Review

Improving Information Communication in Emerging 6G Scenarios: A Review of Semantic Communications for the Future Internet

Evelio Astaiza Hoyos , Héctor Fabio Bermúdez-Orozco * , Nasly Cristina Rodriguez-Idrobo

Posted Date: 24 February 2026

doi: 10.20944/preprints202602.1442.v1

Keywords: semantic communications; Shannon information theory; deep joint source–channel coding; generative semantic communications; task-oriented communications; future internet and 6G networks

 doi:10.20944/preprints202602.1442.v1

Review

Improving Information Communication in Emerging 6G Scenarios: A Review of Semantic Communications for the Future Internet

Evelio Astaiza Hoyos [1], Héctor Fabio Bermúdez-Orozco [1,*] and Nasly Cristina Rodriguez-Idrobo [2]

[1] Electronic Engineering Programme, Faculty of Engineering, University of Quindío, Armenia 630004, Quindío, Colombia

[2] Occupational Health and Safety Program, Faculty of Health Sciences, University of Quindío, Armenia 630004, Quindío, Colombia

* Correspondence: hfbermudez@uniquindio.edu.co; Tel.: +057 3206671107

Abstract

The evolution of future Internet and sixth-generation (6G) networks is driving a paradigm shift from classical bit-centric communication toward meaning-aware and task-oriented communication models. Traditional information theory, while fundamental for ensuring reliable symbol transmission, does not account for semantic relevance or task effectiveness, which are critical for emerging applications such as autonomous systems, immersive services, and ultra-low-latency communications. This article presents a comprehensive review of Semantic Communications (SemCom) from a future Internet perspective. The review systematically analyses representative extensions of classical information theory aimed at quantifying semantic information, including semantic information measures, semantic channel capacity, and semantic rate–distortion formulations. In addition, the main mathematical and computational frameworks enabling practical semantic communication systems are examined, including the Information Bottleneck principle, learning-based end-to-end communication architectures, and reinforcement learning approaches for task-oriented optimization under network constraints. The review further discusses the role of semantic metrics, contextual modelling, and task-driven performance evaluation in the design of semantic-aware communication systems. The analysis identifies key open challenges, particularly the lack of a unified theoretical framework, the need for robust and context-aware semantic performance metrics, and the integration of semantic awareness into network-level design. Overall, this review highlights Semantic Communications as a promising paradigm for future Internet and 6G networks, where communication efficiency is increasingly determined by semantic relevance and task effectiveness rather than bit-level fidelity alone.

Keywords: semantic communications; Shannon information theory; deep joint source–channel coding; generative semantic communications; task-oriented communications; future internet and 6G networks

1. Introduction

The rapid evolution of future Internet architectures and sixth-generation (6G) communication systems is challenging the traditional principles that have governed communication system design for decades. While classical information theory provides a rigorous and mathematically elegant framework for reliable symbol transmission [1], emerging applications increasingly demand communication systems capable of interpreting information, exploiting context, and supporting task-oriented objectives under stringent latency, reliability, and energy constraints. These requirements expose the limitations of conventional bit-centric communication models, which deliberately exclude semantic and pragmatic aspects of communication [2], motivating the exploration of new paradigms that explicitly account for semantic relevance and task effectiveness, particularly in the context of

 doi:10.20944/preprints202602.1442.v1

future Internet and 6G networks [3]. This article provides a comprehensive review of the theoretical foundations and enabling frameworks of Semantic Communications, with a focus on their implications for future Internet architectures.

1.1. Motivation: Beyond the Shannon Paradigm

Traditional communication systems have been predominantly designed under the framework of Claude E. Shannon's information theory, whose primary objective is the accurate and efficient transmission of symbols or bits over a noisy channel [1]. Within this framework, information is quantified in terms of uncertainty reduction based on the statistical occurrence of symbols, and communication performance is evaluated independently of the meaning conveyed by those symbols. This abstraction corresponds to the technical level of communication (Weaver's Level A) [2] and has enabled the development of modern digital communication systems by establishing fundamental limits such as channel capacity and rate–distortion trade-offs.

A crucial aspect of Shannon's formulation is the explicit exclusion of semantic and pragmatic aspects of communication from the engineering problem of reliable transmission. As originally stated, the semantic meaning of messages is considered irrelevant to the technical problem addressed by information theory [2]. While this assumption has been extraordinarily successful for conventional communication scenarios, it becomes increasingly restrictive in emerging environments where performance is inherently tied to interpretation, decision-making, and task execution.

The telecommunications landscape is now undergoing a rapid transformation driven by 6G networks and future Internet applications, including autonomous systems, intelligent robotics, immersive environments, remote industrial control, and the Internet of Everything [4]. These applications generate massive volumes of data and impose stringent requirements on latency, reliability, and energy efficiency, while their success depends fundamentally on the correct interpretation of information rather than on exact symbol reconstruction.

In such scenarios, the brute-force transmission of raw data—despite being optimized according to Shannon's principles—approaches both theoretical and practical limits, leading to inefficiencies in bandwidth usage, latency, and energy consumption [4] . This intrinsic limitation of the classical paradigm motivates a fundamental shift in perspective: from the traditional engineering question of *"how to transmit?"* to the meaning-oriented question of *"what should be transmitted?"*[5]. From a future Internet and network design perspective, this shift challenges the foundations of communication architectures by requiring efficiency metrics that account for semantic relevance and task effectiveness rather than solely throughput and error rates.

Semantic Communications (SemCom) address this challenge by proposing the selective transmission of information that is relevant or meaningful for the receiver or the task at hand, while discarding semantically irrelevant details [4]. By prioritizing meaning over syntactic accuracy, SemCom promises significant reductions in communication resource consumption and improved system efficiency, particularly in resource-constrained and ultra-low-latency environments.

1.2. Semantic Communications versus Technical Communications

In contrast to technical communication, Semantic Communications (SemCom) are defined as a paradigm centred on the successful transmission and interpretation of meaning rather than the exact replication of individual symbols or bits [2,6]. This perspective departs from Shannon's syntactic abstraction of information [4], where semantics are deliberately excluded from the communication model, and instead aligns with emerging theoretical extensions that explicitly incorporate semantic relevance into information processing. Early efforts to formalize semantic information emphasized the relationship between transmitted symbols, contextual knowledge, and meaning consistency [7], while more recent contributions propose quantitative frameworks for semantic information measures, semantic channel capacity, and semantic rate–distortion formulations that extend classical information-theoretic principles to the semantic domain [7,8].

 doi:10.20944/preprints202602.1442.v1

Consequently, the ultimate objective of a SemCom system is not perfect bit-by-bit reconstruction but semantic equivalence or task-relevant interpretation at the destination. This objective is often operationalized through learning-based architectures that jointly optimize representation, compression, and inference under task constraints [9,10]. In contrast, technical communication focuses on accurate physical signal delivery and evaluates performance using syntactic metrics such as bit error rate (BER) or symbol error rate (SER), with achievable transmission rates bounded by Shannon capacity [1]. Although this framework has enabled decades of reliable digital communication, it does not account for whether the received information is meaningful or sufficient for task execution.

The fundamental distinction between these paradigms therefore lies in their optimization objectives and performance metrics. While technical communication seeks to maximize spectral efficiency and symbol fidelity, semantic communication aims to preserve meaning and support task execution, employing metrics related to semantic similarity, task accuracy, contextual consistency, or overall communication effectiveness [3,11]. This shift implies a redefinition of the communication chain as an end-to-end intelligent system, often modelled through deep neural networks, Information Bottleneck principles, or reinforcement learning–based task optimization [12,13]. At the network level, semantic awareness directly influences resource allocation, latency management, and cross-layer optimization strategies, particularly in AI-native 6G architectures and Future Internet environments where efficiency is increasingly determined by task success probability rather than bit-level fidelity [14,15].

1.3. Weaver's Levels of Communication (Contextualizing Levels B and C)

The conceptual framework proposed by Warren Weaver in 1949 is fundamental for contextualizing the scope of semantic communications, as it distinguishes three levels of communication problems [2]:

1.3.1. Level A (Technical Problem): This level asks, "How accurately can the symbols of communication be transmitted?" It addresses the fidelity of physical signal transmission, independently of content. Shannon's mathematical theory of communication provides the foundational framework for this level [1].

1.3.2. Level B (Semantic Problem): This level asks, "How precisely do the transmitted symbols convey the desired meaning?" It focuses on the correspondence between the sender's semantic intention and the receiver's interpretation, addressing challenges such as ambiguity, contextual dependency, and knowledge consistency. Early theoretical efforts to formalize semantic information emphasized precisely this relationship between symbols, contextual knowledge, and meaning coherence [7], laying the groundwork for contemporary semantic information measures and semantic channel models.

1.3.3. Level C (Effectiveness or Pragmatic Problem): This level asks, "How effectively does the received meaning affect conduct in the desired way?" It evaluates communication in terms of its impact or final outcome, particularly regarding goal achievement or behavioural influence at the receiver.

