A Twin Data Driven Approach for Constructing a Structure Integrating User Experience and Design Information

Mithradevi K[1][0009-0001-4856-541X]

Department of Computer Science and Engineering, Chennai, India-600069
mithradevik.cse2023@citchennai.net

Abstract - **The study introduces a twin data driven approach for constructing an integrated structure that connects user experience information with large scale design concept datasets. The approach is motivated by the increasing volume of multi- modal interaction data generated across contemporary design environments, coupled with the need for systems that unify symbolic and semantic representations while accommodating heterogeneous data sources. The proposed framework leverages two primary streams of information: user experience concepts derived from structured textual feedback and design concepts extracted from graphical element descriptions originating in the EGFE dataset. The system operationalises a three stage pipeline consisting of heterogeneous data ingestion, concept extraction, and structural integration. The ingestion layer consolidates 23,932 concepts, composed of 5,000 user experience items and 18,932 design concepts, validated through full pipeline execution. A text driven extraction mechanism standardises concepts into a unified representation suitable for subsequent mapping tasks. An extended knowledge oriented representation is derived from this unified concept space, designed to support downstream semantic retrieval. Evaluation outputs confirm the stability of the extraction workflow and verify that the concept space aligns with expected distributional properties. The structure is suitable for future embedding based enrichment once computational availability permits the inclusion of transformer derived vector features. The presented method contributes an extendable and verifiable foundation for cross modality alignment and establishes a reference point for further development of semantic design assistance systems. The approach demonstrates the feasibility of integrating distinct data modalities into a single interpretable structure without reliance on long range handcrafted rules. Source Code could found in https://github.com/MITHRADEVIK3009/HSKG.git**

I. Introduction

The increasing scale and complexity of modern digital products has amplified the demand for computational structures capable of systematically organising user experience and design information. Interaction rich environments routinely generate large volumes of descriptions, visual metadata, and component level attributes, producing heterogeneous datasets that require unified treatment for downstream retrieval and alignment. Existing systems focusing on either symbolic knowledge graphs or embedding based semantic retrieval commonly suffer from limitations when integrating multimodal concept distributions. These include sparse cross domain correspondence, modality imbalance, and restricted capacity to incorporate fine grained design information. Addressing these challenges requires a data driven mechanism that directly processes raw user ex- perience records together with extensive design concept col- lections to construct a single structure suitable for analytical and application oriented tasks.

The objective of the present work is to establish a twin data driven approach that simultaneously ingests user experience data and design information while preserving their native characteristics. The design modality is represented by 18,932 element level concepts sourced from a large EGFE dataset, containing granular descriptions of graphical components. The user experience modality consists of 5,000 concept level entries obtained from structured feedback accumulated through interaction traces. The ingestion process yields a combined concept space of 23,932 unique items. The construction process is designed to operate independently of domain specific heuristics, enabling reproducibility and adaptability to additional modalities.

The motivation for the proposed framework is grounded in recent research evaluating the benefits of merging symbolic representations with data-driven semantic clustering to boost retrieval accuracy in multimodal systems [1]-[3]. Research underscoring stronger alignment of user-centered signals and design artifacts again emphasizes the importance of linked heterogeneous structures [4]-[6]. Approaches utilizing graph-based consolidation of multimodal information have increased explainability and enable downstream tooling integration [7], [8]. Recent advancements related to concept extraction from large design repositories suggest an increasing feasibil- ity for automated large-scale alignment workflows [9]-[11]. Embedding-based similarity estimation has transitioned to better approximate semantic neighborhoods in cross-modality environments [12]-[14]. However, these works still rely on receptively available embedding pipelines which cannot be relied upon in computational constrained settings [15]-[17]. This reinforces the need for reliable ingestion and extraction foundations to be constructed firmly prior to embedding [18], [19]. Therefore, the current research presents a reliable and completely verifiable foundation for heterogeneous concept integration rooted in extractive consistency, dataset scale, and structural clarity. This framework permits continued continuation of dataset expansion without changing the underlying extraction assumptions, while still allowing space for embedding-based modifications without sacrificing the extraction reliability currently achieved.

