VISUAL DEPENDENCE AND UTILITY IN MULTIMODAL KNOWLEDGE GRAPH COMPLETION: A CONTROLLED PILOT STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal knowledge graph completion experiments often use changes in aggregate ranking metrics to assess the value of visual information. Interpreting these changes requires specifying which inputs, model parameters, and evaluation labels remain fixed. We report a controlled pilot study that examines these choices through three diagnostics: alpha-channel image preprocessing, positive-label filtering with fixed scores, and visual-package reassignment in separately trained MyGO models. Correcting transparent-image handling changes hundreds of source images, while absolute mean new-target MRR changes remain below 7×10^{-5} in our fixed linear-transfer setup under its original validation filter. Adding withheld training positives to the evaluation filter changes MRR without changing model scores and reverses one descriptive fusion comparison. In a separate transductive experiment, five new visual mappings reduce the original-image model's MRR by 0.0544–0.0587. The permanently permuted model changes by less than 0.0006 in either direction, despite more than 58% of query ranks changing under every mapping. A zero-visual control remains exactly invariant. These observations distinguish input dependence, net ranking utility, and evaluation sensitivity within the tested configurations. The experiments are exploratory: the MyGO comparison uses one training seed and 100 epochs, and the evaluation data have informed development. The study motivates reporting query-level responses alongside aggregate metrics and identifying the intervention that supports each claim.

1 INTRODUCTION

Knowledge graph completion ranks candidate entities for incomplete relational facts. Multimodal models additionally associate entities with images or text, introducing several routes through which side information can affect a ranking. An image can provide content useful for a relation, distinguish entities, change optimization, or alter the representation geometry used by the decoder. Evaluation also depends on which known positive answers are excluded from the candidate competition. A small change in a final mean reciprocal rank (MRR) therefore leaves several explanations open.

Existing work already provides strong foundations for examining these explanations. Randomized side information and retraining have been used to test the value of numerical literals in knowledge graph prediction (Blum et al., 2024). Remove-and-retrain analyses motivate retraining-based controls for perturbation-induced distribution shift (Hooker et al., 2019). Knowledge graph evaluation under incomplete positive labels is an established concern (Yang et al., 2022; Zhou et al., 2022). Our study applies these ideas to a concrete multimodal experimental pipeline and reports the resulting scope limitations explicitly.

We investigate three questions. First, how much does an identified image-decoding problem change downstream predictions after its inputs are repaired? Second, how much can a ranking metric change when additional positive labels are used only by the evaluator? Third, can a small aggregate response to visual reassignment coexist with substantial query-level sensitivity? These questions concern different interventions and use two experimental setups. The first two examine frozen-feature transfer

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

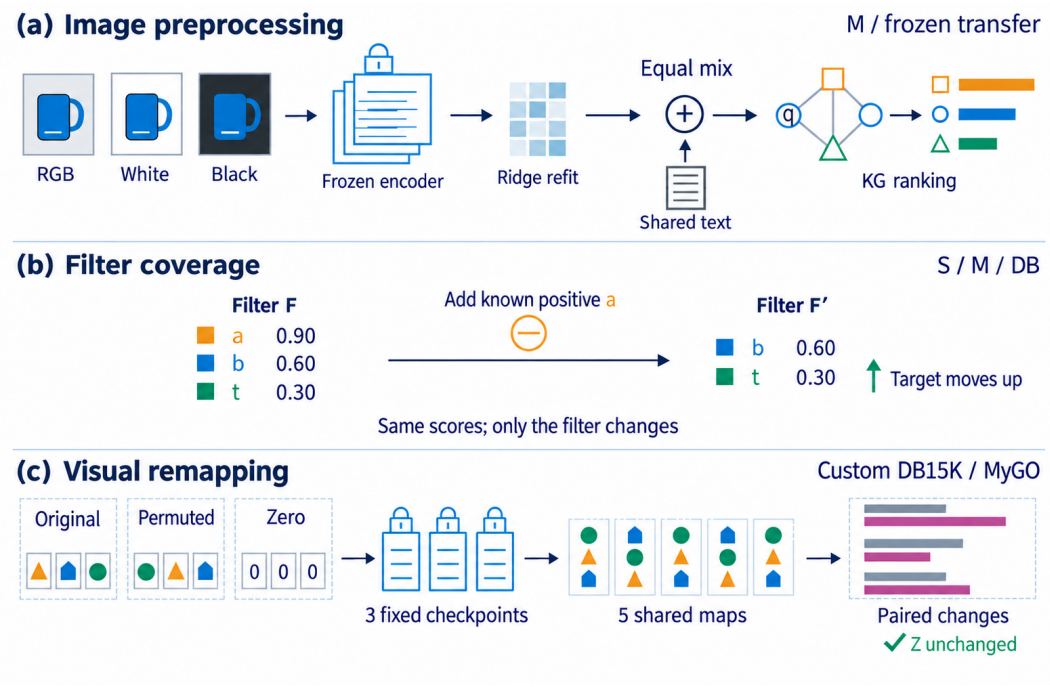

Figure 1: Three interventions at distinct stages of the experimental pipeline. A: on M, background treatments feed frozen visual encoders, followed by base-entity visual ridge refitting and equal-weight fusion with shared text predictions. Only unseen entity vectors are replaced; base entity and relation vectors remain fixed. B: on S, M, and DB, identical scores yield rankings under F (base–base training and eligible validation positives) or F' (also including withheld training positives). C: a separate custom transductive DB15K split supplies original-input, permanently permuted, and zero-visual MyGO checkpoints (Zhang et al., 2025), evaluated with five shared new mappings. Reordered tiles illustrate visual-package reassignment at fixed entity slots. Each row retains its own protocol. Pictures, matrices, response marks, and numerical scores in B are illustrative; they contain no experimental measurements or real sample images.

to structurally unseen entities. The third examines a transductive MyGO model (Zhang et al., 2025) in which all entity identifiers have trainable representations.