Semantic Communications (SemCom) directly target Level B by seeking engineering solutions to transmit meaning efficiently and accurately. However, there exists an inseparable and motivating connection to Level C. Many of the applications driving SemCom research—such as remote control systems, autonomous driving, immersive services, and AI-enabled decision-making—are inherently goal-oriented, where success is measured not by symbol fidelity but by task effectiveness and action accuracy [9,15,16]. In this context, semantic relevance becomes a performance-determining factor in emerging 6G and Future Internet architectures [15].

The analysis of this hierarchical structure reveals that although SemCom primarily addresses Levels B and C, it cannot disregard Level A. Technical precision at the physical layer inevitably influences the ability to preserve meaning and, consequently, to achieve the desired pragmatic outcome [9,10]. This hierarchical interdependence constitutes the underlying principle that justifies

 doi:10.20944/preprints202602.1442.v1

joint and end-to-end optimization approaches, where Level B optimization must be conditioned on Level C objectives while remaining constrained by Level A limitations. Such cross-layer integration is particularly relevant in AI-native 6G networks, where semantic awareness may guide resource allocation, latency management, and adaptive transmission strategies [14,15].

This article is structured as follows. Section II examines the theoretical foundations of semantic information theory, exploring attempts to quantify meaning through semantic information measures, semantic rate–distortion theory, and semantic channel capacity, and contrasting these concepts with their classical Shannon counterparts [7,8].

Section III reviews the key mathematical and computational approaches that enable the practical realization of semantic communication systems. This includes the Information Bottleneck principle, deep learning architectures such as autoencoders and transformer-based models knowledge representation frameworks, and reinforcement learning–based optimization techniques for task-oriented communication under network constraints.

Section IV identifies and analyses the fundamental mathematical challenges and open research problems currently facing semantic communications, highlighting the need for unified theoretical models, robust semantic performance metrics, and scalable network-level integration.

Finally, Section V concludes the review by summarizing the current state of mathematical foundations in semantic communications and outlining future theoretical research directions toward fully semantic-aware and AI-native Future Internet and 6G systems.

2. Foundations of Semantic Information Theory

The mathematical formalization of meaning constitutes the most significant challenge for semantic information theory. Although a unified and universally accepted framework has not yet emerged, several rigorous extensions of Shannon's foundational concepts have been proposed to incorporate semantic relevance into information-theoretic analysis [1,7]. These efforts aim to move beyond purely syntactic representations and address the quantification of meaning within communication systems.

2.1. Semantic Information Measures: Quantifying Meaning

The primary difficulty in quantifying "meaning" lies in its inherently complex, subjective, and context-dependent nature. Unlike syntactic information, which is objectively measured through statistical entropy, semantic information depends on prior knowledge, contextual models, and interpretative frameworks shared between communicating agents [7].

Early attempts to formalize semantic information were grounded in logical probability rather than Shannon's statistical probability. Xin and Fan [17] proposed measuring the semantic information $I(e)$ of a proposition e as inversely proportional to its logical probability $m(e)$, representing the plausibility of e being true across possible worlds:

$$I(e) = -\log m(e) \quad (1)$$

However, this formulation produced a well-known paradox: logical contradictions, having $m(e) = 0$, were assigned infinite information [17]. This limitation highlighted the difficulty of directly translating logical semantics into a robust engineering framework and motivated the search for alternative formulations.

Subsequently, alternative formulations were introduced to explicitly formalize task relevance within semantic information theory. Among these, Xiao Chang Lu [18] proposed the *G*-measure of semantic information, which links semantic content to belief updating under task-dependent hypotheses.

Let θ denote the semantic class associated with hypothesis y. The prior probability that an instance X belongs to the semantic class θ is defined as:

$$T(\theta) \equiv P(X \in \theta) \quad (2)$$

Similarly, the conditional truth function is defined as:

$$T(\theta \mid x) \equiv P(X \in \theta \mid X = x) \quad (3)$$

 doi:10.20944/preprints202602.1442.v1

The semantic information G conveyed by hypothesis y (or its associated class θ about a specific instance x is then defined as:

$$\mathrm{I}(x;\theta) = \log \frac{T(\theta)}{T(\theta|x)} \quad (4)$$

This formulation captures the idea that semantic information is intrinsically related to the extent to which an observation x updates the prior belief $T(\theta)$ associated with a hypothesis y. A condition where $T(\theta \mid x) < T(\theta)$ may even indicate semantic misinformation, reflecting the qualitative impact of semantics beyond purely statistical interpretation [19,20].

The quantification of semantic uncertainty—often referred to as semantic entropy—has led to multiple definitions, reflecting the theoretical fragmentation of the field. A rigorous formulation is based on the concept of synonymous typical sets [21], This theory introduces a synonym mapping function $f: U \to \tilde{U}$, where multiple syntactic sequences $u \in U$ may map to the same semantic meaning $\tilde{u} \in \tilde{U}$. The semantic entropy $H_s(\tilde{U})$ is then defined over the distribution of semantic meanings $\tilde{u}$:

$$H_s(\tilde{U}) = -\sum_{\tilde{u}\in\tilde{U}} P(\tilde{u}) \log P(\tilde{u}) \quad (5)$$

A key result established within this framework is that semantic entropy is always less than or equal to Shannon entropy, $H_s(\tilde{U}) \leq H(U)$. This rigorously demonstrates that exploiting semantic redundancy—such as synonymy—can reduce source uncertainty beyond what is achievable through the elimination of purely statistical redundancy via Shannon source coding. Other proposals include clustering-based semantic entropy formulations, such as:

$$\bar{E}(T) = -\sum_{j} \frac{n_j}{n} \log \frac{n_j}{n} \quad (6)$$

Where semantic uncertainty is defined according to grouping structures rather than symbol frequency distributions.

The diversity of approaches—ranging from logical frameworks and knowledge-base (KB) models to fuzzy probability formulations—for measuring semantic information underscores that, unlike Shannon's theory, which is grounded in a unique set of statistical axioms, the concept of "meaning" in engineering contexts is inherently relative [22].

Knowledge-base–driven measures further imply that semantic uncertainty is not an intrinsic property of the source itself, but rather a relational property dependent on the shared knowledge between the transmitter and the receiver. Table 1 summarizes the fundamental distinctions between the Shannon and Semantic paradigms.

Table 1. Comparison between Shannon and Semantic Information Measures.

Characteristic	Shannon Information (Level A)	Semantic Information (Level B/C)
Primary Objective	Accurate transmission of symbols/bits	Accurate transmission and interpretation of meaning; task effectiveness
Probabilistic Basis	Statistical (frequency of occurrence)	Logical, fuzzy, task-oriented, knowledge-based, synonym-based, clustering-based
Core Measure	Shannon Entropy $H(X)$	Semantic Entropy (multiple definitions: $(H_S(e),\ H_{C_j}(\subseteq),$ $H_s(\tilde{U}), \bar{E}(T),$ KLE, etc.)
Mutual Information	$I(X;Y) = H(X) - H(X/Y)$	
Key Challenge	Approaching channel capacity limits	Universal definition/quantification of "meaning"; context and knowledge dependence
Example Metrics	BER, SER, Bit rate, Channel capacity C	G-measure, Semantic similarity (e.g., BERTScore), Task accuracy, H_s, semantic capacity C_s, semantic rate–distortion $R_s(D)$

 doi:10.20944/preprints202602.1442.v1

Table 1 highlights the conceptual and operational divergence between Shannon's syntactic paradigm and the emerging semantic communication framework. While Shannon information theory provides a mathematically rigorous foundation for symbol transmission and channel capacity optimization, it deliberately abstracts away meaning and contextual interpretation. In contrast, semantic information paradigms redefine communication objectives by incorporating task relevance, contextual knowledge, and interpretative alignment between transmitter and receiver. This shift implies not only new entropy formulations and performance metrics but also a transformation in optimization strategies, moving from layer-centric reliability to end-to-end, goal-oriented intelligent systems.

2.2. Semantic Rate–Distortion Theory

Semantic rate–distortion theory extends the fundamental limits of data compression into the domain of meaning, seeking the optimal trade-off between transmission rate R and semantic fidelity, quantified through semantic distortion D [1,21]. While classical rate–distortion theory defines distortion at the symbol level, semantic rate–distortion reframes the problem by evaluating preservation of meaning rather than syntactic accuracy.

The most critical component of this framework is the definition of an appropriate semantic distortion measure $d_s(s, \hat{s})$, since traditional metrics such as Mean Squared Error (MSE) or Bit Error Rate (BER) are inadequate for capturing human perception, contextual relevance, or task effectiveness. Semantic distortion must instead evaluate similarity at the level of meaning.

2.2.1 Embedding-Based Metrics (Text): Modern approaches leverage deep learning models to generate contextual vector representations (embeddings) of meaning.

- BERTScore: This metric compute cosine similarity between token embeddings of a candidate sentence and a reference sentence. Precision P, recall R, and the averaged F1-score are derived from maximum token-wise similarities. The F1-score is defined as [23]:

$$F1_{\text{BERT}} = \frac{2P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \quad (7)$$

- Sentence-BERT (S-BERT): Siamese network architectures are used to generate sentence-level embeddings. The similarity between two sentences u and v is computed via cosine similarity of their embedding vectors [24]:

$$\text{Similitud}(u, v) = \frac{u \cdot v}{|u||v|} \quad (8)$$

These embedding-based measures approximate semantic alignment by evaluating geometric proximity in representation space.

