Akanksha Aher, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Information Bottleneck Theory in Multimodal AI: Principles, Architectures, and Emerging Research Directions

Akanksha Aher[1], Swaraj Rasam[2], Dr. Jasbir Kaur[3], Ifrah Kampoo[4]

[1,2]Student of Master of Computer Applications (MCA), Guru Nanak Institute of Management Studies, Matunga, Mumbai, India

[3]Director, GNIMS B-School, Head of Information Technology and HR, Guru Nanak Institute of Management Studies, Matunga, Mumbai, India

[4]Assistant Professor, Guru Nanak Institute of Management Studies, Matunga, Mumbai, India

Abstract- Artificial intelligence today faces a fundamental challenge: modern AI systems are trained on enormous quantities of data, yet most of that data is irrelevant to the task at hand. The Information Bottleneck (IB) principle offers a powerful theoretical answer — it teaches AI to retain only what is genuinely useful for a task while discarding everything else. This survey examines how IB applies to multimodal foundation models that process images and text simultaneously, such as CLIP, BLIP-2, LLaVA, and Flamingo. We explain the theory in accessible terms, review how these models apply IB in practice, and discuss real-world benefits, applications, and open challenges. Our goal is to bridge information-theoretic foundations and engineering practice for the MCA research community.

Keywords: Information Bottleneck, Multimodal AI, Foundation Models, CLIP, Representation Learning, Explainable AI, Cross-modal Fusion.

I. INTRODUCTION

We live in a world of information overload. Every second, millions of photographs are uploaded, billions of words are written, and terabytes of video are streamed. The AI systems of today are built on this data — yet most of it is irrelevant to any given task. Research in both neuroscience and computer science has shown that the ability to ignore irrelevant information is just as critical as the ability to absorb new information.

A student who memorises every word of a textbook without grasping the concepts will struggle in an exam that demands reasoning. An AI model that memorises training patterns without extracting essential meaning will fail on unseen data. This is the problem the Information Bottleneck (IB) principle was designed to solve.

IB gives AI a principled way to decide what information is worth keeping and what should be discarded. It was introduced by Tishby, Pereira, and Bialek in 1999 at the intersection of information theory and machine learning. In 2015, Tishby and Zaslavsky showed that deep neural networks naturally behave in ways IB theory predicts — sparking a wave of research that continues today.

Simultaneously, AI was being transformed by the rise of multimodal systems — AI that processes multiple types of data at once, just as humans do. Today, models such as CLIP [1], BLIP-2 [2], LLaVA [3], and Flamingo [4] can simultaneously read images and understand language. Yet they face a profound challenge: combining an image with millions of pixels and a ten-word caption requires principled compression and fusion — precisely what IB provides.

This survey explains IB theory in plain language, reviews its application to multimodal foundation models, and discusses real-world advantages and open challenges. It is aimed at MCA students and researchers entering this field.

II. BACKGROUND AND MOTIVATION

A. The Problem Without IB

Standard deep neural networks are trained to minimize prediction error on training data. Without IB constraints, they risk overfitting — memorising

irrelevant surface details rather than learning the underlying concept. A model trained on indoor photos of cats may fail on outdoor photos because it inadvertently learned that carpets and indoor lighting are associated with cats.

This problem intensifies in multimodal AI. An image may contain millions of pixel values; a caption describing it may contain only fifteen words. A naive model overwhelmed by visual data will effectively ignore the text — a phenomenon called cross-modal information imbalance [14]. The model learns an unbalanced representation that performs poorly when the critical information is in the less data-dense modality.

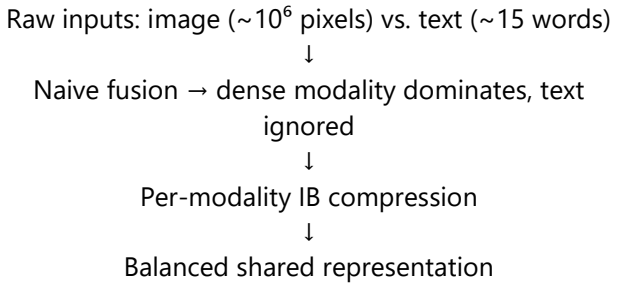

Fig. 1. Cross-modal information imbalance: dense visual data can overwhelm sparse text, while per-modality IB compression yields a balanced shared representation.