Our principal observation concerns the third question. In a permanently permuted visual model, the net MRR response to five further visual mappings is close to zero, while each mapping changes the ranks of over 58% of validation queries. The original-image model has a consistently larger adverse response to the same mappings. These measurements establish conditional sensitivity patterns of the saved models. Retraining variability, optimization progress, and distribution shift remain unresolved explanations of the difference. The image and filtering diagnostics describe additional reasons to keep experimental interventions separate when interpreting aggregate metrics.

Figure 1 locates each intervention relative to learning, scoring, and evaluation. The detailed MyGO computation and control flow appears in Figure 3 in the appendix. The model components are inherited from MyGO; the reassignment protocol defines the diagnostic intervention.

This paper is a pilot empirical report. It provides controlled measurements, executable checks, and explicit limits on their interpretation. The results support neither a state-of-the-art performance claim nor a general conclusion about whether visual semantics are useful across knowledge graphs. Establishing a distinct scientific contribution beyond the cited control methodology requires further evidence.

Related multimodal approaches. NativE uses adaptive fusion and adversarial generation to address heterogeneous and missing modalities (Zhang et al., 2024). SNAG incorporates noise into multimodal representation learning (Chen et al., 2025). The publisher abstract of SAFKGC de-

scribes confidence-aware suppression of irrelevant multimodal information and a consistency objective (Wang et al., 2026); our comparison with that work is limited to its available abstract and introductory description. These approaches provide direct context for modality reliability and noise. The present pilot estimates responses to fixed input assignments and training controls; it does not demonstrate superiority over these methods or establish an exhaustive absence of prior visual-randomization studies.

2 INTERVENTIONS AND MEASURED QUANTITIES

Let $s_\theta(q, e; x)$ be a model's score for query q, candidate entity e, and entity-associated visual packages x. Let t_q be the target and F_q a set of other positive targets excluded from evaluation. We follow standard positive-filtered evaluation (Bordes et al., 2013), using average ranks for ties:

$$r_F(q) = 1 + \sum_{e \neq t_q,\ e \notin F_q} \left[\mathbf{1}\{s_e > s_{t_q}\} + \tfrac{1}{2}\mathbf{1}\{s_e = s_{t_q}\}\right], \qquad \mathrm{MRR}_F = \frac{1}{|Q|}\sum_{q \in Q} \frac{1}{r_F(q)}. \tag{1}$$

All candidates in the corresponding entity vocabulary participate before filtering. We distinguish the following measurements.

Input sensitivity at fixed parameters. A visual mapping π assigns visual packages to entities and gives query-level differences $d_q(\pi) = \mathrm{RR}(q; \theta, \pi x, F) - \mathrm{RR}(q; \theta, x, F)$. We report the signed mean $\bar{d}$, mean absolute change $a = |Q|^{-1}\sum_q |d_q|$, rank-change frequency, and top-ranked-candidate change frequency. When $a > 0$, the cancellation fraction is

$$c = 1 - \frac{|\bar{d}|}{a}. \tag{2}$$

This is a descriptive arithmetic decomposition. A value near one means that positive and negative reciprocal-rank changes largely cancel in their mean. It neither identifies a latent mechanism nor supplies an equivalence test. Rank changes far down the list can have small reciprocal-rank effects, so rank-change frequency needs the accompanying magnitude statistics.

Utility under retraining. Comparing models trained with original, permanently permuted, or zeroed visual inputs changes both the training information and fitted parameters. Such a comparison measures the effect of the complete training intervention at its specified budget. It cannot isolate semantic information from identity cues, optimization, or compensating use of other modalities. Fixed-parameter reassignment also changes the input distribution relative to training; this limitation is central to remove-and-retrain work (Hooker et al., 2019).

Sensitivity to positive-label coverage. For fixed scores and nested filters $F_q \subseteq F'_q$, Equation 1 gives the exact identity

$$r_F(q) - r_{F'}(q) = \sum_{e \in F'_q \setminus F_q} \left[\mathbf{1}\{s_e > s_{t_q}\} + \tfrac{1}{2}\mathbf{1}\{s_e = s_{t_q}\}\right] \geq 0. \tag{3}$$

Thus, enlarging this filter mechanically weakly increases each query's reciprocal rank. The magnitude and the difference between two models remain empirical questions. We use the identity as an implementation check. It is an elementary property of filtered ranking, consistent with prior evaluation studies (Yang et al., 2022; Zhou et al., 2022).

3 EXPERIMENTAL SETUP

3.1 FROZEN-FEATURE TRANSFER WITH CONTROLLED ENTITY ARRIVAL

We use multimodal CoDEx-S and CoDEx-M resources, denoted S and M, and a cleaned DB15K graph, denoted DB. CoDEx provides the underlying knowledge graph and entity descriptions (Safavi & Koutra, 2020); the multimodal extension provides the image-associated CoDEx resources (Zhu et al., 2025). S and M overlap in all 1,383 S entities; M is a scale-related comparison from the same

Table 1: Frozen-transfer validation protocols. Training counts refer to visible base–base facts. S and M share entities and graph provenance. Each fact yields head and tail queries.

Graph	Entities	Relations	Train	New target	New context	Base only
S	1,383	40	9,700	124	116	1,008
M	7,697	49	37,363	281	271	4,122
DB	12,842	279	42,057	2,280	1,924	10,516

source. DB provides a separate graph source. Our splits simulate structurally unseen entities in a static graph. They do not contain observation timestamps or establish real-world temporal evolution.