2.2.2 Perceptual and Task-Based Metrics (Image/Video): For visual data, perceptual metrics such as LPIPS (Learned Perceptual Image Patch Similarity) or distances between feature vectors extracted from deep neural networks (DNNs) are employed [10,25]. These metrics capture perceptual-level similarity rather than pixel-wise fidelity.

Alternatively, semantic distortion may be quantified through downstream task performance, such as classification accuracy degradation or control cost variation in multi-agent systems [26,27]. In such settings, distortion is directly linked to decision-making effectiveness, aligning with the Level C (pragmatic) dimension of Weaver's hierarchy.

The semantic rate–distortion function $R_s(D)$ defines the minimum rate required to represent a source S with reconstruction $\hat{S}$ such that the expected semantic distortion does not exceed a threshold D:

$$R_s(D) = \min_{p(\hat{S}|S):E[d_s(s,\hat{s})]\leq D} I\left(S;\hat{S}\right) \quad (9)$$

Where $I\left(S;\hat{S}\right)$ denotes mutual information between source and reconstruction. Zhang and Niu demonstrated that $R_s(D) \leq R(D)$, where $R(D)$ is the classical Shannon rate–distortion function. This inequality formally indicates that exploiting semantic redundancy—such as synonymy—enables more efficient compression for a given level of meaning fidelity compared to purely syntactic

 doi:10.20944/preprints202602.1442.v1

compression schemes [21]. In other words, when meaning equivalence is permitted, fewer transmitted bits may be required to achieve acceptable task performance.

In practice, semantic coding is implemented using deep learning–based Joint Source-Channel Coding (JSCC), particularly through autoencoder architectures that directly optimize a semantic distortion-aware loss function in an end-to-end manner [10,28]. These models implicitly approximate the optimal trade-off described by $R_s(D)$ without requiring explicit analytical solutions [9,10].

Such end-to-end semantic optimization is especially relevant in AI-native 6G networks, where communication objectives may be dynamically adapted to task requirements and environmental conditions [14,15].

Table 2 illustrates that semantic distortion is inherently dependent on data modality and task definition. Unlike classical distortion measures defined over symbol differences, semantic distortion metrics evaluate representation-level similarity, perceptual alignment, or task performance degradation. Embedding-based metrics approximate semantic equivalence in representation space, perceptual metrics capture human-aligned similarity for visual data, and task-oriented formulations quantify the functional impact of information on downstream objectives. This diversity reinforces the notion that semantic fidelity is inherently context-dependent and cannot be captured through a single universal distortion metric.

Table 2. Representative Semantic Distortion Metrics.

Data Type	Metric	Mathematical Formulation / Concept	References
Text	BLEU	Weighted n-gram precision with brevity penalty	[29]
Text	BERTScore	Cosine similarity between contextual embeddings (BERT) with greedy matching; evaluated via precision, recall, and F1-score	$F1 = \frac{2P \cdot R}{P + R}$
Text	S-BERT Similarity	Cosine similarity between sentence-level embeddings (S-BERT)	$\text{Similitud}(u, v) = \frac{u \cdot v}{\lvert u \rvert \lvert v \rvert}$
Image/Video	LPIPS	Distance between deep feature representations (DNN activations) extracted from image patches; learned perceptual similarity metric	Perceptual learned metric
Task-Oriented	Control Cost	Accumulated cost in a control task using the received information as input	[26]
Task-Oriented	IB Distortion	Information Bottleneck Lagrangian: $I(X; Z) - \beta\, I(Z; Y)$	Compression–relevance trade-off

2.3. Semantic Channel Capacity

The concept of semantic channel capacity, denoted C_s, seeks to define the maximum rate at which semantic information can be reliably transmitted over a noisy channel [1,21]. It generalizes Shannon's classical channel capacity C by incorporating semantic equivalence into the decoding process rather than requiring strict symbol-level reconstruction.

Within the framework proposed by Zhang and Niu [21], based on synonym mapping, semantic channel capacity is defined in terms of semantic mutual information, denoted I_s (or I^s). Formally:

$$C_s = \max_{f_{xy}} \max_{p(x)} I_s(\tilde{X}; \tilde{Y}) \qquad (10)$$

where the maximization is performed over both the input distribution $p(x)$ and admissible joint synonym mappings f_{xy}, which map syntactic sequences to their corresponding semantic representations.

2.3.1 Semantic Channel Coding Theorem: The associated semantic channel coding theorem establishes that reliable semantic transmission—defined as the probability of semantic error tending to zero—is achievable if the semantic entropy rate of the source satisfies:

$$H_s(\tilde{U}) < C_s \qquad (11)$$

This condition parallels Shannon's classical channel coding theorem but replaces syntactic entropy with semantic entropy and classical mutual information with semantic mutual information.

 doi:10.20944/preprints202602.1442.v1

2.3.2 Theoretical Gain over Shannon Capacity: This theoretical gain over the classical Shannon limit arises from the exploitation of **semantic redundancy**, particularly synonymy. Because multiple syntactic symbol sequences may correspond to the same semantic meaning, channel-induced symbol errors do not necessarily result in semantic errors. If the semantic decoder can correctly map noisy syntactic sequences to the intended meaning, semantic reliability can be preserved even when the bit-level reconstruction is imperfect [21]. Consequently, semantically reliable communication may be achieved at bit rates exceeding the classical Shannon capacity, provided that semantic equivalence classes are properly defined and exploited at the receiver.

2.3.3 Structural Implications: This result suggests that transmission limits are not determined solely by the physical properties of the channel—such as signal-to-noise ratio—but also by the structural properties of the semantic source and its redundancy patterns. In other words, communication limits become jointly dependent on channel statistics and semantic structure.

However, the practical application and explicit quantification of C_s depend critically on the ability to formally define and operationalize semantic equivalence for complex data sources such as natural language, images, and multimodal streams. This remains an open research challenge, particularly in large-scale 6G and AI-native network environments where semantic representations are dynamically learned rather than pre-defined [15,30].

3. Key Mathematical and Computational Approaches

The implementation of semantic communication systems requires advanced mathematical tools capable of handling meaning extraction and nonlinear optimization, particularly because a complete analytical theory of semantic information remains under development. In practice, semantic communication relies on principled information-theoretic frameworks combined with learning-based approximations that enable tractable optimization in high-dimensional settings.

3.1. The Information Bottleneck Principle for Semantic Relevance

The Information Bottleneck (IB) principle, introduced by Tishby et al. [12], provides a rigorous information-theoretic framework for defining and extracting semantically relevant information. The core objective is to identify a compressed representation Z of an input variable X that preserves maximal information about a relevant variable Y(e.g., a class label, task output, or intended meaning), while minimizing the information retained about X itself.

Mathematically, this is formulated through the minimization of the following Lagrangian functional over conditional distributions $p(z \mid x)$:

$$\mathcal{L}[p(z \mid x)] = I(X;Z) - \beta I(Z;Y) \qquad (12)$$

Where $I(\cdot;\cdot)$ denotes mutual information, and $\beta \geq 0$ is a Lagrange multiplier controlling the trade-off between compression (minimizing $I(X;Z)$) and relevance (maximizing $I(Z;Y)$ [12,31].

Within the semantic communication framework, the IB principle provides an operational definition of semantics: information is relevant if it is predictive of the task outcome Y. In this context:

X represents the source signal, Y represents the communication objective (corresponding to Level C in Weaver's hierarchy), and Z is the compressed semantic representation to be transmitted.

The optimal semantic encoder $p(z \mid x)$ thus extracts only those features of X that are predictive of Y, discarding task-irrelevant variability. This aligns naturally with goal-oriented communication paradigms, where transmission efficiency is measured by task success rather than symbol reconstruction fidelity [3,32].

Direct optimization of the IB Lagrangian is typically intractable in high-dimensional settings. Therefore, variational approximations such as the Variational Information Bottleneck (VIB) are employed, where neural networks are used to approximate mutual information terms through computable bounds. These variational approaches enable scalable optimization within deep learning frameworks [9].

Extensions of the IB principle include:

 doi:10.20944/preprints202602.1442.v1

- Robust Information Bottleneck (RIB): Incorporates robustness constraints to handle channel noise and adversarial perturbations.
- Graph Information Bottleneck (GIB): Adapts the IB framework to structured data such as graphs, relevant for knowledge-based semantic representations.

Such extensions are particularly important in 6G and Future Internet scenarios, where communication occurs over noisy, heterogeneous, and dynamically adaptive environments [14,15].

3.2. Deep Learning as a Mathematical Tool

Deep Learning (DL) has become the dominant computational framework for implementing semantic communication systems. From a mathematical perspective, deep neural networks act as universal function approximators capable of learning highly nonlinear transformations required for semantic encoding and decoding [9].

In practical SemCom architectures, DL models implement:

- **Semantic encoders:** Mapping raw input X to compressed semantic representations Z.
- **Channel-aware modules:** Joint source-channel optimization.
- **Semantic decoders:** Reconstructing meaning or directly performing downstream inference tasks.