B. The Human Analogy

The IB principle mirrors how humans learn. Before an exam, you do not memorise every word you have read. Your brain compresses experiences into essential concepts — it extracts what helps answer questions and lets go of the rest. This deliberate forgetting is not a loss; it is the foundation of flexible, generalizable reasoning.

IB teaches AI the same discipline. By forcing information through a narrow conceptual bottleneck, only the most task-relevant knowledge survives. What passes through is not raw data but understanding.

III. THE INFORMATION BOTTLENECK PRINCIPLE

A. What Is Mutual Information?

Before understanding the bottleneck, we need the concept of mutual information from Shannon's information theory [21]. Mutual information measures how much knowing one variable tells you about another. Dark overcast sky and rain share high mutual information — one predicts the other reliably. Sock colour and weather share almost none. IB builds on this: find a compressed representation of the input that keeps high mutual information with the desired output (the prediction target) while having low mutual information with the raw input (discarding noise). The tightness of the bottleneck is controlled by a single tuning parameter that the designer adjusts for each task.

B. The Bottleneck Concept

Imagine water flowing through a pipe. Install a narrow section in the middle — a bottleneck — and only a limited flow passes through. The narrow section filters. In IB, the bottleneck is the compressed internal representation the AI creates from its input. Only the most task-relevant information survives the squeeze; irrelevant details dissipate.

The designer controls how tight the bottleneck is. A narrower bottleneck produces a more abstract, generalisable representation. A wider bottleneck retains more detail but risks noise. Finding the right balance for each application is the core design challenge.

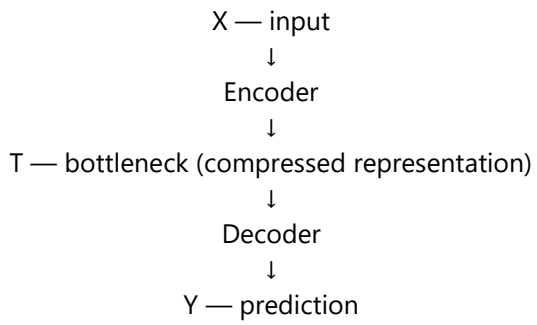

Fig. 2. The Information Bottleneck principle. The encoder compresses input X into a bottleneck T and

the decoder predicts Y — minimising I(X;T) while maximising I(T;Y), i.e. min I(X;T) − β·I(T;Y).

C. Two Phases of Deep Learning

In 2015, Tishby and Zaslavsky [5] showed that training a deep neural network appears to proceed in two phases. In the fitting phase, the network rapidly learns to associate inputs with outputs — mutual information with the target label increases quickly. In the compression phase, the network begins forgetting irrelevant details — mutual information with the raw input decreases while prediction ability is preserved.

This mirrors human learning: rapid absorption of new material followed by the gradual consolidation of core concepts. It is during the compression phase that the model acquires the ability to generalise beyond its training examples.

Phase 1 — Fitting: I(X;T)↑ and I(T;Y)↑
↓
Phase 2 — Compression: I(X;T)↓, I(T;Y) preserved
↓
Generalisation beyond training data

Fig. 3. The two phases of deep learning on the information plane (illustrative): a fitting phase followed by a compression phase that drives generalisation.

D. Deep Variational IB (Deep VIB)

Making IB tractable for large neural networks was achieved by Alemi et al. at Google in 2017 [6]. Their Deep VIB encodes each input not as a fixed value but as a probability distribution [24]. The model then samples from this distribution to generate its internal representation. The deliberate randomness forces selectivity: only robust, reliable information survives repeated sampling. This approximation made IB practical for real AI systems and demonstrated improved robustness against adversarial attacks compared to standard models.

An analogy helps here: instead of the AI recording a precise, fixed description of each input, it records a blurry, probabilistic sketch. Details that appear consistently across many samples are drawn sharply; accidental details that vary from sample to sample blur away. The resulting representations are far more generalisable than precise memorisations — just as a caricature captures the essence of a face more reliably than a high-resolution photograph for the purpose of recognition.

Input x
↓
Encoder
↓
Distribution (μ, σ)
↓
Sample z ~ N(μ, σ)
↓
Decoder → ŷ

Fig. 4. Deep Variational IB. The encoder outputs a distribution (μ, σ) rather than a fixed code; sampling z forces only robust features to survive.