Entities are assigned to base, validation-newcomer, and test-newcomer groups. Training uses facts with two base endpoints. Validation excludes facts incident to test newcomers. We report three roles: a newcomer target, a base target with a newcomer query context, and two base endpoints. The full entity vocabulary remains the candidate universe, including entities held out from structural training. All unseen entity vectors are predicted from side information. Table 1 gives the validation scope used in this report.

ComplEx (Trouillon et al., 2016) and DistMult (Yang et al., 2015) provide structural teacher representations. Each teacher is trained with seeds 20–24 for 100 epochs; base-only validation ranking among base candidates selects the checkpoint. Separate ridge regressions map frozen text and visual features to teacher entity vectors using base entities. The ridge penalty is 2, with an unpenalized intercept, float64 fitting, and float32 prediction. Base entity vectors remain unchanged. We compare text-only prediction with equal-weight text/visual fusion; missing images use the text prediction. The scale is fixed at one.

For S and M, text features use the 384-dimensional `all-MiniLM-L6-v2` checkpoint in the Sentence Transformers framework (Reimers & Gurevych, 2019),[1] and CLIP ViT-B/32 visual features have 512 dimensions (Radford et al., 2021). M additionally uses 4,096-dimensional VGG16 fc7 features (Simonyan & Zisserman, 2015). DB uses existing 768-dimensional text and 4,096-dimensional VGG feature resources. The exact originating DB text checkpoint and the legacy image encoder's version, layer, and preprocessing configuration are unresolved; this restricts end-to-end reproduction from raw inputs. Neither feature-source differences nor S/M dependence permit attributing a cross-graph difference to graph size alone.

3.2 TRANSDUCTIVE MYGO PILOT

The MyGO experiment uses a custom split of 89,197 distinct DB15K facts: 71,357 training, 8,920 validation, and 8,920 held-out test facts. It has 12,842 entities and 279 relations. Every entity identifier is available to the transductive model. Evaluation uses 17,840 directed validation queries and all 12,842 candidates, filtered by training and validation positives. This protocol differs from the frozen-transfer DB experiment.

We use the official MyGO implementation at revision `d2f6a7a7fc79`, with a compacted frozen text-token dictionary that preserves every accessed embedding value. Three arms use original visual packages (O), a fixed package permutation throughout training (P), or zeroed visual embeddings (Z). Training uses seed 2024 and 100 epochs per arm. The original implementation's 1,500-epoch budget provides context for this short pilot; all three selected checkpoints occur at epoch 100. Their comparisons therefore lack evidence of convergence. The original arm used two CPU threads during training and the controls used one; all new evaluations use one. We preserve this execution difference as a limitation.

Five new mappings, with seeds 271832–271836, are fixed before the new evaluations and shared across the three saved checkpoints. Each mapping reassigns the original visual packages among 12,570 nonmissing entities within active-token-count strata. A single-cycle derangement is generated in each eligible stratum; the procedure does not sample uniformly from all possible derangements. The 272 missing-image entities stay fixed. Package reassignment changes both query-side and candidate-side representations. The zero-input hook is reinstalled on loading Z and zeroes the

[1] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

Table 2: M newcomer-target MRR under the original filter, averaged across five teacher seeds (281 queries per seed). Background variants are all retained. Text-only means are 0.007246 (ComplEx) and 0.007117 (DistMult). These values describe one fixed readout configuration.

Teacher	Visual encoder	Old RGB	White	Black
ComplEx	CLIP	0.002887	0.002898	0.002893
DistMult	CLIP	0.002847	0.002870	0.002843
ComplEx	VGG	0.002514	0.002578	0.002531
DistMult	VGG	0.002438	0.002506	0.002456

output of the visual-token lookup; projection biases, positional embeddings, visual slots, and masks remain. Additional implementation details appear in Appendix B.

3.3 Scope and statistical units

The frozen-transfer comparisons have five training seeds on shared graphs and queries. We report sample standard deviations and paired effects. The specified alpha-repair family contains 72 metric comparisons; the filter family contains 66 MRR comparisons. Exact two-sided sign-flip tests enumerate 32 seed sign assignments, giving a minimum attainable p-value of 0.0625, before Holm correction. These small-sample tests cannot establish equivalence. MyGO's five mappings describe inference-input variation conditional on one checkpoint per training arm; they are not five independent training runs. All diagnostics use previously examined validation resources. New mapping seeds provide fresh interventions on existing checkpoints, without independent data confirmation. The workers for the new filtering and reassignment diagnostics do not decode test triples.

4 Results

4.1 Repairing alpha-channel preprocessing

Our original image loader converted files directly to RGB, discarding transparency. Among 7,697 M images, 850 had an alpha channel and 706 contained nonopaque pixels. Compositing on white changes 682 decoded RGB images; compositing on black changes 617. Seventy images that were constant after the old conversion become nonconstant under white compositing. This is a defect in our preprocessing path. We make no attribution of this behavior to a benchmark author's encoder.

Both background choices are retained. White compositing changes 674 CLIP and 431 VGG embedding rows; black changes 607 and 355. Source pixel changes and embedding changes differ because subsequent image transforms can remove or collapse changed content. Exact duplicate-embedding entity counts for CLIP are 152, 91, and 158 under old, white, and black inputs; the corresponding VGG counts are 165, 95, and 165. Every exact-embedding duplicate group inspected in the old-input audit shares the actual transformed input pixels. These observations do not establish semantic encoder collisions.

We refit the same frozen-transfer readouts for all six visual settings, yielding 60 fusion conditions and ten shared text conditions. Under the original validation filter, the paired mean MRR changes for newcomer targets range from -0.00000407 to $+0.00006796$. The largest absolute mean change over all roles is 0.00070148 for a new-context condition. All 60 fusion conditions remain below the paired text-only model on newcomer-target MRR. Table 2 shows the newcomer-target means; full role-level results remain part of the experiment record. Repairing the source handling has a measurable representation effect, while the fixed downstream configuration remains weak on these targets.

