectively recover index information in another, we show the cross-context $R^2$ scores in Figure 6 (2) in Appendix A, indicating that the $R^2$ scores typically fall between 0.45 and 0.8, indicating a moderate level of cross-context generality.

**Semantic Information in CBR** To test whether the CBR subspace encode semantics or is it merely positional, we conduct the analysis shown in Figure 7 in Appendix (§A), which indicates that the CBR subspace also captures semantic information.




(1) Domain: *country* (2) Domain: *job* (3) Domain: *city*

Figure 2: **Top:** PLS goodness-of-fit when predicting entity and relation indices $[ei, ri]$ from Llama3-8B-Instruct attribute activations across three domains. For comparison, we also fit a Principal Component Analysis regression (PCA) and include random-label controls. **Bottom:** Visualization of attribute activations projected onto the top two PLS components, showing a grid-like separation by entity index ($ei$) and relation index ($ri$).

## 3 Causal Effect of the CBR

**Effect on attribute prediction.** To understand if and how models use the CBR subspace for relational binding, we perform causal interventions via activation patching (11; 12; 13; 14; 15; 16; 17), using a one-shot query setting as shown below (1). We opt for one-shot queries since alternatives such as providing preceding context (2) and using a direct query (3) would allow the model to rely on superficial cues and alternative strategies such as context matching via induction heads (18).

(1) **Query** (one-shot): Based on the context, given like Sean to Perm, Jose to?

(2) **Context**: Sean, who hails from Phoenix, currently resides in Perm. ... Meanwhile, Jose was born in Austin and is now living in Berlin. ...

(3) **Query** (direct): Based on the context, Jose is now living in?

We now causally intervene on the CBR subspace by patching activations along the top two PLS directions, which we hypothesize to encode entity and relation indices.

Concretely, we uniformly sample $10^4$ points (denoted $p_j$) from the range defined by the minimum and maximum values of the learned CBR subspace obtained through the projection matrix $W_{\text{CBR}}$. The embedding of each point is denoted as $\mathbf{h}_{p_j}$. For each attribute instance in 50 randomly selected samples (e.g., "Berlin" in Sample 2), we gradually move its activation (denoted as $\mathbf{h}_{ei,ri}$) towards one of the sampled target points according to Equation 2, where $\alpha$ is a hyperparameter and $\mathbf{h}^*_{ei,ri}$ denotes the updated activation of an attribute, effectively sweeping the activation of attribute across the CBR plane. At each step, we compute the logit score of the corresponding attribute predicted by the LLM.

For instance, given a Context 2 and a Query 1 for "Berlin", we patch its activation in the Context and observe the logit score of the corresponding attribute "Berlin". The resulting logit landscape is shown in Figure 3.

$$\mathbf{s}_{ei,ri\to p_j} = \mathbf{h}_{p_j} - W_{\text{CBR}}\mathbf{h}_{ei,ri}, \quad (1)$$

$$\mathbf{h}^*_{ei,ri} = \mathbf{h}_{ei,ri} + \alpha W^T_{\text{CBR}}\mathbf{s}_{ei,ri\to p_j} \quad (2)$$

The logit landscape shows that the CBR subspace is partitioned into "cells" arranged in a grid-like pattern, with each cell corresponding to a specific entity–relation index pair $[ei, ri]$. Within each cell, the attribute bound to that particular index achieves the highest logit score, and the logit value decreases smoothly as the patched activation moves away from center of the cell.

**Control: Intervention along random directions.** To test whether LLMs rely on the CBR subspace for relational binding, we also perturb attribute activations along the CBR directions using Equation 4 and compare the re-

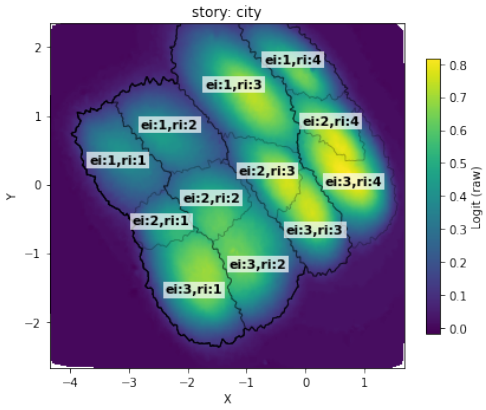



Figure 3: Logit landscape of attribute predictions resulting from causal interventions in the CBR subspace on Domain:*city*. Each "cell" corresponds to an entity–relation index pair $[ei, ri]$, with boundaries marking where the predicted attribute switches. See detailed explanation in §3.