End-to-end architectures—particularly autoencoder-based Joint Source-Channel Coding (JSCC)—optimize semantic distortion objectives directly, effectively approximating the theoretical semantic rate–distortion trade-off derived in Section II [9,10].

Moreover, reinforcement learning–based approaches enable adaptive semantic transmission policies under dynamic network constraints, making them suitable for task-oriented 6G systems with strict latency and reliability requirements [27].

Figure 1 illustrates a generic deep learning–based semantic communication architecture, where the semantic encoder, channel model, and semantic decoder are trained jointly under a task-driven loss function.

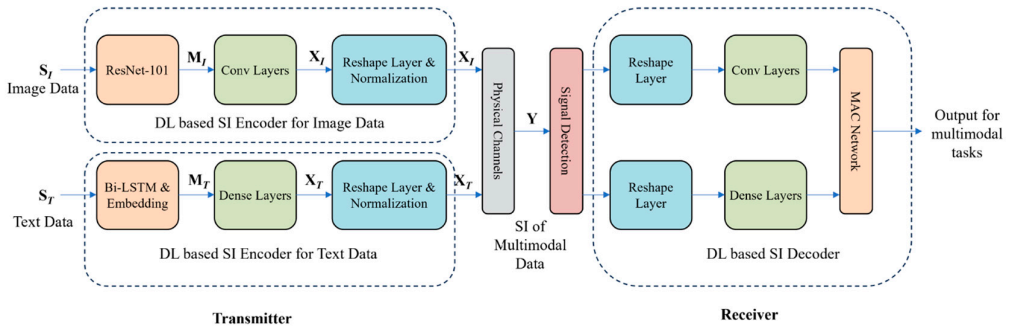

Figure 1. Semantic Transmission and Reception Chain. Source [33].

3.2.1. Autoencoders and Joint Source–Channel Coding (JSCC): Autoencoder architectures constitute the core of many practical semantic communication systems, implementing end-to-end Joint Source–Channel Coding (DeepJSCC). In this framework, the communication system is modelled as a single differentiable pipeline composed of:

- a semantic encoder $f_\theta(\cdot)$,
- a channel layer (e.g., additive Gaussian noise $y = (x + n)$,
- and a semantic decoder $g_\phi(\cdot)$.

The system is trained by minimizing a loss function $\mathcal{L}(s, \hat{s})$ that is directly tied to the desired semantic fidelity metric (e.g., BERTScore, LPIPS, or classification loss). Unlike classical communication systems based on source–channel separation, this approach jointly optimizes compression and robustness to channel impairments [10,28].

 doi:10.20944/preprints202602.1442.v1

Formally, the end-to-end objective can be expressed as:

$$min_{\theta,\emptyset} E_{s,n}\left[\mathcal{L}\left(s, g_{\emptyset}\left(h(f_{\theta}(s), n)\right)\right)\right] \quad (13)$$

where $h(\cdot)$ models the stochastic channel transformation.

A major advantage of DeepJSCC is that it implicitly learns representations that are resilient to channel noise and avoids the classical cliff effect, which characterizes traditional digital communication systems relying on strict source–channel separation [1]. In contrast to conventional coding schemes, semantic autoencoder-based systems degrade gracefully under adverse channel conditions, maintaining task-level performance even when bit-level fidelity is compromised.

This paradigm operationalizes the semantic rate–distortion trade-off discussed in Section II, enabling empirical approximation of $R_s(D)$ through gradient-based optimization [34].

3.2.2. Transformers and Attention Mechanisms

For sequential data such as text and speech, Transformer architectures play a central role due to their self-attention mechanism. Attention enables the model to dynamically weigh the relative importance of different elements within an input sequence when constructing contextualized representations.

The fundamental formulation of scaled dot-product attention is given by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (14)$$

where Q, K, and V denote the Query, Key, and Value matrices, respectively, d_k represents the dimensionality of the key vectors [33].

Through this mechanism, Transformers capture long-range semantic dependencies and contextual relationships, making them particularly suitable for semantic encoding tasks in natural language and multimodal communication systems.

In semantic communication frameworks, attention mechanisms facilitate context-aware compression by selectively emphasizing task-relevant tokens or acoustic features. When combined with end-to-end channel-aware training, Transformer-based encoders can learn semantic representations optimized jointly for meaning preservation and transmission efficiency [19,35].

3.3. Generative Semantic Communications (Gen-SemCom) Enabled by Diffusion Models

Generative Semantic Communications (Gen-SemCom) [36] represent an emerging paradigm driven by Generative Artificial Intelligence (GenAI), and are considered a key enabler for achieving the extreme efficiency targets envisioned for 6G networks. Unlike conventional semantic communication systems that reconstruct or classify received signals, Gen-SemCom reformulates decoding as a posterior inference problem.

In this framework, the receiver leverages a rich, pre-trained semantic distribution—referred to as a **Semantic Knowledge Base (SKB)**—to synthesize high-quality content from minimal semantic cues transmitted over the channel. Rather than transmitting full-resolution representations, the transmitter conveys compact semantic tokens or latent variables that condition the generative process at the receiver.

Formally, if S denotes the original semantic source and Z represents the transmitted semantic embedding, the receiver performs posterior inference of the form:

$$\widehat{S} \sim p(S \mid Z, SKB) \quad (15)$$

where the SKB encodes prior semantic knowledge learned through large-scale pre-training.

Diffusion models and other generative architectures approximate this conditional distribution by iteratively refining latent representations toward semantically plausible outputs. In contrast to deterministic reconstruction, the output is sampled from a learned semantic manifold, thereby exploiting prior knowledge to compensate for reduced transmitted information.

3.3.1. Complexity–Efficiency Trade-Off: This paradigm introduces a fundamental **complexity trade-off:**

- Channel complexity is reduced (fewer transmitted bits).
- Computational complexity at the receiver is increased.

 doi:10.20944/preprints202602.1442.v1

In other words, communication efficiency is achieved by shifting the burden from transmission resources (bandwidth, power) to computation resources (model inference, generative sampling). This shift is particularly aligned with AI-native 6G architectures, where edge/cloud intelligence and powerful receivers are expected to be available [14,15].

As illustrated conceptually in Figure 2, Gen-SemCom transforms communication into a cooperative inference process between transmitter and receiver, where meaning reconstruction is guided by learned semantic priors rather than strict symbol recovery.

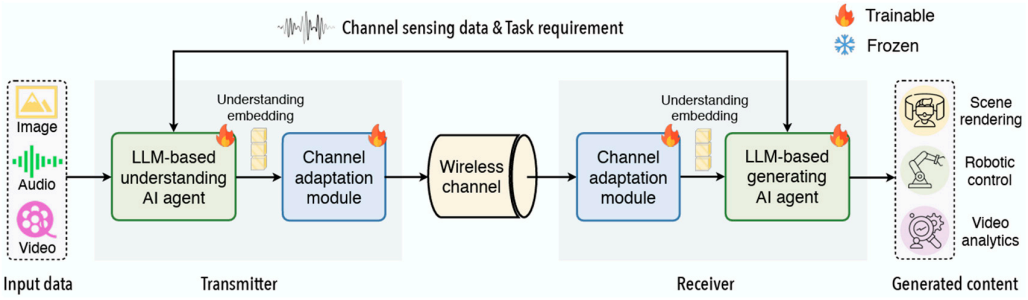

Figure 2. Semantic Communication System Architecture. Source [36].

3.3.2. Implications for 6G

Gen-SemCom aligns with several defining characteristics of 6G systems: AI-native network design, Ultra-low-latency goal-oriented communication, Extreme spectral and energy efficiency, Integration of communication and computation

By exploiting generative priors, Gen-SemCom potentially pushes semantic rate–distortion and semantic capacity limits beyond what deterministic encoders can achieve, especially for multimodal and high-dimensional data.

However, this paradigm also raises open challenges, including: Reliability guarantees under generative sampling, Hallucination risks in safety-critical applications, Synchronization of semantic knowledge bases and Standardization of semantic priors.

3.3.3. Mathematical Core of Gen-SemCom Diffusion Models

The mathematical core of Generative Semantic Communications (Gen-SemCom) lies in diffusion models, which are grounded in the theory of Stochastic Differential Equations (SDEs) and Ordinary Differential Equations (ODEs).

- Forward Diffusion Process Modeling: The forward diffusion process—defined as the progressive injection of Gaussian noise into the original data x_0—can be described by a stochastic differential equation (SDE), which in its general form is expressed as:

$$dx_t = f(t)\, x_t dt + g(t) d\omega_t \tag{16}$$

where x_t denotes the data at time t, $f(t)x_t dt$ represents the drift coefficient, $g(t)d\omega_t$ is the diffusion coefficient, $d\omega_t$ denotes a Wiener process. A commonly used infinitesimal formulation for noise injection is:

$$dx_t = -\frac{1}{2}\beta(t)x_t dt + \sqrt{\beta(t)}d\omega_t \tag{17}$$

where $\beta(t)$ defines the variance schedule of the injected noise.

This forward process gradually transforms structured data into pure noise, typically converging to a Gaussian distribution $x_T \sim N(0, I)$.

