Multimodal Pictogram Embeddings for Augmentative and Alternative Communication: A Comparative User Study

João Victor da Silva[1] **, Silas A. R. Dos Santos,**[1]**,**
André C. A. Nascimento [1]**, Taciana P. Falcão** [1]

[1] Departamento de Computação, Universidade Federal Rural de Pernambuco (UFRPE)
Recife, Brazil

{joao.vsilva4, silas.augusto, andre.camara, taciana.pontual}@ufrpe.br

Abstract. *Augmentative and Alternative Communication (AAC) systems rely on retrieving appropriate pictograms from large, multilingual databases, a task for which classical methods such as TF-IDF are efficient but semantically shallow, often failing when queries and pictogram descriptions diverge lexically. Despite recent progress in dense and multimodal representations, the literature still lacks a systematic comparison of textual and multimodal embeddings for the pictogram* retrieval *task, particularly under real usage conditions. We address this gap by comparatively evaluating three representation approaches over the ARASAAC corpus of 13,586 pictograms: TF-IDF, dense multilingual text embeddings (Sentence Transformers), and multimodal embeddings from a domain-adapted CLIP model. Our CLIP variant replaces the original English text encoder with a multilingual one and is fine-tuned on all ARASAAC pictogram–description pairs, extending linguistic coverage to Portuguese while preserving text–image alignment; this fine-tuning yields relative gains above 155% in P@1 and 167% in MRR over a zero-shot baseline. We further combine offline retrieval metrics with a user study of 34 participants spanning literal (single-concept) and compositional (multi-concept) query scenarios. The central finding is a consistent* performance inversion*: TF-IDF dominates literal queries (P@4 = 0.912, mean rating 4.625/5), whereas fine-tuned CLIP is markedly superior for compositional queries (P@4 = 0.806, mean rating 4.170/5), with text embeddings occupying an intermediate position in both settings. These results indicate that no single representation is universally optimal for AAC pictogram retrieval and motivate adaptive, query-aware hybrid architectures that route each query to the most suitable method.*

1. Introduction

Augmentative and Alternative Communication (AAC) systems play a fundamental role in supporting individuals with speech and language impairments, enabling the expression of needs, emotions, and ideas through pictograms. Despite their social and pedagogical relevance, traditional pictogram sets are generally restricted to a predefined vocabulary and image set, which limits expressive and communicative personalization [Pereira et al. 2022a, Moadeli 2022]. This limitation is particularly critical in everyday contexts that demand flexibility and adaptation to personal situations.

Recent advances in generative AI have enabled the creation of pictograms from user-provided textual descriptions [Rombach et al. 2022]. Although promising, such solutions face significant portability and scalability challenges: most AAC platforms are

designed for offline use on resource-constrained devices (e.g., tablets and smartphones), making the direct integration of large generative models impractical [Brown et al. 2020]. Thus, there is a practical need for lighter and more efficient methods that balance performance with accessibility.

Traditionally, pictogram organization and retrieval in AAC systems have relied on classical information retrieval methods such as Term Frequency–Inverse Document Frequency (TF-IDF). While simple and computationally efficient, TF-IDF treats text in a purely statistical manner without considering semantics or context [Manning et al. 2008]. This limitation can be observed in the retrieval results for the Portuguese query "animal de estimação" ("pet"). As shown in Figure 1, TF-IDF retrieves results largely driven by lexical overlap with the query and its associated metadata, including "pet store" (rank 2), "collar" (rank 3), and "son" (rank 5). In contrast, the Sentence Transformer retrieves pictograms that remain more closely related to the broader concept of pets and animals, including "animal welfare association", "pets", "land animals", "animal husbandry", and "animal handler". The example illustrates the difference between lexical matching and semantic retrieval: while TF-IDF is strongly influenced by individual terms occurring in the indexed search strings, dense embeddings can capture broader semantic relationships between the query and the pictogram descriptions. It is worth noting that the original query is preserved along with the English translations of the labels for the recovered pictograms.