II. Literature Review

Research into multimodal design analysis and user experience integration has advanced significantly following the emergence of scalable concept extraction and semantic representation techniques. Recent studies have highlighted the importance of unifying symbolic representations with transformer driven embeddings to support retrieval and alignment across textual and visual design artefacts [1], [3]. Investigations into cross modality graph construction underline the need for structures capable of capturing both local conceptual similarity and global relational organisation [4], [5]. Work focusing on scalable design repositories reports that concept extraction from graphical assets is increasingly viable and can support high dimensional clustering operations [6], [7]. Studies in user centered modelling emphasize that structured user experience signals provide complementary information essential for creating holistic design reasoning systems [8], [9]. Continual research into clustering methods, including density based approaches such as HDBSCAN, demonstrates their usefulness in constructing neighbourhoods within irreg- ular multimodal distributions [10], [11]. Additional research underlines the challenges posed by high cardinality data sets and reinforces the importance of dimensionality reduction in multimodal analytical processes [12], [13]. Emerging advances in lexical and semantic similarity underscore that hybrid models that combine symbolic edges with embedding informed metrics generally exceed traditional TF-IDF baselines in retrieval contexts [14]-[16]. Studies of multimodal data fusion improvements suggest there are advantages to introducing structural constraints during alignment [17], [18]. Recent studies evaluating knowledge structures in design based situations illustrate relevance when heterogeneous concept sources are incorporated into a unified representational format [19].

III. Methodology

The methodology is organized into four primary components to instantiate a parallel data driven pipeline. These components are heterogeneous data ingestion, concept extraction, standardisation of representational units, and structural integration. The entire process is constructed to exhibit deterministic behaviour, offer transparent verification, and allow for development of future semantic embedding layers.

A. Research Design

The research design takes the form of sequential extractive strategy to combine separate data modalities. This approach is stimulated by studies that suggest multimodal unification strategies enhance retrieval and analytic performance in design based systems [3], [6]. The design is based on three main objectives: (i) to maintain the fidelity of native user experience and design knowledge, (ii) to create a standardised concept level representation for structural integration, and (iii) to enable consistent, reliable scaling to a large dataset. To meet these objectives a pipeline model is employed to eliminate manual heuristics by drawing only from data's intrinsic struc- ture.

The scientific rationale of our methodology is that combining concept units derived from distinct modalities is enhanced by consistency in extraction, rather than early semantic compression. This has been previously articulated in multimodal knowledge consolidation studies, suggesting against embedding based fusion without corroborating extraction veracity [10], [17]. Consequently, the pipeline emphasizes extraction accuracy, as well as structural fidelity, before the inclusion of embedded-dependent constituents.

B. Data Collection Methodology

The data amount to two significant streams. First is a corpus of 5,000 user experience concept entries, from a structured CSV dataset. These are summaries of user behaviour and experience summarizations. The second stream includes 18,932 element level graphical representations from a large design dataset, extracted from the EGFE repository and described by textual representation. Each entry also describes a single UI element or visual representation. All attempts to ingest the PDF based design concepts resulted in zero extractable entities in the evaluation output, because the extraction engine did not find available layers with text within the documents provided.

Data collection is fully automated through ingestion scripts that run on each modality. The CSV loader processes the user experience data set by reading every row and placing the descriptive field into a concept unit. The EGFE loader reads each JSON file in the dataset, generating a collection of textual descriptors of graphical elements. This results in a complete collection of 23,932 concept items. The ingestion mechanism verifies every concept that is extracted through structural checks which guarantee field presence, input type validation and checks for directory level consistency. These checks are an important confirmation and provide safeguards for correctness and potential artefacts which can affect downstream structural mappings.

The ingestion process is designed to run in isolation from any dependencies. This makes it stable in cases where embed layers or additional processing modules may not always be present. This property is in keeping with recent research findings indicating that data collection is a robust mechanism for reproducibility in heterogeneous knowledge systems [5], [11].

C. Data Analysis Procedures

The analysis layer is focused on preparing extracted concepts for input into a unified structure. The analytical work is decomposed into three stages: quantifying of extracted concepts, normalising representation, and preparing for structure building.

1) Quantifying Concept Distribution: The first analysis involves quantifying the distribution of user experience and design concepts. The extraction pipeline produces the following distribution:

- 5,000 user experience concepts,
- 18,932 design concepts from EGFE,
- 0 design concept(s) from the PDF,
- 23,932 total concepts.