- Decoding as Reverse Inference (Reverse Process): Semantic decoding in Gen-SemCom is achieved through the reverse diffusion process, which is learned to progressively remove noise and reconstruct x_0 from a noisy representation.

In classical diffusion models, reconstruction begins from pure Gaussian noise $x_T \sim \mathcal{N}(0, \mathcal{I})$. In Gen-SemCom, however, the reverse process is conditioned on a compressed and noisy semantic latent representation Z, transmitted through the communication channel.

 doi:10.20944/preprints202602.1442.v1

To ensure robustness, the wireless channel noise is explicitly mapped to the forward diffusion process. This design choice allows the receiver to employ the learned reverse diffusion process for progressive denoising, thereby reconstructing signals with high perceptual fidelity—even under extremely noisy channel conditions.

Thus, communication becomes an inference problem:

$$\hat{x}_0 \sim p(x_0 \mid Z) \tag{18}$$

where the generative prior is shaped by the semantic knowledge base and diffusion dynamics.

- Hybrid Optimization for Bandwidth and Perceptual Fidelity: The optimization of Gen-SemCom systems requires a hybrid loss-function design. For example, a Variational Autoencoder (VAE)–based loss may be used to enforce compression and bandwidth constraints, while a guidance-based or perceptual loss is employed to enhance generation quality.

This joint objective enables simultaneous optimization of transmission efficiency and perceptual fidelity, aligning semantic reconstruction quality with communication resource constraints.

- Technical Pillars for 6G Deployment: The successful deployment of Gen-SemCom in 6G environments relies on several key technical pillars:
  - **Conditional Diffusion** – Ensures that the generative process is aligned with the intended semantic message or task objective.
  - **Efficient Diffusion** – Reduces computational latency and inference complexity to meet real-time communication requirements.
  - **Generalized Diffusion** – Enables adaptability across multiple data modalities and application domains.

Together, these elements support the transition toward AI-native communication systems in which communication, inference, and content generation are tightly integrated.

3.4. Knowledge Representation and Reasoning

Semantic interpretation is intrinsically dependent on context and on the background, knowledge shared between communicating agents. For this reason, knowledge representation and reasoning techniques constitute essential complements to deep learning–based models, as illustrated conceptually in Figure 3.

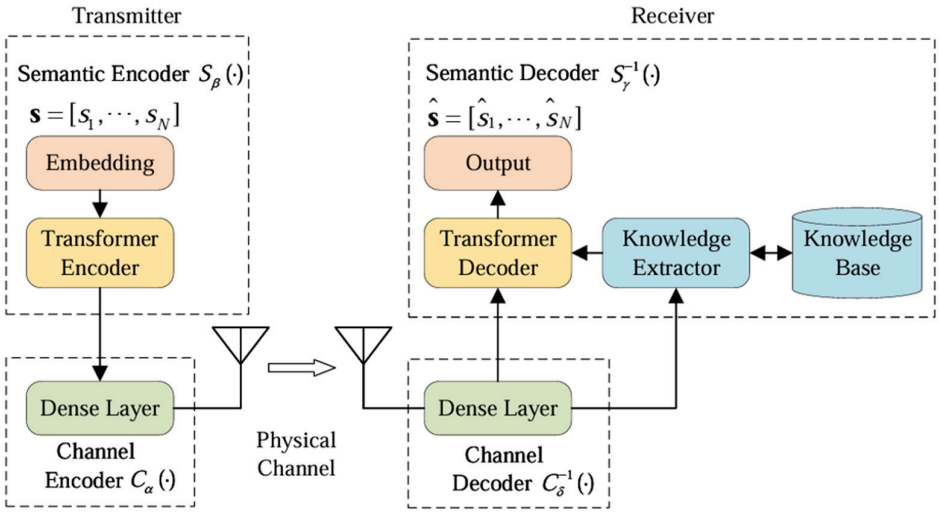

Figure 3. Knowledge-Based Semantic Communication System. Source [37].

3.4.1. Knowledge Graphs (KGs): Knowledge Graphs (KGs) represent structured knowledge through entities (nodes) and their relationships (edges), typically modeled as triples (h, r, t), where h denotes the head entity, r the relation, and t the tail entity.

Within Semantic Communication (SemCom), KGs serve multiple roles:

 doi:10.20944/preprints202602.1442.v1

- Transforming raw data sources into structured semantic representations,
- Supporting semantic disambiguation,
- Assisting the decoder in interpreting received information through structured inference.

By exploiting known relational dependencies encoded in a KG, semantic uncertainty can be reduced and encoding efficiency improved [37]. In this sense, KGs operationalize the concept of shared semantic knowledge, which was previously discussed as a critical factor in defining semantic entropy and semantic capacity.

Moreover, KGs enable structured priors that can constrain generative or predictive models, aligning decoded outputs with domain-consistent semantic structures.

3.4.2. Formal Logic and Symbolic Reasoning: Formal logic and symbolic reasoning provide the mathematical language required for rigorous inference over structured semantic representations. Logical rules—often expressed as Horn clauses—can be applied over knowledge graphs to derive new conclusions or validate semantic consistency.

Methods such as:

- Inductive Logic Programming (ILP).
- Markov Logic Networks (MLNs).

allow reasoning under uncertainty and enable the discovery of new facts while maintaining formal interpretability.

These approaches introduce a layer of traceability and logical rigor into semantic decision-making processes, which is particularly relevant for safety-critical 6G applications such as autonomous systems and industrial control.

Current research increasingly focuses on neuro-symbolic approaches, which aim to integrate the representational power of neural embeddings with the transparency and explicit reasoning capabilities of symbolic methods [38]. Such integration enhances system robustness against unseen data distributions and semantic noise, while also addressing explainability concerns.

3.4.3. Relevance to Semantic Communication: In semantic communication systems, knowledge representation mechanisms enable:

- Context-aware encoding,
- Semantic consistency verification,
- Improved robustness to channel distortions,
- Reduced semantic ambiguity during decoding.

By combining neural semantic embeddings with structured knowledge inference, SemCom systems can achieve higher reliability and better generalization compared to purely end-to-end neural architectures [39–42].

3.5. Optimization Techniques

Optimization plays a central role in translating semantic communication (SemCom) objectives into tractable mathematical problems, enabling the identification of optimal system parameters that maximize semantic performance metrics under resource constraints.

Reinforcement Learning (RL) constitutes a key optimization tool, particularly when the objective function—i.e., the semantic performance metric—is non-differentiable with respect to system actions, or when the communication channel or task environment is dynamic or partially unknown [13]. Recent studies have demonstrated the effectiveness of RL-driven task-oriented semantic transmission strategies in wireless networks, where policies are optimized directly from semantic feedback rather than bit-level reconstruction accuracy [16,43–45].

In RL, an agent learns a policy $\pi(a \mid s)$ that maximizes the expected cumulative reward $J(\pi)$ [46]:

$$J(\pi) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] \quad (19)$$

where r_t denotes the reward at time step t and $\gamma \in [0,1]$is the discount factor.

 doi:10.20944/preprints202602.1442.v1

The design of the reward function r_t is critical, as it must accurately capture task success or communication effectiveness (corresponding to Level C in Weaver's hierarchy). In semantic communication systems, the reward may be derived from:

- Semantic similarity metrics (e.g., BERTScore),
- Task accuracy (e.g., classification performance),
- Control cost in cyber-physical systems,
- Latency-constrained task completion metrics.

By directly linking optimization to task outcomes, RL provides a natural mechanism for pragmatic communication, where the goal is not symbol fidelity but action effectiveness.

Policy-gradient–based methods such as REINFORCE, Proximal Policy Optimization (PPO), and Actor–Critic algorithms are commonly employed to optimize adaptive encoding strategies and resource allocation policies directly from semantic feedback. This enables systems to learn efficient semantic encoding policies based on real task outcomes rather than bit-level similarity alone.

Such adaptive optimization mechanisms are particularly relevant for AI-native 6G networks, where dynamic spectrum allocation, edge intelligence, and context-aware transmission policies are required to meet stringent latency and reliability constraints.

Table 3 summarizes the principal mathematical and computational tools that underpin modern semantic communication systems. Unlike classical communication frameworks—primarily concerned with symbol fidelity and channel capacity—semantic communication integrates information-theoretic relevance (IB), end-to-end representation learning (DeepJSCC), contextual modeling (Transformers), structured knowledge reasoning (KGs), and task-driven adaptive optimization (RL). This convergence of information theory, machine learning, and symbolic reasoning reflects the inherently interdisciplinary and AI-native nature of semantic communication in future Internet and 6G environments.

Table 3. Mathematical and Computational Tools for Semantic Communications.