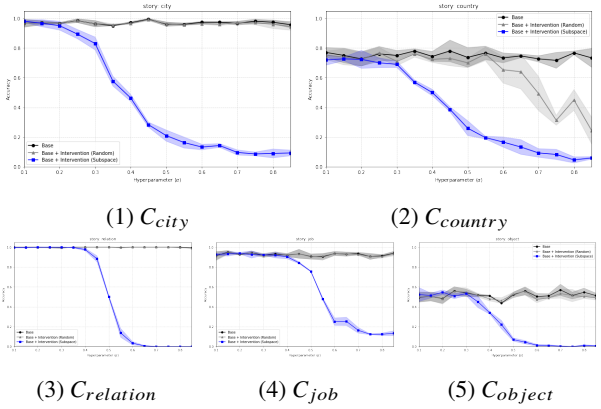



Figure 4: Effect of perturbing activations along the CBR subspace versus a random subspace on Llama3-8B-Instruct. Qwen3-8B exhibits similarly.

sults with perturbations along a random subspace defined by a randomly generated projection matrix in Equation 3. We use the same one-shot query setting to query each attribute in the context.

$$\mathbf{h}^{*}_{ei,ri} = \mathbf{h}_{ei,ri} + \alpha W^{T}_{\text{rand}}(W_{\text{CBR}}\mathbf{h}_{ei,ri}), \quad (3)$$

$$\mathbf{h}^{*}_{ei,ri} = \mathbf{h}_{ei,ri} + \alpha W^{T}_{\text{CBR}}(W_{\text{CBR}}\mathbf{h}_{ei,ri}) \quad (4)$$

If the LLM indeed uses the CBR subspace to make predictions, perturbing activations along this subspace should degrade performance. Figure 4 shows the results as a function of perturbation strength. We observe that as the perturbation weight increases, the accuracy of attribute predictions decreases significantly. These results prove that LLMs utilize the CBR subspace when predicting attributes: disruptions in this subspace directly impair model performance, whereas unrelated directions do not.

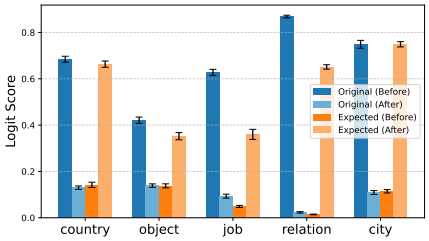



Figure 5: Activation patching for steering CBR indices on Llama3-8B-Instruct. Qwen3-8B behaves similarly.

**CBR Subspace based Mechanism** To understand how LLMs use the CBR subspace to retrieve the correct attribute given a context (e.g., Sample 2) and a query (e.g., Query 1), we propose a high-level mechanism illustrated in Figure 1 (d). When answering a one-shot relational query, the model appears to perform two parallel operations: (i) it extracts relation index information (e.g., $ri : 2$) from the attribute exemplar provided in the one-shot part, and (ii) it extracts entity index information (e.g., $ei : 2$) from the query entity itself. These two indices together define a point corresponding to a cell in the CBR subspace, which is then used to retrieve the answer from the context.

To test this mechanism, we perform several Activation Patching (AP) interventions that steer model activations along specific CBR subspace directions. One such intervention is Relation-index steering, which shifts $ri$ in the one-shot attribute from $ri : j$ to $ri : j + 1$, as illustrated in Figure 8 in Appendix (§A).

Following (14), we evaluate the effect of steering by measuring the change in logit scores for both the original correct answer and the expected answer after intervention. As shown in Figure 5, steering along the CBR subspace direction consistently suppresses the logit of the original answer and increases the logit of the expected answer, precisely in line with the intended index manipulation.

## 4 Conclusion

In this work, we investigated how LLMs internally represent relational binding in discourse. To this end, we proposed the Cell-based Binding Representations (CBR) framework and applied PLS to identify a low-dimensional subspace that linearly encodes entity and relation indices. Causal interventions show that manipulating the CBR subspace reliably alters LLMs predictions, supporting a CBR subspace based mechanism in LLMs.