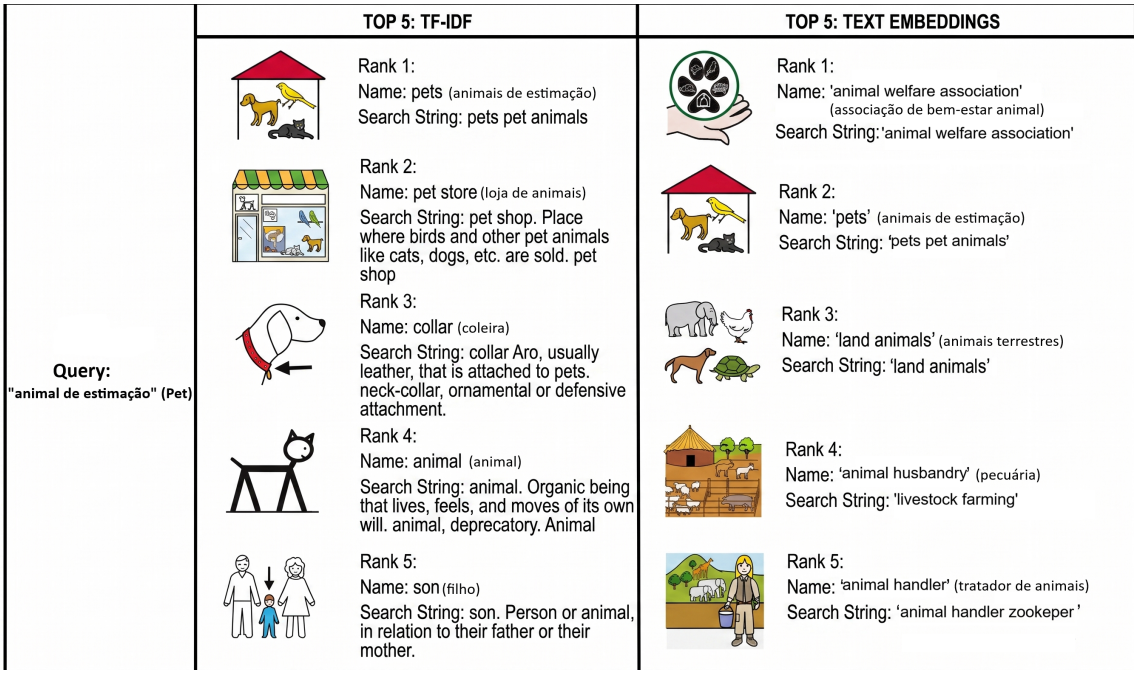


Figure 1: Comparison of pictogram retrieval results for the original Portuguese query "animal de estimação" ("pet") using TF-IDF and multilingual text embeddings.

Deep-learning models based on Transformer architectures have brought significant advances in dense text representations. Models such as Sentence Transformers [Reimers and Gurevych 2019] generate multilingual semantic embeddings that preserve relationships between words and concepts in a shared vector space, which is particularly relevant for multilingual pictogram databases. Furthermore, multimodal models such

as CLIP (Contrastive Language–Image Pretraining) [Radford et al. 2021] project images and texts into a common vector space, enabling cross-modal retrieval. In the AAC context, this approach is particularly promising because it combines the visual expressiveness of pictograms with the semantic richness of text, overcoming limitations of purely textual models [Draffan et al. 2023, Pereira et al. 2022a, Cabello et al. 2018].

Despite these advances, the literature lacks a systematic comparison of textual and multimodal representations for multilingual pictogram retrieval. Existing AAC studies have primarily focused on keyword-based retrieval or next-pictogram prediction, leaving the practical comparison of sparse, dense-text, and multimodal representations underexplored [Cabello et al. 2018, Pereira et al. 2022b, Draffan et al. 2023].

We therefore compare three approaches: (i) TF-IDF over textual descriptions; (ii) multilingual Sentence Transformer embeddings using *paraphrase-multilingual-MiniLM-L12-v2*; and (iii) multimodal embeddings from fine-tuned CLIP. Our contributions are threefold: (i) we quantify the effect of domain-specific fine-tuning by comparing zero-shot and fine-tuned CLIP; (ii) we identify a query-type-dependent performance inversion between TF-IDF and fine-tuned CLIP; and (iii) we release 612 annotated relevance judgments from 34 participants over the ARASAAC corpus as a benchmark for Portuguese-language pictogram retrieval.

2. Background

2.1. TF-IDF Representations

TF-IDF [Salton and Buckley 1988] represents documents as sparse vectors weighted by the importance of each term. We compute $\mathrm{TFIDF}_{t,d} = \mathrm{TF}(t,d) \times \mathrm{IDF}(t,D)$, where $\mathrm{IDF}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$ [Robertson 2004]. We use cosine similarity for ranking, following prior AAC pictogram-retrieval work [Pereira et al. 2022b]. Each pictogram search string concatenates its ARASAAC keywords and meanings, without additional stop-word removal or stemming, since these fields are short and keyword-dense. The vectorizer uses unigrams over the full corpus. Although computationally efficient, TF-IDF is limited in capturing semantic relationships such as synonyms and related concepts [Manning et al. 2008].