Tool / Approach	Key Mathematical Concept / Formulation	Role in Semantic Communication	Key Benefit
Information Bottleneck (IB)	Minimize $\mathcal{L}[p(z \mid x)] = I(X;Z) - \beta I(Z;Y)$	Extract task-relevant features; semantic compression	Theoretical foundation for relevance; rate–relevance trade-off
Deep Learning (AE/JSCC)	$\min E\,[\mathcal{L}(s,\hat{s})] sobre \theta$	End-to-end implementation; implicit learning of coding and robustness Z	Joint optimization; avoids *cliff-effect*
Deep Learning (Transformers)	$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$	Capture context and long-range dependencies in sequences	Effective modelling of semantic relationships
Knowledge Graphs (KGs)	Triples *(h,r,t)*; graph/logical structure	Explicit knowledge representation; disambiguation; reasoning	Incorporation of world knowledge; interpretability
Reinforcement Learning (RL)	Maximize $J(\pi) = E_\pi[\sum \gamma^t r_t]$	Optimization for non-differentiable objectives (semantic similarity, task success)	Direct optimization of semantic metrics; adaptability

 doi:10.20944/preprints202602.1442.v1

4. Mathematical Challenges and Open Problems

Despite the remarkable progress driven by deep learning, research in Semantic Communications (SemCom) continues to face fundamental mathematical challenges that require rigorous theoretical investigation.

4.1. Toward a Unified Theory of Semantic Information

The primary theoretical obstacle is the absence of a unified and universally accepted framework for semantic information. Currently, multiple formulations coexist—based on logical probability, synonym mappings, knowledge-based representations, and task-oriented relevance measures. This theoretical fragmentation has been explicitly highlighted in recent surveys on semantic communications [47], and prevents the establishment of rigorous fundamental limits comparable to Shannon entropy and channel capacity.

The difficulty stems from the inherently relative nature of meaning. As discussed in foundational philosophical analyses of semantic information [48], semantic content depends on context, interpretation, and prior knowledge, making it intrinsically challenging to formulate a universal axiomatic framework.

Recent efforts toward formalizing semantic entropy and semantic channel capacity [22] represent important steps in this direction; however, a mathematically unified theory that reconciles logical, probabilistic, and task-oriented perspectives remains an open problem. Future research must explore either consistent extensions of Shannon's framework into the semantic domain or more abstract mathematical structures capable of relating heterogeneous semantic formulations.

4.2. Robust and Universal Semantic Metrics

The optimization of SemCom systems requires distortion or similarity metrics that are robust, computable, and strongly correlated with human perception and task utility. Although current metrics—such as BERTScore and LPIPS—are widely adopted, they depend on deep embedding models and are therefore domain-specific and potentially opaque.

Recent task-oriented semantic communication frameworks emphasize that system performance critically depends on the design of semantic-aware evaluation metrics [49]. For example, adaptive Information Bottleneck–guided JSCC explicitly links semantic relevance to rate–distortion optimization under bandwidth constraints [4], while task-oriented explainable semantic communication models integrate interpretability into performance evaluation [43].

However, formal quantification of "semantic noise"—the distortion of meaning rather than symbols—remains unresolved. Without mathematically grounded and standardized semantic metrics, both reinforcement learning–based optimization and objective benchmarking across systems remain problematic.

4.3. Mathematical Modeling of Context and Knowledge Dependence

Semantics is inseparably linked to context and shared knowledge. Modeling dynamic contextual factors—situational, linguistic, or environmental—in a mathematically rigorous manner remains highly complex.

Knowledge Graphs (KGs) have been proposed as structured priors to mitigate semantic ambiguity [42,50,51], while neuro-symbolic approaches attempt to combine deep embeddings with explicit logical reasoning to improve robustness and interpretability [52]. Nevertheless, scalable integration of symbolic knowledge structures with sub-symbolic neural models remains computationally demanding.

Furthermore, communication alignment problems can be analyzed through signaling-game frameworks, where multiple equilibria may exist between encoder and decoder policies. Misalignment may lead to semantic failure despite syntactic correctness, limiting interoperability in distributed systems.

 doi:10.20944/preprints202602.1442.v1

In multi-user and distributed network settings, additional theoretical challenges arise from heterogeneous or outdated knowledge bases, which may alter semantic entropy and effective semantic capacity. Formal models of knowledge mismatch and update dynamics are still largely underdeveloped.

4.4. Scalability and Computational Complexity Analysis

The reliance on large-scale deep learning models introduces significant concerns regarding scalability and computational complexity. Training DeepJSCC systems, Transformer-based semantic encoders, or reinforcement learning–based optimization frameworks require extensive computational resources and may suffer from instability.

Recent discussions on AI-native 6G systems emphasize that computational efficiency, latency constraints, and energy consumption must be jointly considered with communication performance [53]. However, a rigorous mathematical framework for analyzing complexity–performance trade-offs in semantic communication systems remains absent.

Moreover, the deployment of semantic communication models on edge devices or IoT sensors is constrained by limited computational and energy budgets. Developing lightweight semantic encoders, efficient generative inference mechanisms, and distributed training strategies tailored to SemCom is therefore imperative.

5. Opportunities and Effectiveness Theory in 6G Networks

Semantic communications constitute a key technological enabler for the vision of sixth-generation (6G) networks, supporting the transition from the Internet of Everything toward the Intelligent Internet of Everything. Their fundamental contribution lies in directly addressing Weaver's Level C: The Theory of Effectiveness (pragmatic communication).

Recent 6G visions emphasize AI-native networking, goal-oriented communication, and task-driven optimization as core design principles [14,53]. In this context, semantic communication shifts the optimization objective from bit fidelity to task success and decision effectiveness.

5.1. Effectiveness Theory and Pragmatic Communication (Level C)

The fundamental objective of pragmatic (effective) communication is to ensure that the transmitted meaning (Level B) influences the behavior or decision of the receiver in the intended manner, thereby achieving the task objective.

Effectiveness becomes the primary performance metric, requiring transmitted information to possess pragmatic semantic significance aligned with the destination's goals. This perspective is increasingly reflected in task-oriented communication frameworks for 6G systems [43].

This theory implies:

- Intention Recognition and Semantic QoS (QoS-S): Semantic communication enables the construction of an intelligence plane within the network capable of differentiating and prioritizing traffic based on message intent or semantic importance. This leads to the concept of Semantic Quality of Service (QoS-S), where network resources are allocated not merely based on throughput or latency, but on semantic criticality and task impact. Such an approach aligns with AI-native 6G architectures that integrate communication, computation, and intelligence [14,54].
- Modeling Effectiveness Errors: Beyond bit errors, SemCom systems must model errors arising from:
    - Misalignment in semantic interpretation,
    - Knowledge base mismatch,
    - Contextual ambiguity,
    - Processing limitations at the receiver.

 doi:10.20944/preprints202602.1442.v1

These errors directly affect Level C effectiveness. Emerging works on task-oriented and explainable semantic communication highlight the necessity of modeling such semantic distortions explicitly [43].

Accordingly, future 6G communication systems envision meaning-driven transmission mechanisms that optimize network resources for next-generation services, as illustrated in Figure 4.

Figure 4. Vision of 6G networks articulated with the Theory of Effectiveness.

5.2. Semantic Communications for Ultra-Reliable Low-Latency Communications (URLLC)

URLLC services in 6G—such as industrial control, autonomous systems, and mission-critical operations—require massive data handling under extreme latency and reliability constraints [55].

Semantic communication addresses these requirements fundamentally:

• Delay Violation Probability (DVP) Minimization

By focusing on goal-oriented content, SemCom achieves substantial compression of transmitted data, reducing wireless resource load and consequently minimizing Delay Violation Probability (DVP). This aligns with recent analyses linking semantic compression to latency-aware optimization [45].

• Semantic-Aware Resource Allocation

Unlike traditional bit-centric communication, SemCom adapts resource allocation—such as transmission power, coding schemes, and modulation formats—according to semantic importance. Task-critical information is prioritized to guarantee control reliability under constrained resources.

Recent task-oriented wireless communication studies confirm that semantic relevance–aware scheduling significantly improves system-level performance [43,44].

• Age of Information (AoI) Alignment

Semantic communication aligns naturally with the Age of Information (AoI) framework, which measures information freshness rather than throughput. By transmitting only semantically relevant updates, SemCom reduces obsolete transmissions and optimizes freshness-aware scheduling [56,57].

5.3. Empowering the Metaverse and Semantic-Centric Digital Twins (SCDT)

The Metaverse and Digital Twin ecosystems impose unsustainable bandwidth requirements due to immersive rendering, multi-sensory streaming, and continuous state synchronization [58].

 doi:10.20944/preprints202602.1442.v1

Semantic communication emerges as a foundational solution:

• Compression for Immersion: SemCom enables efficient compression of sensor data, video streams, and rendering features by transmitting structured semantic representations rather than raw signals. This reduces communication overhead while preserving immersive experience quality.

• Semantic-Centric Digital Twins (SCDT): SemCom facilitates the development of Semantic-Centric Digital Twins (SCDT), where virtual representations are built upon shared semantic concept definitions. Instead of exchanging raw data streams, physical and virtual domains exchange structured meaning representations, enabling efficient cross-domain interaction.

Digital twin architectures for 6G increasingly emphasize semantic interoperability and AI-driven synchronization mechanisms [58].

• Compression–Robustness Trade-Off: A key challenge lies in achieving extreme semantic compression without compromising robustness. Over-aggressive feature abstraction may reduce resilience to noise and degrade immersive experience due to inaccurate reconstruction of critical semantic attributes.

Balancing semantic abstraction and anti-noise capability remains an open research problem, particularly in generative and diffusion-based SemCom systems.