Deep VIB also revealed a practical benefit: models trained with variational IB tend to be naturally calibrated — their confidence scores accurately reflect the probability of being correct. Standard neural networks are often over-confident, assigning near-100% confidence to incorrect predictions. IB-trained models are more honest about uncertainty, which is essential for deployment in safety-critical settings such as medical diagnostics and autonomous vehicles.

IV. MULTIMODAL AI AND FOUNDATION MODELS

A. What Is Multimodal AI?

A modality is a type of information channel: vision (images, video), language (text, speech), audio (sound, music), or sensor data (radar, LiDAR). Traditional AI systems handled only one modality at a time. Multimodal AI integrates multiple modalities simultaneously — just as humans combine sight, hearing, and reading to navigate the world.

The development of multimodal AI was driven by a straightforward observation: the most valuable real-world information often exists at the intersection of modalities. Medical diagnosis requires reading patient history, examining scans, and considering

symptoms together. Content moderation requires understanding image and caption as a pair, not individually.

B. CLIP

CLIP (OpenAI, 2021) [1] trained two neural networks — one reading images [23], one reading text — on 400 million image-text pairs from the internet. It taught them to produce compatible representations: matching image and correct caption should be close together; mismatched pairs should be far apart. The result was remarkable zero-shot ability — CLIP could classify images into categories it was never explicitly trained on, by matching them to text descriptions.

From an IB perspective, CLIP's contrastive training implicitly performs compression. The model must discard image-specific noise that is irrelevant to language and language-specific noise that is irrelevant to vision. What survives is shared semantic meaning — precisely the IB-optimal representation. Research in 2025 (CIBR [9]) formally proved this connection and showed that adding explicit IB regularisation further improves CLIP's zero-shot performance.

Image & text inputs
↓
Vision Encoder || Text Encoder (parallel)
↓
IB-compressed shared semantic space
↓
Aligned image–text embeddings

Fig. 5. Dual-encoder multimodal IB in CLIP. Contrastive training discards modality-specific noise, leaving an IB-optimal shared semantic space.

C. BLIP-2

BLIP-2 (Salesforce, 2023) [2] addressed the cost of training multimodal models from scratch by keeping a pre-trained vision model and a pre-trained language model frozen, and training a lightweight bridging module — the Querying Transformer (Q-Former) [22] — to connect them. The Q-Former takes the rich output of the vision model and selects only the subset of visual information most relevant for the language model.

This selection is a direct practical implementation of IB: the Q-Former is a learned bottleneck compressing visual information into a small number of query vectors. The result is exceptional efficiency — BLIP-2 achieves state-of-the-art performance on visual question answering at a fraction of the computational cost of end-to-end trained systems.

Frozen Vision Model
↓
≈ 256 visual features
↓
Q-Former — 32 learned query vectors (bottleneck)
↓
Frozen Language Model
↓
Text output

Fig. 6. BLIP-2's Q-Former as a learned bottleneck, compressing ≈256 visual features into 32 query vectors passed to a frozen language model.

D. LLaVA

LLaVA (University of Wisconsin, 2023) [3] demonstrated that even a simple linear projection layer connecting CLIP's vision encoder to a large conversational language model (LLaMA/Vicuna) could produce powerful visual instruction following. A user can send LLaVA an image and ask any question about it in natural language. LLaVA's success supports an important IB insight: when each modality has already been compressed through pre-training, even a simple bridge between their representation spaces can enable sophisticated multimodal reasoning.

E. Flamingo

Flamingo (DeepMind, 2022) [4] inserted cross-attention layers into a frozen language model, allowing it to selectively query visual representations at each step of language generation. Rather than forcing all visual information into the language model at once, Flamingo retrieves only the visual information relevant to the current point in reasoning. This selective retrieval is architecturally consistent with IB principles and gave Flamingo exceptional few-shot multimodal learning ability — generalising to new tasks from just two or three examples.

Flamingo's architecture offers an important lesson: the most effective integration of multiple modalities is not brute-force concatenation but intelligent, context-sensitive selection. At each step of generating a language response, Flamingo asks: 'What visual information, if any, is relevant right now?' This question-and-answer structure is IB in action at the architectural level — compress visual information on demand, retain only what is needed for the current reasoning step, discard the rest.