4.2 Changing positive-label coverage with scores held fixed

We compare two evaluator filters. F contains visible base–base training facts and eligible validation facts. F' adds the original training facts whose structure was withheld from the learner. These labels affect only evaluation. The original and enlarged filters use the same queries, candidate vocabulary, entity representations, and unfiltered score matrices. Both retain the evaluated target. Conventional

Table 3: Mean newcomer-target MRR over five teacher seeds. Each row uses identical scores under F and F'. Fusion uses CLIP for S/M and VGG for DB. The full grid, including both M encoders and backgrounds, appears in Appendix C.

Graph	Teacher	Text F	Text F'	Fusion F	Fusion F'
S	ComplEx	.050413	.101264	.038058	.069141
S	DistMult	.052788	.107902	.037744	.072348
M	ComplEx	.007246	.014352	.002887	.007579
M	DistMult	.007117	.015928	.002847	.007524
DB	ComplEx	.000705	.000867	.000312	.000326
DB	DistMult	.000680	.000838	.000300	.000314

Table 4: MyGO validation results. Baseline uses each checkpoint's training input. The remaining columns report mean $\pm$ sample SD across five fixed new mappings: signed and absolute reciprocal-rank changes are multiplied by 10^3; cancellation, changed target ranks, and changed top-ranked entities are percentages. Each arm has one training seed. Cancellation is undefined for Z because all changes are zero.

Arm	Baseline MRR	$10^3\bar{d}$	$10^3 a$	Cancellation	Rank changed	Top-1 changed
O	.275560	-56.273 ± 1.608	100.278 ± 2.158	43.89 ± 0.85	83.18 ± 0.35	59.05 ± 2.40
P	.270705	-0.198 ± 0.354	8.482 ± 0.227	95.81 ± 1.58	58.84 ± 0.29	6.41 ± 0.36
Z	.271193	0 ± 0	0 ± 0	—	0 ± 0	0 ± 0

filtered evaluation can legitimately use evaluator-only positives (Bordes et al., 2013); access to those labels by a learner would be a separate information-flow question.

The full grid contains 110 prediction states: 20 for S, 70 for M, and 20 for DB. Each is evaluated under both filters in three roles, producing 660 metric rows. Equation 3 holds exactly for every query and state; all 70 M rank arrays under the original filter exactly reproduce the prior alpha-sensitivity experiment. Table 3 displays the newcomer-target comparison for the common old-input configurations. The resulting metric changes occur with identical model scores.

The mean fusion-minus-text contrast reverses sign in one examined group: S/DistMult with a base target and newcomer context. Its value changes from $+0.008472$ under F to -0.005433 under F', with a positive-to-negative change in four of five seed-level contrasts. We report this descriptively; it was found within a grid of role-level comparisons. Every newcomer-target fusion mean remains below its text counterpart under both filters. Adding positive labels improves the absolute metric mechanically and does not establish a learned performance gain.

Query composition also needs direct verification. In DB, the target *Island* accounts for 733 of the 2,280 newcomer-target queries (32.15%). All 733 ranks are unchanged by this filter intervention in each of the 20 DB prediction states. This high-frequency target contributes zero to the observed filter-induced MRR change. Frequency concentration alone would therefore give an incorrect explanation of this particular effect. The checkpoint-selection procedure is unaffected by these added positives because it used base-only queries and base candidates; this fact does not validate any separate downstream hyperparameter-selection procedure.

4.3 Visual reassignment and cancellation

The original-input, permanently permuted, and zero-visual checkpoints obtain validation MRRs of 0.275560, 0.270705, and 0.271193, respectively, on their own training mappings (Table 4). These short-budget training outcomes are close in aggregate. The new inference interventions reveal different patterns of conditional dependence.

For P, the five signed MRR changes range from -0.000552 to $+0.000392$, with mean -0.000198 and sample SD 0.000354. Nevertheless, 58.84% of target ranks and 6.41% of filtered top-ranked entities change on average. Mean absolute reciprocal-rank change is 0.008482 ± 0.000227. Across the five mappings, the positive and negative reciprocal-rank contributions average $+0.004142$ and -0.004340 per query, respectively. The mean per-mapping cancellation fraction is 95.81%. Thus the small signed aggregate combines opposing query responses. Changes with $|d_q| \geq 0.01$ occur in

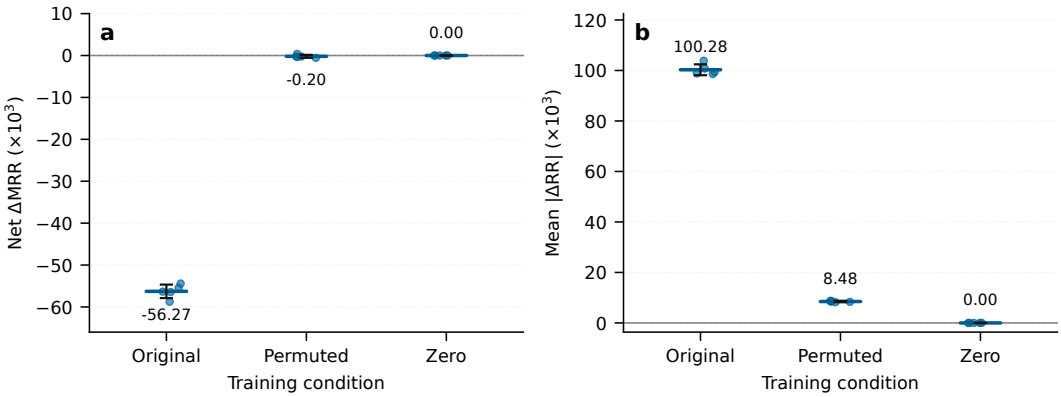

Figure 2: Net and absolute query response to visual reassignment in fixed MyGO checkpoints. Each point is one of five shared mappings; horizontal marks show means and error bars show sample SD across those mappings. Both vertical axes multiply reciprocal-rank quantities by 10^3. These error bars describe mapping variation conditional on the saved model and validation queries.