2.2. Dense Multilingual Embeddings

Transformer-based models produce dense, low-dimensional embeddings that encode distributed semantic information [Reimers and Gurevych 2019]. The *paraphrase-multilingual-MiniLM-L12-v2* model [Reimers and Gurevych 2019], based on the Sentence-BERT (SBERT) architecture, uses siamese networks to derive sentence-level vector representations that can be compared via cosine similarity. It maps sentences in over 50 languages into a shared 384-dimensional vector space, making it suitable for multilingual pictogram databases such as ARASAAC. Compared to larger models, MiniLM offers a favorable trade-off between quality and computational cost—a critical aspect for mobile AAC applications [Reimers and Gurevych 2019]. However, performance may degrade for rare or domain-specific terms [Leys 2022].

2.3. CLIP: Contrastive Language–Image Pretraining

CLIP [Radford et al. 2021] uses dual image and text encoders that map both modalities into a shared vector space through contrastive learning. This enables cross-modal retrieval

by comparing image and text embeddings. Although effective in general domains, its performance may degrade on specialized visual domains such as AAC pictograms, which have a distinctive graphic style [Schall et al. 2024]. Domain-specific fine-tuning can improve alignment between the visual characteristics of the target domain and its textual descriptions [Schall et al. 2024, Lülf et al. 2024].

2.4. Evaluation Metrics

We adopt standard information retrieval metrics. **Precision@K** measures the fraction of relevant results among the top-K retrieved items. **Mean Reciprocal Rank (MRR)** is the average of the reciprocal rank of the first relevant result across all queries: $\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i}$. **Mean Average Precision (mAP)** is the mean of per-query average precisions. In AAC contexts, Precision@5 and Precision@10 are particularly important, since users typically view only the first results in a pictogram grid [Pereira et al. 2022b].

3. Related Work

The intersection of artificial intelligence and AAC systems is a rapidly evolving research area, driven by the need for more fluent, expressive, and accessible communication. Pereira et al. [Pereira et al. 2022b] reviewed computational approaches for pictogram selection in AAC and reported a shift from knowledge-based systems to n-gram models and deep learning methods. Even though the reviewed works primarily address next-pictogram prediction, they are relevant to retrieval-based AAC systems because they show that semantic and contextual representations are increasingly important for selecting appropriate pictograms. Pereira et al. [Pereira et al. 2024] further extended this line of research with PrAACT, showing how Transformer-based methods can support larger AAC communication boards.

Outside the AAC domain, CLIP-based multimodal retrieval in specialized visual domains has been investigated more directly. Schall et al. [Schall et al. 2024] show that fine-tuning CLIP on domain-specific data consistently improves retrieval metrics without degrading the pretrained text–image alignment, a finding that directly motivates our fine-tuning strategy over the 13,586 ARASAAC pairs. Lülf et al. [Lülf et al. 2024] propose CLIP-Branches, which refines multimodal search through iterative user feedback—an approach with clear implications for future AAC personalization, where pictogram preferences vary across users. However, neither line of work targets AAC pictograms specifically, which combine a standardized, low-detail graphic style with multilingual textual metadata, a combination not addressed by general-purpose CLIP fine-tuning studies.

Benevento et al. [Benevento et al. 2025] situate AI as one of the main technological directions in contemporary AAC, highlighting its use in functions such as user-input optimization, communicative content generation, prediction, and adaptability. Their review is relevant because it shows that recent AAC research is moving beyond static symbol systems toward intelligent mechanisms that can better adapt communicative resources to user needs, which directly motivates the search for more effective pictogram representation and retrieval methods. Chang et al. [Chang et al. 2025] present an AAC system for individuals with cerebral palsy that combines handshape recognition, language prediction, and speech synthesis to reduce the physical effort required for communication. This work is related because it exemplifies how AI can support AAC interaction for users with

motor impairments; however, unlike studies focused on pictogram retrieval, it emphasizes input and message generation rather than multimodal representations for AAC databases.