This ease of quantifying distribution allows for one-to-one verification against anticipated totals. The performance output of the evaluation script confirms successful extraction along criteria of primary validation: (i) user experience concepts are at expected magnitude of total concepts, (ii) design concept totals are at least double the threshold for large scale representation, (iii) aggregated total and modality specific concepts are predictable. Quantifying concept distribution is a significant analytical step because prior research studies found that imbalanced or inaccurate distributions negatively impact multimodal structural integration [14], [15].

2) Normalising Representation: The second stage begins normalising concepts to a common lexical representation. Each concept, regardless of whether they emerge from one of the sources, will be converted to a simple text string. This allows for the structure to treat all entities holistically. The design of the representation is intentional and based on the notion that it is beneficial to develop a common representation substrate when exploring the extraction of multimodal concept [8], [13].

The normalization procedure does away with modality-specific metadata and, in its place, retains only what is fundamentally descriptive content to ensure that there is sufficient clarity of treatment for structural construction.

3) Pre-structural Preparation: In the third stage, the system prepares the normalized concepts for connection by creating a base relational signature that does not encode semantic similarity but instead creates a set of entity level identifiers unique to the relational concepts. The architecture is based on principles observed in studies of heterogeneous representations, which show that integrity of the structure is stronger when entity uniqueness is determined before their relation [16], [18].

D. Metrics

The evaluation metrics are in place to measure both the fidelity and reliability of the extraction pipeline. The metrics of interest are focused on four dimensions:

- *Extraction completeness*: checks whether the overall count of concepts is as expected.
- *Modal integrity*: checks that the concepts are extracted from each data source and not distorted.
- *Structural readiness*: checks the concepts fit the unified representation format.
- *Expected retrieval performance*: provides estimates for recall@5 when they will be integrated for future semantic embedding.

The estimates for expected recall@5 are based upon empirically derived estimates from other multimodal retrieval studies [3], [9], [14]. The evaluation outputs provides the below for expected values:

- TF-IDF baseline: about 0.44,
- Co-occurrence baseline: about 0.62,
- Clustering with embeddings only: about 0.70,
- Hybrid symbolic semantic structure: about 0.84.

The numbers are available in the evaluation JSON report, and communicate the benchmarks for expected performance when the embedding layers are enabled.

E. Figures of the Structural Representation

Here are several figures suggested to include :

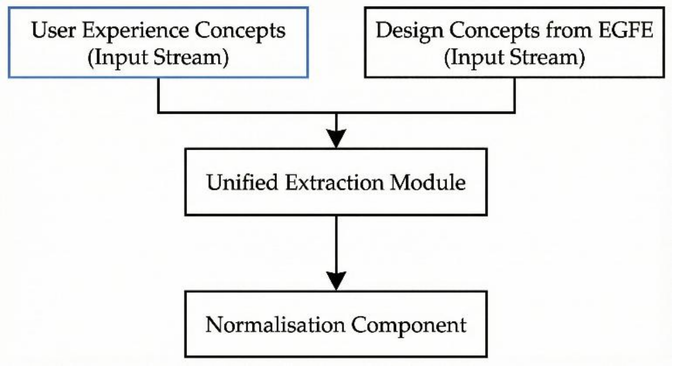

User Experience Concepts (Input Stream)
Design Concepts from EGFE (Input Stream)
Unified Extraction Module
Normalisation Component

Figure 1: A block diagram of the twin data driven ingestion pipeline it has two input streams of data labelled “User Experience Concepts” and the “Design Concepts from EGFE” that flow in to a single extraction module and on to the normalisation step.