7.61–8.26% of queries, showing why the rank-change frequency alone overstates the prevalence of larger reciprocal-rank changes.

For O, every new mapping reduces MRR: the signed changes range from -0.058749 to -0.054415, with mean -0.056273 ± 0.001608. The mean absolute change is 0.100278 ± 0.002158, and 59.05% of top-ranked entities change. Z is exactly invariant under all five mappings, including hashes of all unfiltered scores and intermediate representations, target ranks, target scores, and top-ranked entities. Historical baseline ranks and repeated evaluations with the same mapping also agree exactly.

The training intervention and fixed-model reassignment estimate different quantities. O's adverse response can include distribution shift, and P's residual dependence can involve arbitrary entity-associated inputs, shared visual tokens, or training compensation. The results therefore leave semantic contribution and identity memorization unresolved. This interpretation follows the existing distinctions in randomized side-information controls and remove-and-retrain evaluation (Blum et al., 2024; Hooker et al., 2019). All observations remain conditional on the one-seed, 100-epoch pilot and its previously examined validation set; the five mappings provide no independent training or data replication.

5 Discussion and Limitations

What the combined evidence supports. The three diagnostics locate changes at different parts of an experimental pipeline: image decoding before feature extraction, evaluator labels after scoring, and visual inputs at fixed model parameters. Their numerical effects are conditional on their own protocols. In particular, the alpha study uses a linear transfer readout with unseen-entity structure withheld, while MyGO has trainable representations for every entity. Their results cannot be pooled into an estimate of a universal visual contribution.

Why the MyGO result remains unresolved mechanistically. The O checkpoint responds adversely to visual reassignment, and P exhibits opposing query-level changes with a small net mean. Visual semantics, arbitrary identity signatures, modality compensation, and short-budget optimization can all interact with these observations. The P checkpoint's response establishes input dependence within the tested perturbation family. It does not identify identity memorization. The O-to-P training comparison also changes the distribution that the model learns from. The relationship between perturbation sensitivity and retraining utility is already addressed in prior work on side-information controls and feature removal (Blum et al., 2024; Hooker et al., 2019). Our evidence adds a concrete visual-token pilot, with broader novelty still unestablished.

Remaining confirmation requirements. A stronger empirical claim would require multiple training seeds, matched execution conditions, sufficient optimization budgets, independent graph or split

confirmation, and a mechanism-specific intervention that separates the competing explanations. Five inference mappings cannot replace those training replications. New-target transfer additionally requires careful candidate and filter definitions, fully resolved feature provenance, and confirmation with a capable multimodal baseline in the same inductive protocol. Previous development examined these graphs repeatedly, so conventional test labels alone would not make a follow-up independent without a new analysis design.

Practical reporting implications. An experiment record should name the intervention, retain all predetermined control conditions, and report a query-level response measure alongside aggregate MRR. Positive-filter sources should be documented separately from learner-visible structure. Source-pixel repairs should retain both representation and downstream checks. These practices follow established diagnostic principles; the present observations illustrate their consequences in a specific multimodal knowledge graph workflow.

6 Conclusion

In the tested pipeline, repairing transparent-image handling, expanding evaluator positive labels, and reassigning visual packages produce distinct observable effects. The MyGO pilot shows that near-zero mean response can coexist with widespread query-level rank changes, while the zero-input control stays exactly invariant. The filtering experiment demonstrates score-invariant metric changes and one descriptive comparison reversal. These results provide a bounded empirical record and identify the additional evidence needed for a general account of visual utility in knowledge graph completion.

Reproducibility statement

Protocols, input and code hashes, per-seed metrics, and query-level ranks are retained in the local experiment archive. The filtering experiment checks its rank identity exhaustively and reproduces earlier M ranks exactly. The reassignment experiment checks historical baseline ranks, identical-map reruns, and zero-control invariance. Appendix B specifies the principal implementation choices, and Appendix D reports newcomer-target means and standard deviations. Summary and paired-comparison CSVs accompany the manuscript source package. A public anonymized implementation release has not been prepared for this working draft; exact legacy DB feature provenance remains a documented limitation.

AI use statement

OpenAI ChatGPT/Codex agents were used for literature retrieval and synthesis, hypothesis generation and refinement, experiment design, implementation, data preparation, mathematical descriptions, numerical analysis, figure generation, and manuscript drafting and editing. The two schematics were redrawn with reference-conditioned image generation; agent checks compared their labels and topology with the implementation and protocol. Separate agent passes checked saved artifacts, claims, and cited sources. These checks are automated assistance and do not establish human author verification. This working draft has not yet completed a full human-author review. The reported experimental measurements were computed by executable programs; generated explanatory text was not used as ground-truth experimental evidence. Human authors remain responsible for reviewing and approving any submitted version.

References

Ivana Balazevic, Carl Allen, and Timothy Hospedales. TuckER: Tensor factorization for knowledge graph completion. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5185–5194, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1522. URL https://aclanthology.org/D19-1522/.

Moritz Blum, Basil Ell, Hannes Ill, and Philipp Cimiano. *Numerical Literals in Link Prediction: A Critical Examination of Models and Datasets*, pp. 23–46. Springer Nature Switzerland, November 2024. ISBN 9783031778445. doi: 10.1007/978-3-031-77844-5_2. URL http://dx.doi.org/10.1007/978-3-031-77844-5_2.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf.

