

# Multimodal Entity Indexing: A Conceptual Framework for Autonomous Visual Entities in Generative Engine Optimization

**Author:** Tumengkol, S.S.

**Affiliation:** PahamSEO Research Institute

**Document Type:** Draft Academic Journal Manuscript

**Target Publication:** PahamSEO Research Institute Monograph / Academic Journal Series

## Abstract

The rapid evolution of Large Language Models (LLMs) and Large Multimodal Models (LMMs) has transformed information retrieval from document-centric indexing toward multimodal semantic representation. This conceptual paper proposes **Multimodal Entity Indexing (MEI)** as a framework for understanding visual assets as knowledge-bearing entities within Generative Engine Optimization (GEO). The framework consists of **Autonomous Visual Entity (AVE)**, **Information-to-Pixel Ratio (IPR)**, and **Tri-Anchoring Context System (TACS)**. Rather than presenting empirical proof, the paper establishes a theoretical foundation and a structured research agenda for future multimodal retrieval studies.

**Keywords:** Generative Engine Optimization, Multimodal Embedding, Autonomous Visual Entity, Information-to-Pixel Ratio, Tri-Anchoring Context System, Semantic Search

## 1. Introduction

The landscape of information retrieval is undergoing a paradigm shift, transitioning from lexical, keyword-based search toward deep multimodal semantic representation. The emergence and rapid evolution of Large Language Models (LLMs) and Large Multimodal Models (LMMs) have fundamentally altered how search systems parse, index, and retrieve information.

Traditionally, visual assets (e.g., images, diagrams, infographics) were treated merely as decorative or supporting media attached to textual content. However, in modern AI-driven search environments, visual media carry independent semantic weight. This section introduces **Multimodal Entity Indexing (MEI)** as a conceptual framework designed to reposition visual assets as standalone semantic entities within Generative Engine Optimization (GEO).

## 2. Literature Review

To establish a rigorous theoretical grounding, this framework synthesizes concepts across several domain literatures:

- **Information Retrieval & Lexical-to-Semantic Search:** Historical transition from TF-IDF and inverted index systems to dense vector representations.

- **Knowledge Graphs & Entity-Centric Retrieval:** Modeling knowledge nodes and relationships rather than isolated string queries.
- **Vision–Language Models & CLIP:** Joint embedding spaces that enable cross-modal semantic alignment between text and pixel matrices.
- **Large Multimodal Models (LMMs):** The synthesis of visual reasoning in modern generative engines (e.g., GPT-4o, Google Gemini).
- **Generative Engine Optimization (GEO):** The emerging discipline of optimizing digital content for synthesis by generative search systems.

*Research Gap:* Existing GEO literature primarily focuses on textual optimization, leaving a significant gap regarding the systematic conceptualization of visual entities within multimodal retrieval environments.

### 3. Proposed Conceptual Framework

**Multimodal Entity Indexing (MEI)** serves as an overarching conceptual framework that integrates three core components:

| Component                                  | Core Concept         | Function / Definition                                                                                                                            |
|--------------------------------------------|----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Autonomous Visual Entity (AVE)</b>      | Semantic Identity    | Defines the visual asset as an independent, knowledge-bearing entity capable of standalone semantic representation in a multimodal vector space. |
| <b>Information-to-Pixel Ratio (IPR)</b>    | Semantic Density     | A theoretical model quantifying the density of factual and semantic information contained within a visual asset relative to its pixel footprint. |
| <b>Tri-Anchoring Context System (TACS)</b> | Contextual Alignment | Explains contextual semantic alignment across three textual boundary                                                                             |

| Component | Core Concept | Function / Definition                                    |
|-----------|--------------|----------------------------------------------------------|
|           |              | anchors: Filenames, Deep Alternative Text, and Captions. |

## 4. Conceptual Discussion

The discussion analyzes the broader implications of MEI across several dimensions:

- **Shift to Entity-Centric Optimization:** Moving beyond page-level SEO toward visual entity optimization.
- **Semantic Information Density:** The role of IPR in minimizing visual noise while maximizing factual signal in multimodal embeddings.
- **Contextual Alignment & Hallucination Reduction:** How TACS stabilizes the vector positioning of visual assets, strengthening their integration into multimodal Knowledge Graphs.
- **Generative Search Integration:** The mechanism by which AVEs are cited and synthesized in AI Overviews and generative answers.

*Note:* This paper explicitly positions these concepts as theoretical propositions intended for future empirical validation.

## 5. Implications and Future Research

### 5.1 Implications

- **Generative Engine Optimization (GEO):** Provides a structured framework for optimizing non-textual assets.
- **Digital Publishing & Information Architecture:** Offers guidance on metadata structuring, image compression, and visual content strategy.
- **AI & Retrieval Systems:** Offers insights into multi-agent and LMM retrieval architectures.

### 5.2 Future Research Agenda

- Empirical validation and mathematical refinement of the **IPR** metric.
- Experimental evaluation of **TACS** under controlled CLIP/LMM retrieval scenarios.
- Longitudinal studies tracking **Autonomous Visual Entities** across generative search engine results.

- Cross-domain testing and extension toward video, audio, and agentic retrieval environments.

## 6. Conclusion

The proposed **Multimodal Entity Indexing (MEI)** framework establishes a conceptual bridge between multimodal representation learning and Generative Engine Optimization. By formalizing AVE, IPR, and TACS, this paper contributes a structured theoretical foundation and a comprehensive research agenda to guide future empirical investigations in AI-driven retrieval systems.

## Statements & Declarations

- **Acknowledgements:** The author acknowledges the PahamSEO Research Institute for providing the conceptual foundation and research environment underlying this manuscript.
- **Conflict of Interest:** The author declares no competing financial or personal interests.
- **Data Availability Statement:** Data sharing is not applicable to this article as no public experimental datasets were generated or analyzed for this conceptual manuscript.
- **Author Contributions:** Conceptualization, methodology, framework development, writing, review, and editing: Tumengkol, S.S.