Prior AAC research has primarily focused on textual models for next-pictogram prediction and on retrieval interfaces based on keywords, categories, or task-specific rules, sometimes adapted to particular contexts such as intercultural communication or curated vocabulary sets. While these studies are important for supporting pictogram selection, they do not investigate how general-purpose textual and multimodal embeddings compare in the pictogram *retrieval* task itself. The present work fills this gap by systematically evaluating three representation approaches—TF-IDF, dense multilingual text embeddings, and fine-tuned CLIP-based multimodal embeddings—on the ARASAAC corpus with 13,586 pictograms, and by examining their practical impact on AAC-oriented pictogram search.

4. Proposed Methodology

Our work is structured in four stages: (i) data collection and preparation; (ii) embedding generation; and (iii) experimental evaluation in a controlled setting. The methodological framework was inspired by Shao et al. [Shao et al. 2019], who demonstrate how to coordinate offline quality assessment with empirical user-satisfaction investigation.

4.1. Data Collection and Preprocessing

The ARASAAC database [Gobierno de Aragón 2024], maintained by the Government of Aragon, provides 13,586 pictograms with associated textual descriptions (keywords and meanings) in multiple languages, including Portuguese and English. For each pictogram, keywords and meanings were extracted from JSON metadata files and concatenated into a *search string* used for indexing. Images were loaded and converted to RGB format. The CLIP processor handles resizing and transformations for the visual encoder, while text is tokenized for each respective model.

4.2. Embedding Generation

Three approaches were employed, reflecting the evolution from sparse statistical to dense multimodal representations:

TF-IDF. Search strings were vectorized using TF-IDF over the full corpus, producing a sparse matrix where each row represents a pictogram and each column corresponds to a vocabulary term [Robertson 2004].

Sentence Transformers. The pre-trained model *paraphrase-multilingual-MiniLM-L12-v2* [Reimers and Gurevych 2019] was used to generate 384-dimensional dense embeddings from each search string. These embeddings were normalized and indexed using FAISS [Meta Engineering 2017] for efficient similarity search.

CLIP Fine-Tuned. The main contribution of this work involves generating multimodal embeddings using CLIP with ViT-B/16 and ViT-B/32 visual encoders [Radford et al. 2021]. Crucially, the original English-only text encoder was replaced by *clip-ViT-B-32-multilingual-v1* [Reimers and Gurevych 2019], which projects multilingual text into the same vector space as the ViT-B/32 visual encoder. This substitution preserves the learned text–image alignment while extending linguistic coverage to

Portuguese. Fine-tuning was performed on all 13,586 pictogram–description pairs from ARASAAC to specialize the model for the AAC pictogram domain.

Figure 2 illustrates the hybrid architecture, and Figure 3 provides a comparative overview of the three embedding pipelines. All embeddings were indexed using FAISS [Meta Engineering 2017] and stored in cloud storage for efficient retrieval during similarity search.

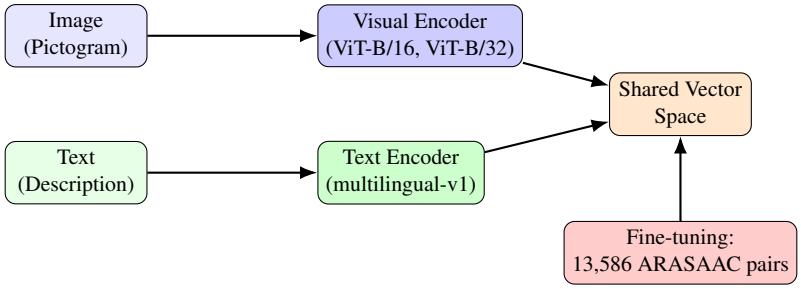


Figure 2: Multimodal embedding generation with the hybrid CLIP architecture, specialized via fine-tuning on ARASAAC pictogram–description pairs.

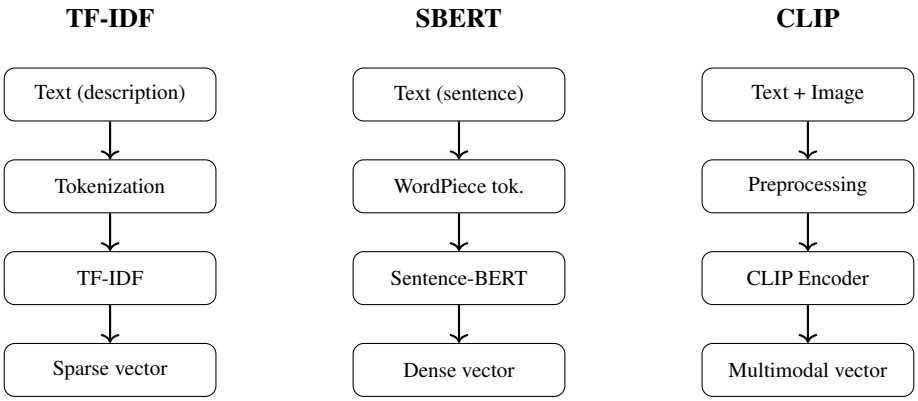


Figure 3: Comparative pipeline for textual and multimodal embedding generation.