Zhuo Chen, Yin Fang, Yichi Zhang, Lingbing Guo, Jiaoyan Chen, Jeff Z. Pan, Huajun Chen, and Wen Zhang. Noise-powered multi-modal knowledge graph representation framework. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 141–155, Abu Dhabi, UAE, jan 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.11/.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S. Rosenblum. *MMKG: Multi-modal Knowledge Graphs*, pp. 459–474. Springer International Publishing, 2019. ISBN 9783030213480. doi: 10.1007/978-3-030-21348-0_30. URL http://dx.doi.org/10.1007/978-3-030-21348-0_30.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Tara Safavi and Danai Koutra. CoDEx: A Comprehensive Knowledge Graph Completion Benchmark. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8328–8350, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.669. URL https://aclanthology.org/2020.emnlp-main.669/.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2071–2080, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/trouillon16.html.

Lei Wang, Richong Zhang, Junfan Chen, Jaein Kim, Wei Xiao, and Hui Liu. Revealing multimodal information trade-off in multimodal knowledge graph completion. *Expert Systems with Applications*, 322:132003, Aug 2026. ISSN 0957-4174. doi: 10.1016/j.eswa.2026.132003. URL http://dx.doi.org/10.1016/j.eswa.2026.132003.

Bishan Yang, Wen tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases, 2015. URL https://arxiv.org/abs/1412.6575.

Haotong Yang, Zhouchen Lin, and Muhan Zhang. Rethinking knowledge graph evaluation under the open-world assumption. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8374–8385. Curran Associates, Inc., 2022. doi: 10.52202/068431-0609. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/378226e5df7eded3e401de5c9493143c-Paper-Conference.pdf.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. Native: Multi-modal knowledge graph completion in the wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, pp. 91–101. ACM, July 2024. doi: 10.1145/3626772.3657800. URL http://dx.doi.org/10.1145/3626772.3657800.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. Tokenization, fusion, and augmentation: Towards fine-grained multi-modal entity representation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12): 13322–13330, April 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i12.33454. URL http://dx.doi.org/10.1609/aaai.v39i12.33454.

Ying Zhou, Xuanang Chen, Ben He, Zheng Ye, and Le Sun. Re-thinking knowledge graph completion evaluation from an information retrieval perspective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pp. 916–926. ACM, July 2022. doi: 10.1145/3477495.3532052. URL http://dx.doi.org/10.1145/3477495.3532052.

Jing Zhu, Yuhang Zhou, Shengyi Qian, Zhongmou He, Tong Zhao, Neil Shah, and Danai Koutra. Mosaic of modalities: A comprehensive benchmark for multimodal graph learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14215–14224, June 2025.

A MyGO DIAGNOSTIC ARCHITECTURE AND CONTROL FLOW

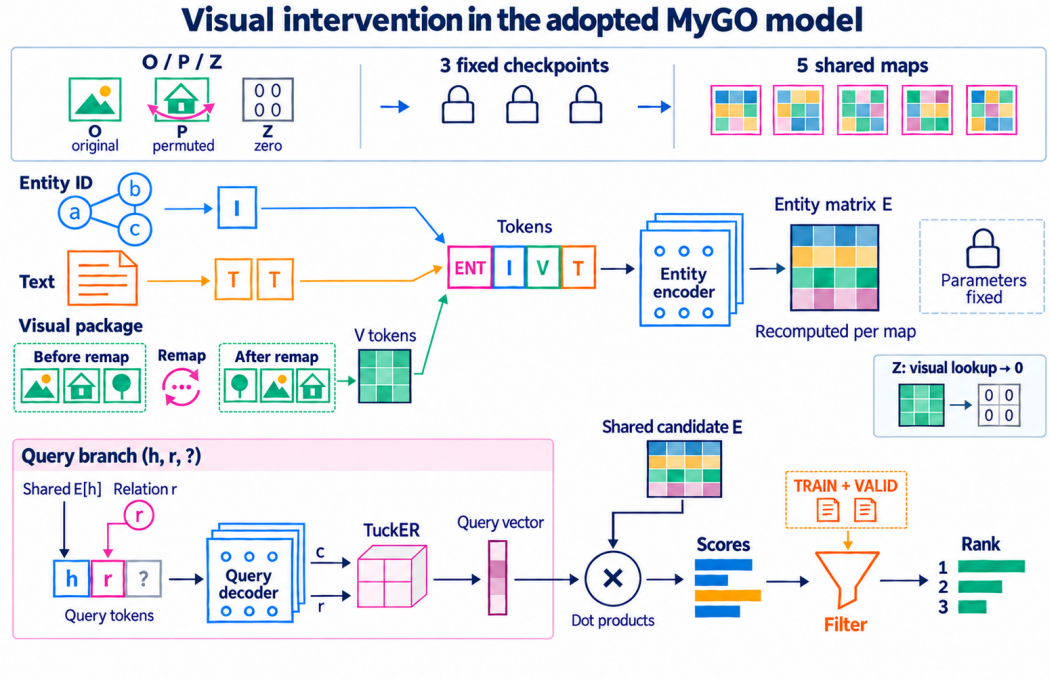

Figure 3: Visual reassignment and scoring in the adopted MyGO model (Zhang et al., 2025). Entity IDs, text tokens, and visual tokens enter the entity encoder. Repeated E labels denote the same recomputed representation matrix, shared by query and candidate branches. Query-decoder outputs pass through TuckER (Balazevic et al., 2019) and are scored against candidate vectors. Picture columns before and after remapping denote fixed entity slots. Each O/P/Z arm has one seed-2024 checkpoint at epoch 100. Five shared new mappings supply paired interventions against each arm's training-mapping baseline. Saved parameters remain fixed while E is recomputed. TRAIN and VALID positives enter the ranking filter after scoring. Z zeroes the visual-token lookup output; projection biases, positional embeddings, slots, and masks remain. Entity identifiers supply ID tokens, with no additional graph encoder. Pictures, token counts, matrix colors, and ranks are illustrative. This functional schematic summarizes adopted components and omits internal layer details.