6. Conclusion

Semantic communications represent a necessary and profound evolution of communication theory, driven by the demands of task-oriented future networks. Their mathematical foundations are being progressively constructed upon rigorous extensions of classical information theory—such as semantic rate–distortion formulations $R_s(D)$ and semantic channel capacity C_s—which indicate the theoretical possibility of surpassing Shannon limits through the exploitation of semantic redundancy. The Information Bottleneck principle provides a principled framework for relevance extraction, while advanced architectures such as Generative Semantic Communications (Gen-SemCom), enabled by diffusion models and grounded in stochastic differential equations (SDEs), offer a practical pathway toward extreme efficiency and robustness against channel impairments.

By explicitly addressing Weaver's Theory of Effectiveness (Level C), semantic communications reposition network design around task success rather than bit-level fidelity. This shift establishes SemCom as a foundational enabler of critical 6G services, including Ultra-Reliable Low-Latency Communications (URLLC), immersive metaverse environments, semantic-centric digital twins, and intelligent cyber–physical systems. Nevertheless, the practical realization of these theoretical limits currently relies heavily on deep learning–based approximations—such as DeepJSCC and Transformer architectures—which function as powerful yet empirically driven solvers of otherwise intractable optimization problems.

A fundamental tension therefore persists between the pursuit of a unified mathematical theory of semantic information and the dependence on data-driven neural implementations. Resolving theoretical fragmentation, developing robust and universally grounded semantic performance metrics that faithfully capture task utility, and rigorously addressing scalability, computational complexity, latency, and security constraints remain open challenges.

Looking forward, the transition from theoretical promise to large-scale 6G deployment will require coordinated advances in cross-layer semantic-aware architectures and globally harmonized standardization efforts. Future networks must evolve beyond throughput-centric key performance indicators and incorporate semantic effectiveness, task relevance, and freshness-aware metrics into system evaluation frameworks. The formalization of Semantic Quality of Service (QoS-S), semantic-aware resource orchestration, and interoperable knowledge abstraction layers will be critical to ensure scalability and robustness across heterogeneous devices and distributed intelligence planes.

Ultimately, semantic communications constitute not merely an incremental enhancement of classical systems, but a paradigm shift toward meaning-centric network intelligence. Bridging rigorous mathematical foundations, efficient AI-native implementations, and standardized semantic

 doi:10.20944/preprints202602.1442.v1

frameworks will be indispensable for enabling the next generation of intelligent, efficient, and reliable 6G ecosystems.

Author Contributions: Conceptualization, E.A.H; methodology, H.F.B.-O. and E.A.H.; formal analysis, H.F.B.-O. and N.C.R.-I.; investigation, E.A.H. and N.C.R.-I.; resources, H.F.B.-O. and N.C.R.-I.; writing—original draft preparation, E.A.H.; writing—review and editing, H.F.B.-O. and N.C.R.-I.; visualization, E.A.H.; supervision, H.F.B.-O.; project administration, H.F.B.-O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors acknowledge the academic support of the Universidad del Quindío, Colombia. During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-5) for language refinement and structural editing. The authors have reviewed and edited the generated content and take full responsibility for the final version of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shannon, C.E. A Mathematical Theory of Communication. *Bell System Technical Journal* 1948, 27(3), 379–423; 27(4), 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
2. Shannon, C.E.; Weaver, L. The Mathematical Theory of Communication. University of Illinois Press,1950, 34, 312–313.
3. Qin, Z.; Tao, X.; Lu, J.; Tong, W.; Li, G.Y. Semantic Communications: Principles and Challenges. arXiv 2022. doi: 10.48550/ARXIV.2201.01389
4. Zhang, P.; Liu, Y.; Song, Y.; Zhang, J. Advances and Challenges in Semantic Communications: A Systematic Review. Natl Sci Open 2024, 3, 20230029, doi:10.1360/nso/20230029.
5. Iyer, S.; Khanai, R.; Torse, D.; Pandya, R.J.; Rabie, K.M.; Pai, K.; Khan, W.U.; Fadlullah, Z. A Survey on Semantic Communications for Intelligent Wireless Networks. Wireless Pers Commun 2023, 129, 569–611, doi:10.1007/s11277-022-10111-7.
6. Floridi, L. Outline of a Theory of Strongly Semantic Information. Minds and Machines 2004, 14, 197–221, doi:10.1023/B:MIND.0000021684.50925.c9.
7. Coghill, G. Towards a Measure Theory of Semantic Information; 2025, doi: 10.48550/arXiv.2508.00525
8. Fernandes, G.; Fontes, H.; Campos, R. Semantic Communications: The New Paradigm Behind Beyond 5G Tech-nologies 2024.doi: 10.48550/ARXIV.2406.00754
9. Xie, H.; Qin, Z.; Li, G.Y.; Juang, B.-H. Deep Learning Enabled Semantic Communication Systems. IEEE Trans. Signal Process. 2021, 69, 2663–2675, doi:10.1109/TSP.2021.3071210.
10. Bourtsoulatze, E.; Burth Kurka, D.; Gunduz, D. Deep Joint Source-Channel Coding for Wireless Image Trans-mission. IEEE Trans. Cogn. Commun. Netw. 2019, 5, 567–579, doi:10.1109/TCCN.2019.2919300.
11. Weng, Z.; Qin, Z. Semantic Communication Systems for Speech Transmission. IEEE J. Select. Areas Commun. 2021, 39, 2434–2444, doi:10.1109/JSAC.2021.3087240.
12. Tishby, N.; Pereira, F.C.; Bialek, W. The Information Bottleneck Method. arXiv 2000. doi: 10.48550/ARXIV.PHYSICS/0004057
13. Peng, J.; Xing, H.; Xu, L.; Luo, S.; Dai, P.; Feng, L.; Song, J.; Zhao, B.; Xiao, Z. Adversarial Reinforcement Learning Based Data Poisoning Attacks Defense for Task-Oriented Multi-User Semantic Communication. IEEE Trans. on Mobile Comput. 2024, 23, 14834–14851, doi:10.1109/TMC.2024.3447087.
14. Letaief, K.B.; Chen, W.; Shi, Y.; Zhang, J.; Zhang, Y.-J.A. The Roadmap to 6G: AI Empowered Wireless Networks. IEEE Commun. Mag. 2019, 57, 84–90, doi:10.1109/MCOM.2019.1900271.
15. Akyildiz, I.F.; Kak, A.; Nie, S. 6G and Beyond: The Future of Wireless Communications Systems. IEEE Access 2020, 8, 133995–134030, doi:10.1109/ACCESS.2020.3010896.