4.3. Quantitative Evaluation of CLIP Fine-Tuning

Before involving end users, we quantified the impact of domain fine-tuning on CLIP itself, isolating this methodological contribution from the three-way comparison reported in Section 6. We adopted a forced-choice 1-vs-N classification protocol: given a pictogram image, the model must identify the correct textual description among $N = 50$ candidates (1 correct and 49 random distractors drawn from the corpus), ranked by cosine similarity between the visual embedding and each text embedding. Precision@K (P@1, P@5, P@10) and MRR were computed over this protocol, together with top-1 classification accuracy. Table 1 presents the results, comparing the original zero-shot CLIP (English text encoder) against our fine-tuned variant (multilingual text encoder, fine-tuned on all 13,586 ARASAAC pairs).

Fine-tuning improved every metric: P@1 more than doubled (4.27%→10.92%), and P@10 reached 34.47%, meaning the correct pictogram appears among the top-10 results in roughly one-third of queries; MRR improved by 167.9%. While absolute P@1

Table 1: Comparative performance of zero-shot CLIP vs. fine-tuned CLIP on retrieval metrics (P@K and MRR).

Metric	CLIP Zero-Shot	CLIP Fine-Tuned	Relative Gain
P@1	4.27%	10.92%	+155.7%
P@5	9.43%	26.07%	+176.4%
P@10	12.77%	34.47%	+169.9%
MRR	0.0649	0.1739	+167.9%
Accuracy (1-vs-49)	25.40%	60.10%	+136.6%

remains moderate, this must be contextualized: the task involves open retrieval over 13,586 pictograms with high semantic and visual density, since many share ARASAAC's standardized graphic style. In this setting, P@1 values around 10% are consistent with benchmarks reported for specialized multimodal retrieval domains [Schall et al. 2024]. The complementary 1-vs-49 forced-choice accuracy (60.10%) confirms that fine-tuning moved pictograms and descriptions into more discriminative regions of the shared vector space, and it is this specialized model — not zero-shot CLIP — that is used in the three-way comparison of Section 6.

5. User Study

While the offline evaluation measures algorithmic performance, a user study is essential to validate practical utility. We designed a user study with user satisfaction as the primary annotation metric, following established practices in search-system evaluation [Shao et al. 2019].

The study addresses three research questions: **RQ1 (Effectiveness):** What is the relevance of pictograms retrieved by each representation method, as judged by users? **RQ2 (Query-type sensitivity):** Does the relative effectiveness of TF-IDF, dense text embeddings, and multimodal CLIP embeddings depend on whether the query is literal (single-concept) or compositional (multi-concept)? **RQ3 (Reliability):** Are the collected relevance judgments consistent enough, across items and across the objective and subjective measures, to support comparative conclusions between methods?

5.1. Interactive Evaluation Tool

We developed a web-based research prototype that guides participants through a structured workflow. The interaction flow consists of three stages: (1) participant identification using email; (2) task selection and query input; (3) result evaluation.

The main evaluation interface (Figure 4) presents results from three models (A, B, C corresponding to TF-IDF, Text Embeddings, and CLIP Fine-Tuned, respectively) side by side in columns. Below each pictogram, participants provide a binary relevance judgment (relevant/irrelevant). Below each column, a mean star rating (0–5 scale) captures the overall quality perception for that model. Results are automatically stored in a cloud repository.