For the original-input arm, the baseline uses the original entity-to-visual-package association. For the permanently permuted arm, it uses the fixed association present throughout that arm's training. All five alternative mappings are generated from the original package table with the same active-token-count constraints. The zero-input hook remains active in every evaluation of Z. Text inputs, graph query identities, candidate vocabulary, and positive filter are fixed across each checkpoint's comparisons. All 17,840 validation queries contribute to both the signed and absolute response summaries.

B IMPLEMENTATION AND AUDIT DETAILS

Resource provenance. The DB15K graph and legacy visual HDF5 are acquired from the MMKB release associated with MMKG (Liu et al., 2019).[2] Entity IDs follow the MyGO/AdaMF-MAT dictionaries, and the frozen 768-dimensional entity text features come from AdaMF-MAT.[3] Entity alignment back to the source graph was checked by URI; local splits operate on the deduplicated fact set. The frozen-transfer visual resource covers 12,837 entities, with five missing rows represented by zero vectors and availability flags. MyGO's tokenized visual resource has a different coverage of 12,570 entities. The legacy feature extraction gaps described in Section 3 remain unresolved. M's locally extracted VGG16 uses the TorchVision ImageNet-1K V1 weights and fc7 ReLU output in evaluation mode. S/M MiniLM and CLIP checkpoints are pinned to revisions `1110a243fdf4` and `3d74acf9a28c`, respectively; MiniLM uses attention-mask mean pooling and maximum length 256. All nonmissing frozen feature rows are locally L2-normalized.

Frozen-transfer teachers. Teacher entity and relation vectors use 128 real dimensions. Training uses reciprocal queries, Adam with learning rate 0.01, batch size 512, and regularization coefficient 0.001 (complex-magnitude N3 for ComplEx and real L3 for DistMult). Checkpoints are evaluated every ten epochs, with the earliest checkpoint retained on a tie. The two ridge maps are fitted separately on base entities, and the visual fit uses image-available base entities. Text-only prediction replaces unseen entity vectors with the text map. Fusion replaces them with the arithmetic mean of text and visual predictions; missing-image predictions fall back to text. The graph scoring functions and learned relation vectors are then held fixed.

Input controls. The union of images changed by either alpha-compositing background contains 706 entities. Re-encoding the old RGB version for those entities plus 32 fixed unchanged controls produces 738 checks. The maximum absolute discrepancy from stored normalized features is 1.79×10^{-7} for CLIP and 3.40×10^{-7} for VGG, below the prespecified numerical tolerance of 10^{-4}. Entity order, text features, availability masks, and graph arrays are preserved. White and black composites are evaluated together; no background is selected by downstream performance.

MyGO configuration. The pilot uses embedding dimension 256, feed-forward dimension 1,024, two attention heads, one layer per modality encoder, and one decoder layer. It retains eight visual tokens and four text tokens, with dropout values 0.01 (general), 0.6 (entity embedding), 0.3 (visual), and 0.1 (text). Training uses Adam at learning rate 0.001, batch size 2,048, gradient clipping at 0.1, and the implementation's cosine warm-restart schedule with initial period 50 and multiplier 2. The auxiliary-loss coefficient is 0.01. The compacted model has 22,716,928 trainable and 3,933,984 frozen parameters. The frozen dictionary compaction preserves accessed token embeddings; it reduces stored unused rows. Evaluation runs in float32 on one RTX 4090 with PyTorch 2.7.1 and NumPy 2.2.6, using batches of 128 validation triples. The checkpoints save model state without complete optimizer and scheduler state, which prevents an exact training-state continuation from these artifacts alone.

Negative controls and replication units. The reassignment worker recomputes each historical baseline and compares all 17,840 ranks exactly. Same-map reruns compare ranks and hashes of full scores and intermediate entity/relation representations. All five Z mappings preserve these hashes, ranks, target scores, and top-ranked candidates. Full score matrices are hashed during execution; per-query ranks, target scores, and top-ranked candidate identities are saved. The old alternative mapping (seed 271831) informed the question and is excluded from the new five-mapping summary. The original permanent training mapping uses seed 271830. Analysis is conditional on these selected models and graph.

Audit scope. Separate code paths recomputed 2,048 alpha-study ranks and 512 true-score filtering cases, and checked saved summaries against rank arrays. Hash checks establish byte-level continuity, while rank and control checks establish specific numerical properties. These automated audits cannot establish that the research question is novel, that every implementation choice is optimal, or that an unobserved population would exhibit the same effect.