## Acknowledgements

This work was supported by JST CREST Grant Number JPMJCR20D2, JSPS KAKENHI Grant Number 21K17814 and Japan Science and Technology Agency under Grant No. JST BOOST JPMJBY24F9.

## References

[1]Bonnie Lynn Webber, editor. **A Formal Approach to Discourse Anaphora**. Routledge, London, 1 edition, 1979.

[2]Teun Adrianus Van Dijk, Walter Kintsch, et al. Strategies of discourse comprehension. 1983.

[3]Rolf A Zwaan and Gabriel A Radvansky. Situation models in language comprehension and memory. **Psychological bulletin**, Vol. 123, No. 2, p. 162, 1998.

[4]Anne Treisman. The binding problem. **Current opinion in neurobiology**, Vol. 6, No. 2, pp. 171–178, 1996.

[5]Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? **arXiv preprint arXiv:2310.17191**, 2023.

[6]Najoung Kim and Sebastian Schuster. Entity tracking in language models. **arXiv preprint arXiv:2305.02363**, 2023.

[7]Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring latent world states in language models with propositional probes. **arXiv preprint arXiv:2406.19501**, 2024.

[8]Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Representational analysis of binding in language models. **arXiv preprint arXiv:2409.05448**, 2024.

[9]Yoav Gur-Arieh, Mor Geva, and Atticus Geiger. Mixing mechanisms: How language models retrieve bound entities in-context. **arXiv preprint arXiv:2510.06182**, 2025.

[10]Svante Wold, Michael Sjöström, and Lennart Eriksson. Pls-regression: a basic tool of chemometrics chemometr. **Intell. Lab**, Vol. 58, No. 2, pp. 109–130, 2001.

[11]Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. **Advances in neural information processing systems**, Vol. 33, pp. 12388–12401, 2020.

[12]Atticus Geiger, Kyle Richardson, and Christopher Potts. Neural natural language inference models partially embed theories of lexical entailment and negation. **arXiv preprint arXiv:2004.14623**, 2020.

[13]Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. **Advances in Neural Information Processing Systems**, Vol. 34, pp. 9574–9586, 2021.

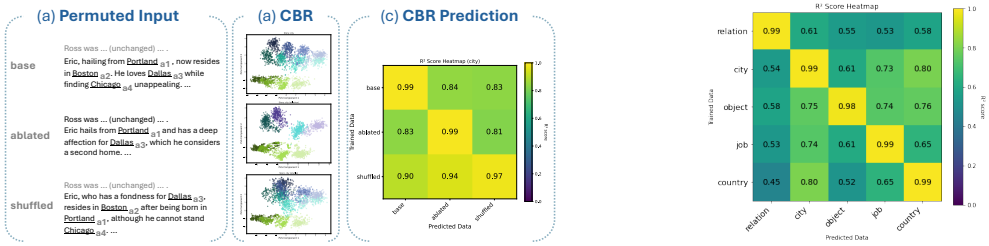
[14]Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. **arXiv preprint arXiv:2211.00593**, 2022.

[15]Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. **arXiv preprint arXiv:2305.15054**, 2023.

[16]Benjamin Heinzerling and Kentaro Inui. Monotonic representation of numeric properties in language models. **arXiv preprint arXiv:2403.10381**, 2024.

[17]Michael Hanna, Ollie Liu, and Alexandre Variengien. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. **Advances in Neural Information Processing Systems**, Vol. 36, , 2024.

[18]Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. **arXiv preprint arXiv:2209.11895**, 2022.




(1) Stability (2) Generality

Figure 6: (1) (a) Samples of Ablated and Shuffled Dataset from $C_{city}$. The ablated sample removes the attribute a2 (i.e., "Boston") and a4 (i.e., "Chicago"), while the shuffled sample alters the order of a1 (i.e., "Portland") and a3 (i.e., "Dallas"). (1) (b) Visualization of the CBR subspace before and after ablating (or shuffling) attributes and corresponding relations. The essential geometric structure remains unchanged, indicating that removing (or shuffling) relational content that does not affect entity and relation indices preserves the underlying CBR subspace. (1) (c) Cross-dataset $R^2$ scores for index prediction using projection matrices learned from the original, ablated, and shuffled datasets on Llama3-8B-Instruct. All scores remain high (around 0.8), demonstrating strong consistency and transferability of the learned CBR projection across these perturbed datasets. In (2) $R^2$ from Llama3-8B-Instruct, indicating cross-context generality, where $W_{CBR}$ learned from one context (column) is used to predict the indices of another (row).