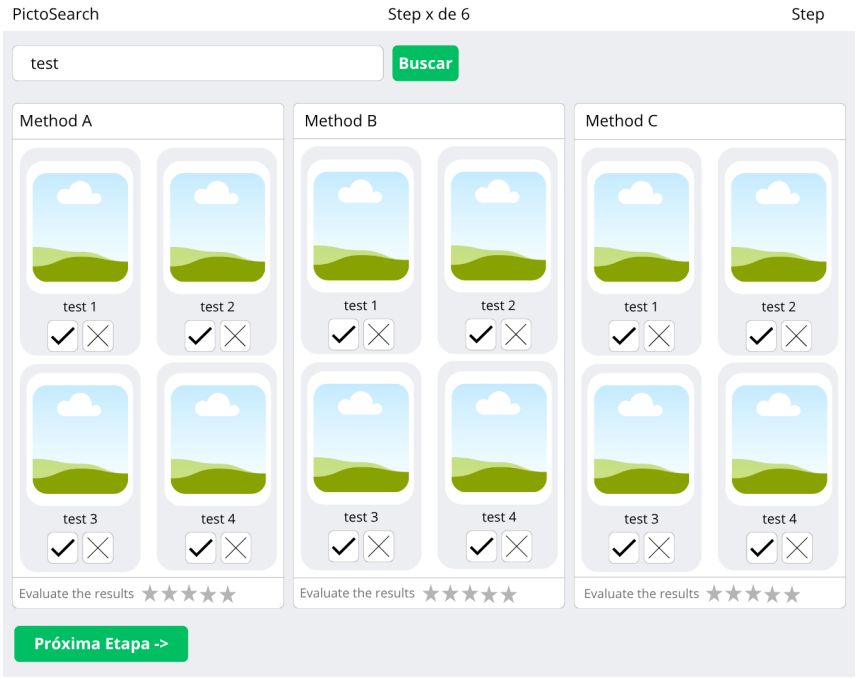


Figure 4: Evaluation interface showing side-by-side model results with per-item binary ratings and per-model star ratings.

5.2. Participants and Protocol

A total of 34 participants took part in the study. They were recruited through WhatsApp groups and e-mail, with the access link shared publicly to encourage voluntary participation. Each participant completed a single online session, lasting approximately 15 minutes, in which six search tasks were evaluated across the three models.

The sample was intentionally heterogeneous, including participants from different academic and professional backgrounds, such as technology, health, and other areas, as well as individuals with and without formal training in the subject domain. The age range varied from 18 to 60 years old, providing a broad set of perspectives for the evaluation.

The six tasks were divided into two scenarios: **Scenario A**, with literal/simple queries (Tasks 1–3), and **Scenario B**, with compositional/semantic queries (Tasks 4–6). The first scenario used single-concept terms directly present in ARASAAC metadata, while the second included multi-concept descriptions requiring contextual and compositional understanding.

Data collection combined interaction logs stored as JSON files, which recorded the queries, retrieved results, and model-specific ratings, with metric calculations and subjective star-based ratings.

6. Results and Discussion

6.1. Objective Retrieval Metrics (K = 4)

Table 2 presents the Information Retrieval metrics computed over the ordered lists returned by each model in both scenarios. A striking *performance inversion* is observed across scenarios. In Scenario A, TF-IDF achieves P@4 = 0.912 and MRR = 0.941,

while CLIP Fine-Tuned obtains only P@4 = 0.461. This reverses entirely in Scenario B, where CLIP Fine-Tuned reaches P@4 = 0.806 and MRR = 0.923, far above TF-IDF's P@4 = 0.268. Sentence Transformers consistently occupy an intermediate position, suggesting that semantic text embeddings partially capture compositional meaning but do not yet reach the level of visual–textual alignment provided by fine-tuned CLIP.

Table 2: Retrieval metrics (K = 4) by scenario and model. n = number of search instances, 34 participants × 3 queries per task × 3 models = 306 searches per scenario.

Scenario	Model	P@4	MRR	mAP	NDCG@4
A — Literal (Tasks 1–3)	TF-IDF	**0.912**	**0.941**	**0.938**	**0.985**
	Semantic (SBERT)	0.630	0.729	0.721	0.953
	CLIP Fine-Tuned	0.461	0.525	0.525	0.924
B — Compositional (Tasks 4–6)	TF-IDF	0.268	0.525	0.525	0.968
	Semantic (SBERT)	0.397	0.632	0.569	0.855
	CLIP Fine-Tuned	**0.806**	**0.923**	**0.913**	**0.954**

6.2. Subjective Evaluation (Star Ratings)

Table 3 presents mean star ratings (0–5 scale) and 95% confidence intervals (CI) obtained from participant evaluations. The 95% CI indicates the range within which the true population mean is estimated to lie with 95% confidence, computed via t-distribution given the sample size.

Table 3: Subjective star ratings (mean ± 95% CI) by scenario and model.