[2] `https://github.com/mniepert/mmkb`

[3] `https://github.com/zjukg/AdaMF-MAT`

C NEWCOMER-TARGET RESULTS ACROSS THE FIXED-SCORE FILTER GRID

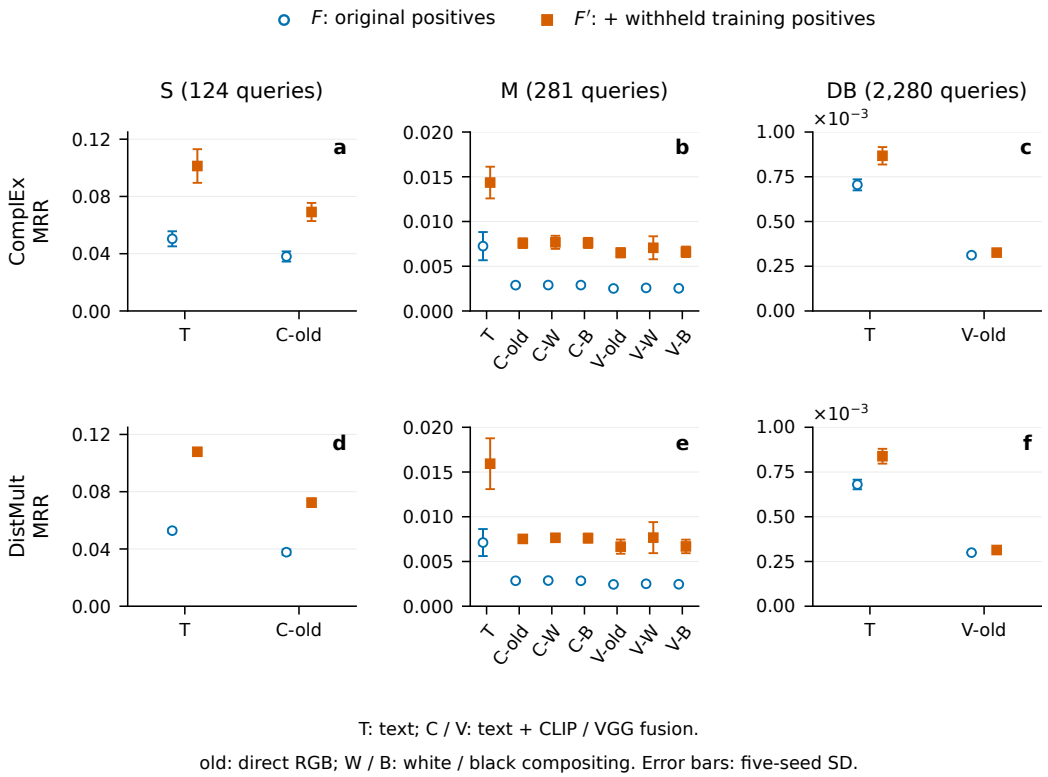

Figure 4: Newcomer-target results for all input configurations in the fixed-score filter grid. Points show five-seed mean MRR and error bars show sample standard deviation. The intervention changes evaluator positive sets with model scores fixed. Panels separate graph and teacher; all panels show newcomer-target queries. M includes both encoders and both repaired backgrounds. Absolute metric increases are implied by filter-set inclusion, while magnitudes depend on positive-target scores.

D SEED VARIATION IN NEWCOMER-TARGET MEASUREMENTS

Tables 5 and 6 report all newcomer-target means and sample standard deviations. The five seeds share data, features, and queries; variability is conditional on that shared material. Per-role summaries and paired comparisons accompany the source package.

Table 5: M alpha-input diagnostic, original filter. Each row summarizes five teacher seeds and 281 newcomer-target queries per seed.

Teacher	Input	MRR mean ± SD
ComplEx	Text	0.007246 ± 0.001577
DistMult	Text	0.007117 ± 0.001509
ComplEx	CLIP / old	0.002887 ± 0.000139
DistMult	CLIP / old	0.002847 ± 0.000163
ComplEx	CLIP / white	0.002898 ± 0.000146
DistMult	CLIP / white	0.002870 ± 0.000149
ComplEx	CLIP / black	0.002893 ± 0.000134
DistMult	CLIP / black	0.002843 ± 0.000155
ComplEx	VGG / old	0.002514 ± 0.000140
DistMult	VGG / old	0.002438 ± 0.000116
ComplEx	VGG / white	0.002578 ± 0.000178
DistMult	VGG / white	0.002506 ± 0.000146
ComplEx	VGG / black	0.002531 ± 0.000146
DistMult	VGG / black	0.002456 ± 0.000126

Table 6: All newcomer-target configurations under the two fixed-score filters. Each entry is mean ± sample SD over five teacher seeds. Rows within each graph use the same validation queries and candidate universe.

Graph/teacher	Input	MRR, F	MRR, F'
S/C	text	0.050413 ± 0.005263	0.101264 ± 0.011754
S/D	text	0.052788 ± 0.001385	0.107902 ± 0.001657
S/C	CLIP / old	0.038058 ± 0.003550	0.069141 ± 0.006341
S/D	CLIP / old	0.037744 ± 0.001895	0.072348 ± 0.002883
M/C	text	0.007246 ± 0.001577	0.014352 ± 0.001772
M/D	text	0.007117 ± 0.001509	0.015928 ± 0.002841
M/C	CLIP / old	0.002887 ± 0.000139	0.007579 ± 0.000540
M/D	CLIP / old	0.002847 ± 0.000163	0.007524 ± 0.000437
M/C	VGG / old	0.002514 ± 0.000140	0.006505 ± 0.000514
M/D	VGG / old	0.002438 ± 0.000116	0.006659 ± 0.000792
M/C	CLIP / white	0.002898 ± 0.000146	0.007674 ± 0.000732
M/D	CLIP / white	0.002870 ± 0.000149	0.007650 ± 0.000386
M/C	CLIP / black	0.002893 ± 0.000134	0.007606 ± 0.000564
M/D	CLIP / black	0.002843 ± 0.000155	0.007610 ± 0.000498
M/C	VGG / white	0.002578 ± 0.000178	0.007063 ± 0.001284
M/D	VGG / white	0.002506 ± 0.000146	0.007666 ± 0.001735
M/C	VGG / black	0.002531 ± 0.000146	0.006610 ± 0.000586
M/D	VGG / black	0.002456 ± 0.000126	0.006683 ± 0.000758
DB/C	text	0.000705 ± 0.000031	0.000867 ± 0.000049
DB/D	text	0.000680 ± 0.000027	0.000838 ± 0.000041
DB/C	VGG / old	0.000312 ± 0.000011	0.000326 ± 0.000013
DB/D	VGG / old	0.000300 ± 0.000003	0.000314 ± 0.000004

C = ComplEx; D = DistMult. Text is shared across visual variants within each graph and teacher. The S/M visual source overlaps, and DB uses a different feature provenance.