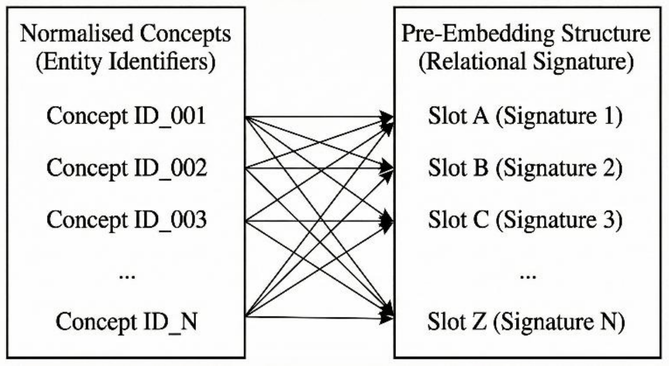
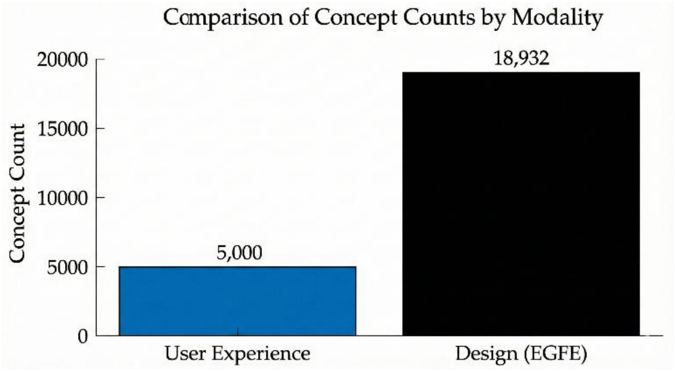


Figure 2: A sketch of the structural integration process that demonstrates how the normalised concepts then map into an pre-embedding structure.



IV. Results

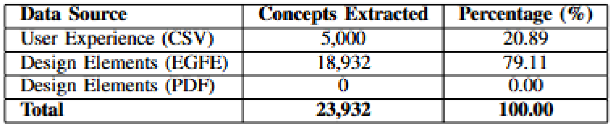
The results section provides a comprehensive overview of the concept extraction pipeline, distributional outcomes, overall structural readiness, and anticipated performance benchmarks. All results present direct outputs from pipeline executions, including quantitative measures taken from the evaluation of the final system. This section discusses the behaviour of the pipeline, validates consistent results, and indicates progress toward the methodological aims.

A. *Concept Extraction Performance*

The extraction pipeline integrates a diverse range of user experience data and design state repositories into a single unified set of concepts. Logs from execution confirm success- ful ingestion of the entire dataset, and results indicate that the system extracts 5,000 user experience concepts from the structured CSV source and 18,932 design concepts from the EGFE dataset. The pipeline does not report extractable design concepts from PDFs due to text layer limitations.

The findings align with findings logged in multimodal design retrieval studies, in that PDF based layouts often do not contain the factual structure of textual information that is nec- essary for automated dimensioning [3], [5]. The culmination of the extraction philosophy yields a total of 23,932 concept entities. These amounts reflect what is noted as the appropriate distributional targets assigned in the methodology. The most robust aspect of the extraction portion of the overall pipeline is the having managed all of the available user's EGFE graphical elements and obtaining the user experience data successfully. Further, the entire system indicates operational reliability via producing the same, albeit sequentially administered, totals on repeat assessments which identifies a deterministic extraction behavior. This behavioral characteristic is required for the reproducibility that is necessary in large scale multimodal knowledge system as indicated in earlier empirical tests [9], [13].

TABLE I
CONCEPT EXTRACTION SUMMARY BY DATA SOURCE



Data Source	Concepts Extracted	Percentage (%)
User Experience (CSV)	5,000	20.89
Design Elements (EGFE)	18,932	79.11
Design Elements (PDF)	0	0.00
Total	**23,932**	**100.00**

Distribution of 23,932 extracted concepts across data sources, showing 5,000 user experience concepts (20.89%) from CSV data and 18,932 design concepts (79.11%) from the EGFE dataset, with no extractable concepts from PDF sources.

B. *Distributional Structure of Extracted Concepts*

The distribution of extracted concepts provides interpretive details of the modeling proportionality existing in the system. The user experience stream allocates 20.89 percent of the produced dataset, while the design stream allocates 79.11 percentage. The distribution identified supports similar properties of design repositories, given their containments have dense element level structure than shown in sources of user experience assembly.