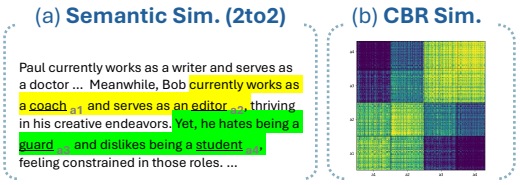



Figure 7: Semantic Similarity Pattern and CBR Subspace Similarity, where a1 and a2 share similar relation (i.e., "current job"), as do a3 and a4 (i.e., "disliked job"). The heatmap shows the subspace similarity among them.

**Semantic Information in CBR Subsapce** To examine whether the CBR subspace captures semantic information in addition to indices, we construct an additional dataset in which relations exhibit controlled patterns of semantic similarity. As shown in Figure 7 (a), the first and last two relations share similar meaning (denoted as 2to2). we project attribute (e.g., "coach" denoted as $a1$) activations into the CBR subspace via the original $W_{\text{CBR}}$ (§2) and compute pairwise cosine similarities among the projected vectors, shown in Figure 7 (b). The CBR-projected representations clearly reproduce the intended semantic similarity structure, indicating that CBR subspace embeds semantic information.


Figure 8: Causal Intervention.

**Causal Intervention** on the CBR subspace reveals the CBR subspace based mechanism. Steering different components of the CBR subspace produces systematic changes in model behavior. For example, manipulating the relation index in the one-shot attribute activation (e.g., shifting from $ri:1$ to $ri:2$ in the activation of "Australia") redirects the model toward predicting attributes associated with the intervened relation (e.g., changing the output from "China" to "France").

# A Appendix

**Stability and Generality of CBR Subspace** We illustrate the results for Generality and Stability analysis of CBR Subspace in Figure 6, indicating high stability and moderate cross-context(or domain) generality. **Stability and Generality of CBR Subspace** We analyze the semantic information encoded by CBR subspace in Figure 7. **Causal Intervention Method** Figure 8 explains the method of causal intervention on the CBR Subspace. **Additional Examples** Sample 4 and 5 are additional samples selected from Domain: *job* and Domain: *relation* respectively.

(4) Sean is currently a guard$_{[ei:1,ri:1]}$, aspiring to be a builder$_{[ei:1,ri:2]}$ someday. He once held the position of a student$_{[ei:1,ri:3]}$ but found no joy in being a chef$_{[ei:1,ri:4]}$. Luke, on the other hand, works as a coach$_{[ei:2,ri:1]}$ and dreams of becoming an actor$_{[ei:2,ri:2]}$. His previous role was as a judge$_{[ei:2,ri:3]}$, yet he has no fondness for being an artist$_{[ei:2,ri:4]}$. Lastly, Sam is a teacher$_{[ei:3,ri:1]}$ who hopes to transition into a writer$_{[ei:3,ri:2]}$. Before this, he worked as a drive$_{[ei:3,ri:3]}$r, and he particularly disliked being a manager$_{[ei:3,ri:4]}$. Each of them navigates their careers, ... (from Domain: *job*)

(5) Roy, happily married to Ella$_{[ei:1,ri:1]}$, is the proud father of Ed$_{[ei:1,ri:2]}$. He learned from Jim$_{[ei:1,ri:3]}$ and reports to Ross$_{[ei:1,ri:4]}$ at work. Meanwhile, Mike shares his life with Kate$_{[ei:2,ri:1]}$ and they have a son named Rob$_{[ei:2,ri:2]}$. Under the guidance of Mark$_{[ei:2,ri:3]}$, Mike navigates his career under the supervision of Dave$_{[ei:2,ri:4]}$. Lastly, Jay is devoted to Lila$_{[ei:3,ri:1]}$ and they have a child$_{[ei:3,ri:2]}$, Luke. He was educated by James$_{[ei:3,ri:3]}$ and works under Leo$_{[ei:3,ri:4]}$. Together, these families weave a tapestry of connections ... (from Domain: *relation*)