Scenario	Model	Mean ($\bar{x}$)	95% CI	Median	Std. Dev. (SD)
A — Literal	TF-IDF	**4.625**	[4.429, 4.821]	5.0	0.965
	Semantic (SBERT)	3.438	[3.109, 3.766]	4.0	1.621
	CLIP Fine-Tuned	2.787	[2.460, 3.115]	3.0	1.599
B — Compositional	TF-IDF	1.947	[1.693, 2.202]	1.0	1.249
	Semantic (SBERT)	2.448	[2.192, 2.704]	2.0	1.264
	CLIP Fine-Tuned	**4.170**	[3.965, 4.375]	4.0	1.001

The subjective ratings mirror the objective metrics with remarkable consistency. TF-IDF drops from $\bar{x} = 4.625$ in Scenario A to $\bar{x} = 1.947$ in Scenario B ($\Delta = -2.678$), while CLIP Fine-Tuned rises from 2.787 to 4.170 ($\Delta = +1.383$). This inversion pattern is directly observable when the metrics are calculated, being a central qualitative finding of the study.

6.3. Per-Task Breakdown

Table 4 provides a granular view of P@4 and mean star ratings per individual task, enabling identification of edge cases within each scenario.

Among the notable outliers is Task 3 ("Cadeira de rodas" - wheelchair), where Sentence Transformers substantially underperforms the other approaches (P@4 = 0.265 versus 0.882 for TF-IDF and 0.816 for CLIP Fine-Tuned), likely due to the compound

Table 4: Per-task metrics P@4 and mean star ratings (⋆) for each model.

	TF-IDF		SBERT		CLIP FT	
Task	P@4	⋆	P@4	⋆	P@4	⋆
1 — Pineapple	**0.934**	**4.750**	0.713	3.719	0.132	1.633
2 — watching TV	**0.919**	**4.688**	0.912	**4.688**	0.434	2.625
3 — wheelchair	**0.882**	**4.438**	0.265	1.906	0.816	4.031
4 — family gathered at the table	0.206	1.781	0.169	1.531	**0.875**	**4.375**
5 — playing with friends	0.304	2.097	0.404	2.500	**0.765**	**3.968**
6 — starry sky	0.294	1.969	0.618	3.312	**0.779**	**4.161**

nominal concept ‘wheelchair” having multiple fragmented translations in ARASAAC. In contrast, CLIP Fine-Tuned achieves P@4 = 0.816, approaching the performance of TF-IDF while relying on multimodal representations, suggesting that visual information helps bridge lexical variability and improve retrieval robustness.

The pattern that emerges from the two scenarios can be summarized as a mechanism, not merely a coincidence of metrics: in Scenario A, TF-IDF benefits directly from rich, well-translated textual metadata, so when the query and the pictogram label coincide lexically, the conceptual neighborhood introduced by CLIP’s embeddings competes with, and loses to, the precision of literal matching. In Scenario B, the logic inverts for the opposite reason: queries describing scenes composed of multiple concepts exceed what TF-IDF’s lexical decomposition can capture, and only CLIP’s visual encoder, specialized on ARASAAC’s own pictograms, can associate the query with the depicted scene. SBERT’s behavior in both scenarios is consistent with this reading: by capturing textual semantic similarity without visual grounding, it outperforms TF-IDF when lexical decomposition fails, but does not reach CLIP’s image–text alignment, hence its stable, intermediate position in both regimes.

7. Conclusion

We presented a comparative study of three embedding approaches for pictogram retrieval in AAC: TF-IDF, dense multilingual text embeddings (Sentence Transformers), and fine-tuned multimodal CLIP embeddings. Our offline evaluation on the ARASAAC corpus (13,586 pictograms) demonstrated that fine-tuned CLIP yields substantial improvements over the zero-shot baseline, with relative gains exceeding 155% in P@1 and 167% in MRR. More importantly, a user study with 34 participants reveals a complementary performance inversion: TF-IDF dominates for literal, single-concept queries (Scenario A: P@4=0.912, mean star rating =4.625), while fine-tuned CLIP is markedly superior for compositional, multi-concept queries (Scenario B: P@4=0.806, mean star rating=4.170). Sentence Transformers consistently occupy an intermediate position in both scenarios.