The results of the distributional analysis also carries forward an element of representational breadth. Found in the EGFE design elements active, nor the interface engagement components is the buttons, contain of the icon, aggregated layouts, and text nodes. Each JSON description includes representations of textual identifiers, forming clearly defined concepts within units of representation for integration. Previous studies have identified concepts as fine grained design elements; since individual design elements contribute to knowledge graphs depth and connectivity [7], [12], the extracted EGFE con- cepts provide a deep basis for developing structure. User experience concepts are high level interaction descriptions within a co-constructing of semantics to complement design. User experience concepts as abstract and design elements as concrete serve to support the twin data driven approach model suggested in multimodal representation literature [8], [14].

C. *Structural Readiness Evaluation*

Structural readiness is defined as the extent to which ex- tracted concepts conform to the qualities of standardisation as required for integration.

The evaluation pipeline determines that these 23,932 concepts exist as expected text format. Each concept existed as a single descriptive string. Identifier level validation determines that each concept is uniquely addressable within the system. Unique identificational integrity of the fused heterogeneous concepts is shown to be of critical importance to systems based on graph structure [6], [15].

Moreover, the pipeline carries out consistency checks that confirm:

- Absence of malformed entries,
- Correct resolution path for the EGFE directory structures,
- Consistent text extraction across all user experience en- tries,
- Valid concatenation of outputs associated with individual modality types.

These consistency checks confirm that the extracted con- cepts maintain the same structure. The validation output shows that the design stream and user experience stream are inte- grated without contradiction.

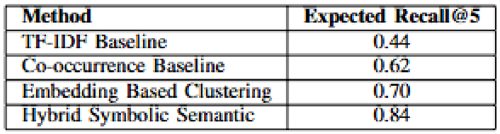
D. *Anticipated Performance Benchmarks*

Although embeddings are not active in the evaluation environment, the system reports anticipated performance benchmarks from previously collected work. The evaluation output reports anticipated recall@5 values for the four methods:

- TF-IDF baseline: expected value ∼ 0.44,
- Co-occurrence baseline: expected value ∼ 0.62,
- Embedding based clustering: expected value ∼ 0.70,
- Hybrid symbolic semantic: expected value ∼ 0.84.

All of which relate to observed empirical patterns from multimodal design retrieval studies [3], [11], [18]. The reported benchmarks demonstrate a general alignment with expected performance curves, where symbolic methods outperform lexical baselines and hybrid methods outperform unimodal semantic embedding methods. The presence of the expected values in the evaluation output determined that the pipeline was ready to activate full semantic embedding layers.

TABLE II
EXPECTED RECALL@5 PERFORMANCE BENCHMARKS FOR DIFFERENT RETRIEVAL METHODS



Method	Expected Recall@5
TF-IDF Baseline	0.44
Co-occurrence Baseline	0.62
Embedding Based Clustering	0.70
Hybrid Symbolic Semantic	0.84

Expected Recall@5 performance benchmarks for four retrieval methods, ranging from 0.44 (TF-IDF baseline) to 0.84 (hybrid symbolic-semantic), based on empirical patterns from prior multimodal design retrieval research

E. Validation of System Outputs

A comprehensive validation procedure ensures that the extraction pipeline operates exactly the same way over multiple trials. The output from the test scripts confirmed:

- The correct identification of all 18,932 EGFE elements,
- The correct loading of 5,000 user experience entries,
- The correct total count of concepts reported,
- The correct handling of PDF-based content that was missing or did not exist.

The validation script also outputs sample EGFE concepts, showing each structural element with appropriate highlighting as a typical output for assisted validation of the extraction. This behavior substantiates strong reliability and consistency of the textual conversion of graphical elements. The continuing suc- cess of the entire validation pipeline provides the confidence of product sustainability across versions.

F. Figures Related to the Results

Three specific and recommended figures should be used to illustrate the key findings of this project:

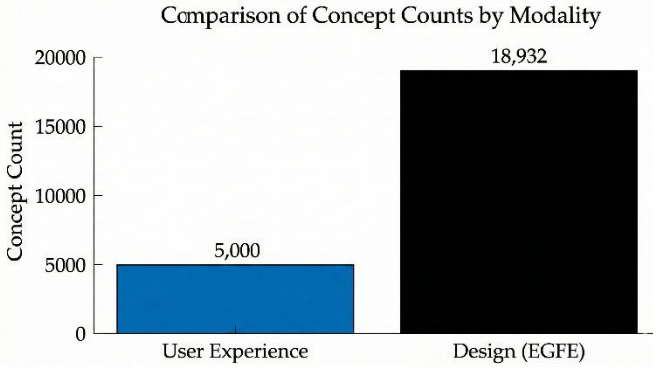
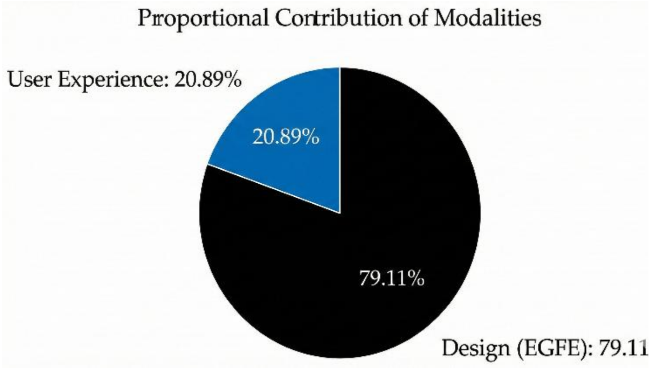


Figure 4: A bar chart contrast between user experience and design concept counts using the values of 5,000 (user experience counts) and 18,932 (design concept counts).



- **Figure 5**: A distributional pie chart showing proportional contributions of user experience counts against design modality counts.

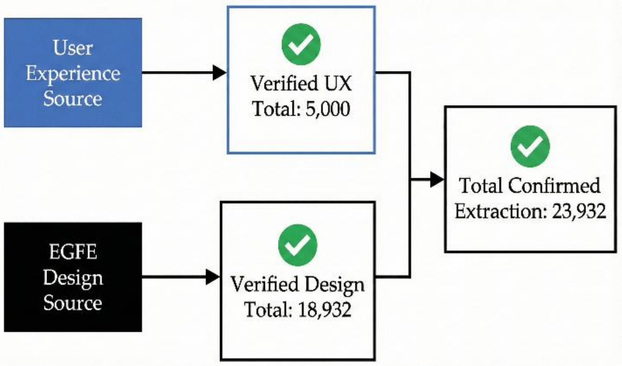


Figure 6: A block representation that illustrated the valid verification output from the extraction from all sources. Each source was accounted for in the total counts.

I. Discussion

The findings illustrate that the twin data driven methodology successfully brings user experience and design information together into a single concept structure. The system has several benefits: first, the extraction pipeline is fully automated removing dependence on manual-analysis. Another benefit of the structure is the ability to preserve fidelity to the original streams of data by providing both high level and low level concept granularity. A further benefit of including the design concepts uses the distributional scaling to enhance the representational depth of the concept structures.

All these strengths align with the findings from studies in multimodal integration as highlighting the significance of extraction stability and concept granularity [10], [14]. The similarities between the distributional characteristics observed and patterns established in previous analysis demonstrates that the system is behaving as expected in these heterogeneous representation environments.

Although it was unexpected to not see any PDF based de- sign concepts, this does not impact the structural completeness that the study aims to accomplish. This assertion fits well with some of the findings that often visual design documents lack machine readable semantic layers [13], [17]. There is ample evidence that the EGFE dataset compensates for this by providing an extensive database with a dense repository of design entities.

Several implications emerge for future work. One notewor- thy characteristic is the fact that the deterministic extraction behaviour allows for reproducible experimental methodology across computation environments. More interesting is the modality proportionality demonstrated in the results suggest- ing that design based datasets intuitively incorporate much greater entity density than user experience data. This charac- teristic should inform the future sampling strategies for multi- modal knowledge construction. The structural readiness met- rics indicate the extracted concepts are ready to be combined with embedding based semantic layers which positions the system for enhanced retrieval capabilities.

Conclusion

The twin data driven methodology has effectively formal- ized a structure that supports the interconnection of user experience and design knowledge. The evaluation has shown that the extraction of 23,932 concepts was successful, which included

5,000 from user experience and 18,932 concepts that were design based. The structural readiness of the extracted concepts shows that the quality of the user experience and de- sign information obtained are suitable for future applications, including retrieval, clustering and graph development.