 doi:10.20944/preprints202602.1442.v1

16. Wang, Y.; Li, R.; Wang, C.; Ye, J.; Feng, C.; Guo, S. Collaborative Learning for Task-Oriented Semantic Commu-nications: Overcoming Data Mismatch Between Transceivers. IEEE Open Journal of the Communications Society 2025, 6, 5778–5794, doi:10.1109/OJCOMS.2025.3586462.
17. Xin, G.; Fan, P. EXK-SC: A Semantic Communication Model Based on Information Framework Expansion and Knowledge Collision. Entropy 2022, 24, doi:10.3390/e24121842.
18. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A Survey on Evaluation of Large Language Models. ACM Trans. Intell. Syst. Technol. 2024, 15, 1–45, doi:10.1145/3641289.
19. Xie, H.; Qin, Z.; Tao, X.; Letaief, K.B. Task-Oriented Multi-User Semantic Communications. IEEE J. Select. Areas Commun. 2022, 40, 2584–2597, doi:10.1109/JSAC.2022.3191326.
20. Lu, C. Using the Semantic Information G Measure to Explain and Extend Rate-Distortion Functions and Maxi-mum Entropy Distributions. Entropy 2021, 23, doi:10.3390/e23081050.
21. Niu, K.; Zhang, P. A Mathematical Theory of Semantic Communication: Overview 2024. https://arxiv.org/abs/2401.14160v1
22. Hua, S.; Member, G.S.; Sun, Y.; Member, S.; Ma, K.; Imran, M.A. A Mathematical Framework of Semantic Communication Based on Category Theory. 2025. https://arxiv.org/pdf/2504.11334v1
23. Singh, G. BERTScore: Evaluating Text Generation with BERT - Statwiki. https://wiki.math.uwaterloo.ca/statwiki/index.php?title=BERTScore:_Evaluating_Text_Generation_with_BERT
24. Efimov, V. Large Language Models: SBERT - Sentence-BERT. Towards Data Science 2023. https://towardsdatascience.com/sbert-deb3d4aef8a4/
25. Liu, F.; Tong, W.; Yang, Y.; Sun, Z.; Guo, C. Task-Oriented Image Semantic Communication Based on Rate-Distortion Theory 2022. https://arxiv.org/abs/2201.10929
26. Stavrou, P.A.; Kountouris, M. A Rate Distortion Approach to Goal-Oriented Communication. In Proceedings of the 2022 IEEE International Symposium on Information Theory (ISIT); IEEE: Espoo, Finland, June 26 2022; pp. 590–595. doi: 10.1109/ISIT50566.2022.9834593
27. Zhang, G.; Hu, Q.; Qin, Z.; Cai, Y.; Yu, G.; Tao, X. A Unified Multi-Task Semantic Communication System for Multimodal Data. IEEE Trans. Commun. 2024, 72, 4101–4116, doi:10.1109/TCOMM.2024.3364990.
28. Xie, H.; Qin, Z.; Li, G.Y. Task-Oriented Multi-User Semantic Communications for VQA. IEEE Wireless Commun. Lett. 2022, 11, 553–557, doi:10.1109/LWC.2021.3136045.
29. Freitag, M.; Rei, R.; Mathur, N.; Lo, C.; Stewart, C.; Avramidis, E.; Kocmi, T.; Foster, G.; Lavie, A.; Martins, A.F.T. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In Proceedings of the Proceedings of the Seventh Conference on Machine Translation (WMT); Association for Computational Linguistics: Abu Dhabi, United Arab Emirates (Hybrid), 2022; pp. 46–68. doi: 10.18653/v1/2022.wmt-1.2
30. Xin, G.; Fan, P.; Letaief, K.B. Semantic Communication: A Survey of Its Theoretical Development. Entropy 2024, 26, 102, doi:10.3390/e26020102.
31. Wang, Y.; Guo, S.; Deng, Y.; Zhang, H.; Fang, Y. Privacy-Preserving Task-Oriented Semantic Communications Against Model Inversion Attacks. IEEE Trans. Wireless Commun. 2024, 23, 10150–10165, doi:10.1109/TWC.2024.3369170.
32. Zhang, H.; Shao, S.; Tao, M.; Bi, X.; Letaief, K.B. Deep Learning-Enabled Semantic Communication Systems With Task-Unaware Transmitter and Dynamic Data. IEEE J. Select. Areas Commun. 2023, 41, 170–185, doi:10.1109/JSAC.2022.3221991.
33. Wang, Y.; Han, H.; Feng, Y.; Zheng, J.; Zhang, B. Semantic Communication Empowered 6G Networks: Tech-niques, Applications, and Challenges. IEEE Access 2025, 13, 28293–28314, doi:10.1109/ACCESS.2025.3532797.
34. Shi, G.; Xiao, Y.; Li, Y.; Xie, X. From Semantic Communication to Semantic-Aware Networking: Model, Archi-tecture, and Open Problems. IEEE Commun. Mag. 2021, 59, 44–50, doi:10.1109/MCOM.001.2001239.
35. Zhang, P.; Xu, W.; Gao, H.; Niu, K.; Xu, X.; Qin, X.; Yuan, C.; Qin, Z.; Zhao, H.; Wei, J.; et al. Toward Wis-dom-Evolutionary and Primitive-Concise 6G: A New Paradigm of Semantic Communication Networks. Engi-neering 2022, 8, 60–73, doi:10.1016/j.eng.2021.11.003.

 doi:10.20944/preprints202602.1442.v1

36. Ren, J.; Sun, Y.; Du, H.; Yuan, W.; Wang, C.; Wang, X.; Zhou, Y.; Zhu, Z.; Wang, F.; Cui, S. Generative Semantic Communication: Architectures, Technologies, and Applications. Engineering 2025, S2095809925004291, doi:10.1016/j.eng.2025.07.022.
37. Ni, F.; Wang, B.; Li, R.; Zhao, Z.; Zhang, H. Interplay of Semantic Communication and Knowledge Learning. 2024. https://arxiv.org/pdf/2402.03339v1
38. Zhang, J.; Chen, B.; Zhang, L.; Ke, X.; Ding, H. Neural, Symbolic and Neural-Symbolic Reasoning on Knowledge Graphs. AI Open 2021, 2, 14–35, doi:10.1016/j.aiopen.2021.03.001.
39. Farsad, N.; Rao, M.; Goldsmith, A. Deep Learning for Joint Source-Channel Coding of Text. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE: Calgary, AB, April 2018; pp. 2326–2330. doi: 10.1109/ICASSP.2018.8461983
40. Nie, H.; Lu, S.; Wu, J.; Zhu, J. Deep Model Intellectual Property Protection With Compression-Resistant Model Watermarking. IEEE Trans. Artif. Intell. 2024, 5, 3362–3373, doi:10.1109/TAI.2024.3351116.
41. Cheng, S.; Zhang, X.; Sun, Y.; Cui, Q.; Tao, X. Knowledge Discrepancy Oriented Privacy Preserving for Semantic Communication. IEEE Trans. Veh. Technol. 2024, 73, 11637–11646, doi:10.1109/TVT.2024.3381222.
42. Hogan, A.; Blomqvist, E.; Cochez, M.; D'amato, C.; Melo, G.D.; Gutierrez, C.; Kirrane, S.; Gayo, J.E.L.; Navigli, R.; Neumaier, S.; et al. Knowledge Graphs. ACM Comput. Surv. 2022, 54, 1–37, doi:10.1145/3447772.
43. Ma, S.; Qiao, W.; Wu, Y.; Li, H.; Shi, G.; Gao, D.; Shi, Y.; Li, S.; Al-Dhahir, N. Task-Oriented Explainable Semantic Communications. IEEE Trans. Wireless Commun. 2023, 22, 9248–9262, doi:10.1109/TWC.2023.3269444.
44. Sagduyu, Y.E.; Ulukus, S.; Yener, A. Task-Oriented Communications for NextG: End-to-End Deep Learning and AI Security Aspects. IEEE Wireless Commun. 2023, 30, 52–60, doi:10.1109/MWC.006.2200494.
45. Sun, L.; Yang, Y.; Chen, M.; Guo, C.; Saad, W.; Poor, H.V. Adaptive Information Bottleneck Guided Joint Source and Channel Coding for Image Transmission. IEEE J. Select. Areas Commun. 2023, 41, 2628–2644, doi:10.1109/JSAC.2023.3288238.
46. Kågebäck, M.; Carlsson, E.; Dubhashi, D.; Sayeed, A. A Reinforcement-Learning Approach to Efficient Commu-nication. PLoS ONE 2020, 15, e0234894, doi:10.1371/journal.pone.0234894.
47. Niyato, D. Editorial: Third Quarter 2023 IEEE Communications Surveys and Tutorials. IEEE Communications Surveys & Tutorials 2023, 25, i–vi, doi:10.1109/COMST.2023.3301328.
48. Floridi, L. Semantic Conceptions of Information. In; 2008.
49. Chen, Z.; Zhang, Z.; Yang, Z. Big AI Models for 6G Wireless Networks: Opportunities, Challenges, and Research Directions. IEEE Wireless Commun. 2024, 31, 164–172, doi:10.1109/MWC.015.2300404.
50. Ji, G.; He, S.; Xu, L.; Liu, K.; Zhao, J. Knowledge Graph Embedding via Dynamic Mapping Matrix. In Proceed-ings of the Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); Association for Computational Linguistics: Beijing, China, 2015; pp. 687–696.
51. Huang, X.; Zhang, J.; Li, D.; Li, P. Knowledge Graph Embedding Based Question Answering. In Proceedings of the Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining; ACM: Mel-bourne VIC Australia, January 30 2019; pp. 105–113.
52. Garcez, A. d'Avila; Lamb, L.C. Neurosymbolic AI: The 3rd Wave. Artif Intell Rev 2023, 56, 12387–12406, doi:10.1007/s10462-023-10448-w.
53. Letaief, K.B.; Shi, Y.; Lu, J.; Lu, J. Edge Artificial Intelligence for 6G: Vision, Enabling Technologies, and Applica-tions. IEEE J. Select. Areas Commun. 2022, 40, 5–36, doi:10.1109/JSAC.2021.3126076.
54. Calvanese Strinati, E.; Barbarossa, S.; Gonzalez-Jimenez, J.L.; Ktenas, D.; Cassiau, N.; Maret, L.; Dehos, C. 6G: The Next Frontier: From Holographic Messaging to Artificial Intelligence Using Subterahertz and Visible Light Communication. IEEE Veh. Technol. Mag. 2019, 14, 42–50, doi:10.1109/MVT.2019.2921162.
55. Yang, P.; Xiao, Y.; Xiao, M.; Li, S. 6G Wireless Communications: Vision and Potential Techniques. IEEE Network 2019, 33, 70–75, doi:10.1109/MNET.2019.1800418.
56. Kahraman, İ.; Köse, A.; Koca, M.; Anarim, E. Age of Information in Internet of Things: A Survey. IEEE Internet Things J. 2024, 11, 9896–9914, doi:10.1109/JIOT.2023.3324879.

 doi:10.20944/preprints202602.1442.v1

57. Chen, J.; Wang, J.; Jiang, C.; Wang, J. Age of Incorrect Information in Semantic Communications for NOMA Aided XR Applications. IEEE J. Sel. Top. Signal Process. 2023, 17, 1093–1105, doi:10.1109/JSTSP.2023.3282836.
58. Poorzare, R.; Kanellopoulos, D.N.; Sharma, V.K.; Dalapati, P.; Waldhorst, O.P. Network Digital Twin Toward Networking, Telecommunications, and Traffic Engineering: A Survey. IEEE Access 2025, 13, 16489–16538, doi:10.1109/ACCESS.2025.3531947.