Regarding RQ1 (Effectiveness), the results show that no representation method is universally superior. TF-IDF was most effective for literal, single-concept queries, achieving P@4 = 0.912 and a mean rating of 4.625 in Scenario A, whereas fine-tuned CLIP was substantially more effective for compositional queries, achieving P@4 = 0.806 and a mean rating of 4.170 in Scenario B. Regarding RQ2 (Query-type sensitivity), the results provide clear evidence that retrieval effectiveness depends strongly on query type:

the relative ranking of the methods is effectively inverted between literal and compositional scenarios, with TF-IDF moving from the best to the worst-performing method and fine-tuned CLIP showing the opposite pattern. Regarding RQ3 (Reliability), the agreement between objective retrieval metrics and subjective ratings provides convergent evidence that the collected judgments are sufficiently consistent to support the observed comparison, although these annotations reflect the judgments of general participants rather than AAC specialists and should therefore not be interpreted as a formal measure of inter-rater reliability.

These findings highlight that a single method is not universally optimal for AAC pictogram retrieval. Instead, the observed performance inversion suggests that an adaptive hybrid architecture could route queries to the representation method that best matches their semantic structure. Future work will therefore investigate: (i) query-type classification for automatic routing between retrieval methods; (ii) embedding compression techniques for resource-constrained mobile AAC devices; (iii) transfer-learning strategies for adaptation across pictogram databases and user contexts; and (iv) evaluation with additional languages, pictogram sets, and AAC specialists..

Acknowledgments

The authors thank the Livox Application Team for their support and collaboration. This work has also been supported by the following Brazilian research and innovation agencies: FINEP, FACEPE, CAPES and CNPq.

References

Benevento, A. D., Ciulla, G., and Merlo, G. (2025). Future technologies in alternative and augmented communication: a scoping review of innovations. *Frontiers in Communication*, Volume 10 - 2025.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Cabello, L., Lleida, E., Simón, J., Miguel, A., and Ortega, A. (2018). Text-to-pictogram summarization for augmentative and alternative communication. *Procesamiento del Lenguaje Natural*, 61:15–22.

Chang, Y.-S., Lin, C.-W., and Chen, C.-C. (2025). Ai-driven augmentative and alternative communication system for communication enhancement in cerebral palsy using american sign language recognition. *Sensors and Materials*, 37(9):3841–3853.

Draffan, E., Wald, M., Ding, C., and Yin, Y. (2023). Ai supporting aac pictographic symbol adaptations. *Studies in health technology and informatics*, 306:215–221.

Gobierno de Aragón (2024). Arasaac - aragonese center for augmentative and alternative communication. `https://arasaac.org/`. Accessed in: september 9 2025.

Leys, M. (2022). Multilingual domain-specific sentence transformers. `https://blog.ml6.eu/`

multilingual-domain-specific-sentence-transformers-d34f071763d0. Accessed in: 2026-01-28.

Lülf, C., Martins, D., Salles, M., Zhou, Y., and Gieseke, F. (2024). Clip-branches: Interactive fine-tuning for text-image retrieval. *arXiv preprint arXiv:2406.13322*.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Meta Engineering (2017). Faiss: A library for efficient similarity search. https://engineering.fb.com/2017/03/29/data-infrastructure/faiss-a-library-for-efficient-similarity-search/. Accessed in: 2026-02-03.

Moadeli, M. (2022). *Design and Evaluation of a Context Aware AAC Application for Pre-literate Children with Complex Communication Needs*. PhD thesis, University of Toronto.

Pereira, J. A., Macêdo, D., Zanchettin, C., De Oliveira, A. L. I., and Fidalgo, R. D. N. (2022a). PictoBERT: Transformers for next pictogram prediction. *Expert Systems with Applications*, 202:117231.

Pereira, J. A., Medeiros, S. D., Zanchettin, C., and Fidalgo, R. D. N. (2022b). Pictogram Prediction in Alternative Communication Boards: a Mapping Study. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação (SBIE 2022)*, pages 705–717, Brasil. Sociedade Brasileira de Computação - SBC.

Pereira, J. A., Pereira, J. A., Zanchettin, C., and Do Nascimento Fidalgo, R. (2024). PrAACT: Predictive Augmentative and Alternative Communication with Transformers. *Expert Systems with Applications*, 240:122417.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation - J DOC*, 60:503–520.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Schall, K., Barthel, K. U., Hezel, N., and Jung, K. (2024). Optimizing clip models for image retrieval with maintained joint-embedding alignment. *arXiv preprint arXiv:2409.01936*.

Shao, Y., Liu, Y., Zhang, F., Zhang, M., and Ma, S. (2019). On annotation methodologies for image search evaluation. *ACM Trans. Inf. Syst.*, 37(3).