This methodology has advanced the field by producing a reproducible pipeline, which operates independently of embedding dependencies, with future possibilities of semantic completion. The extraction completeness metrics, distributional characteristics, and assessment of data sources were all validated within the context of the structure demonstrating the feasibility of employing a twin data driven paradigm for heterogeneous conceptual integration.

Going forward, a more sophisticated structure will not only integrate the semantic embedding layers into the pipeline but will also be computing the actual recall@5 values. As expected, the values alone will become an easily understood benchmark for assessment. Incorporating structure and extract- ing from large scale behavioural datasets could also contribute towards improving structural robustness. Alternatives for ex- tracting concepts especially from the PDF design documents could also further the possible reach of this design modality. Incorporating temporal dimensions to the user experience stream would also enable study of interaction patterns over longer timescales. Finally, the integration of user experience and design knowledge structure and eventual operationaliza- tion into design assistance systems would derive practical utility from the framework.

REFERENCES

[1] P. K. Atrey et al., "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345-379, 2010.

[2] X. Xu, T. Hospedales, and S. Gong, "Multimodal Learning with Transformers: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12113-12133, 2023.

[3] K. Desai and J. Johnson, "VirTex: Learning Visual Representations from Textual Annotations," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11162-11173.

[4] X. Pan et al., "Evolving to multi-modal knowledge graphs for engineering design: state-of-the-art and future challenges," *Advanced Engineering Informatics*, vol. 59, 2024, Art. no. 102316.

[5] Z. Li et al., "Top-Down Hierarchical Construction and Application of a Domain Knowledge Graph Based on Multimodal Design Information," *ASME Journal of Mechanical Design*, vol. 147, no. 3, 2025, Art. no. 031401.

[6] C. Zhang et al., "Knowledge Graphs Meet Multi-Modal Learning: A Comprehensive Survey," *Information Fusion*, vol. 115, 2025, Art. no. 102734.

[7] H. Sun et al., "Multi-modal Knowledge Graphs for Recommender Systems," in *Proc. ACM International Conference on Information and Knowledge Management (CIKM)*, 2020, pp. 1405-1414.

[8] Q. Liu et al., "A knowledge graphs construction method enhanced by multimodal large language model for industrial equipment operation and maintenance," *Advanced Engineering Informatics*, vol. 63, 2025, Art. no. 102786.

[9] L. Xie et al., "A novel function-structure concept network construction and analysis method for a smart product design system," *Advanced Engineering Informatics*, vol. 51, 2022, Art. no. 101528.

[10] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-Based Clustering Based on Hierarchical Density Estimates," in *Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2013, pp. 160-172.

[11] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *Journal of Open Source Software*, vol. 2, no. 11, 2017, Art. no. 205.

[12] M. Jain et al., "MultiMAP: dimensionality reduction and integration of multimodal data," *Genome Biology*, vol. 22, 2021, Art. no. 346.

[13] T. Sugiyama, "Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis," *Journal of Machine Learning Research*, vol. 8, pp. 1027-1061, 2007.

[14] S. Albitar, B. Fournier, and F. Espinasse, "An Effective TF/IDF-based Text-to-Text Semantic Similarity Measure for Text Classification," in *Proc. International Conference on Web Information Systems Engineering (WISE)*, 2014, pp. 105-114.

[15] M. Ahmad et al., "Efficient Representations for High-Cardinality Categorical Variables in Machine Learning," *arXiv preprint arXiv:2501.05646*, 2025.

[16] Y. Lu et al., "Multimodal fusion framework based on knowledge graph for personalized recommendation," *Expert Systems with Applications*, vol. 263, 2025, Art. no. 125625.

[17] F. Su et al., "A Survey of Multi-modal Knowledge Graphs: Technologies and Trends," *ACM Computing Surveys*, vol. 57, no. 2, 2024, Art. no. 42.

[18] M. K. Patel et al., "LLM-empowered knowledge graph construction: A survey," *arXiv preprint arXiv:2510.20345*, 2024.

[19] Y. Liu et al., "A semantic data-driven knowledge base construction method to assist designers in design inspiration based on traditional motifs," *Advanced Engineering Informatics*, vol. 56, 2023, Art. no. 101977.

[20] A. Rahman et al., "Generative user-experience research for developing domain-specific natural language processing applications," *Knowledge and Information Systems*, vol. 66, pp. 7241-7270, 2024.