This on-demand compression explains Flamingo's few-shot strength. Because the model never overwhelms its language processing with unnecessary visual detail, it can attend carefully to the small number of task-relevant visual cues provided in a few-shot prompt. A model that forced all visual information through at every step would drown these subtle cues in noise.

V. IB IN MULTIMODAL LEARNING: KEY RESEARCH

Table I summarises the key papers surveyed in this work, their publication venues, and their primary contributions. Together they demonstrate the rapid growth of explicit IB applications in multimodal AI between 2023 and 2026.

TABLE I. Summary of Key IB Research Papers

Ref	Paper / Authors	Venue	Key Contribution
[5]	Tishby & Zaslavsky	IEEE ITW 2015	IB two-phase theory for deep networks
[6]	Alemi et al.	ICLR 2017	Deep VIB: practical variational IB
[10]	Wang et al.	IEEE TMM 2023	MIB: 5-property multimodal IB framework
[11]	Wang, Rudner & Wilson	NeurIPS 2024	M2IB: IB-based multimodal attribution
[9]	Hu et al.	arXiv 2025	CIBR: IB regularisation for CLIP
[12]	Chen et al.	ICML 2025	OMIB: per-modality adaptive compression
[13]	Zhu et al.	arXiv 2025	NIB: multimodal interpretability via IB
[15]	Tian et al.	arXiv Feb 2026	Layer-wise visual absorption analysis

A. Multimodal IB Framework (MIB)

Wang et al. (IEEE Transactions on Multimedia, 2023) [10] proposed the Multimodal Information Bottleneck framework as a unified theory for multimodal learning. MIB identifies five properties that a good multimodal representation should satisfy: consistency (same behaviour regardless of which modality provides the information), specificity (each modality's unique contribution is preserved), complementarity (modality-unique information is not lost), sufficiency (all task-needed information is retained), and conciseness (no irrelevant information is included). MIB provides a formal way to measure and optimise for all five simultaneously.

B. Explainability via IB (M2IB)

Wang, Rudner, and Wilson (NeurIPS 2024) [11] proposed M2IB — Multimodal Information Bottleneck Attribution — to explain why a multimodal AI such as CLIP makes a particular decision. Standard models are black boxes. M2IB uses IB compression to identify which visual regions and which words in the text drove a specific prediction. It outperforms existing methods such as SHAP and GradCAM [25] in attribution accuracy and is especially valuable in healthcare, where clinicians need to verify that an AI diagnostic tool is attending to the correct parts of an X-ray or medical report.

C. Cross-Modal Imbalance (OMIB)

Images are information-dense; text is sparse. Schrodi et al. (ECCV 2024) [14] showed that standard training systematically under-represents text in contrastive models. Chen et al. (ICML 2025) [12] addressed this with the Optimal Multimodal IB (OMIB) framework, which dynamically adjusts the compression rate per modality based on how much task-relevant information that modality carries. Images, being information-dense, are compressed more aggressively; text, being sparser, is compressed more gently. This adaptive approach produced consistent improvements on multimodal retrieval and classification benchmarks.

D. Layer-Wise Visual Absorption

Tian et al. (February 2026) [15] performed a layer-by-layer information-theoretic analysis of multimodal LLMs, tracking how visual information propagates through transformer layers and is converted into linguistic reasoning. They found that visual information is not uniformly distributed — specific layers act as natural IB bottlenecks where vision is compressed and absorbed into language. This provides practical architectural guidance: designers can deliberately strengthen these natural bottleneck layers for more efficient and accurate multimodal models.

VI. ADVANTAGES OF IB IN MULTIMODAL AI

A. Improved Generalisation

Because IB discards idiosyncratic training details and retains only universal task-relevant patterns, IB-trained models perform substantially better on unseen data. CIBR [9] demonstrated this with consistent improvements on seven diverse image classification datasets under zero-shot conditions — the model had never been specifically trained for those tasks or domains.

B. Robustness to Adversarial Attacks

Adversarial attacks introduce small, carefully crafted perturbations to inputs that cause standard models to make catastrophically wrong predictions [26] — a sticker on a stop sign fooling a self-driving car's vision system. Because IB models are trained to compress away details irrelevant to the prediction task, such small structured perturbations are treated as noise and discarded. Deep VIB [6] demonstrated significantly higher accuracy under adversarial conditions than standard neural networks.

C. Explainability and Transparency

IB-based models offer a natural pathway to explainability: the information that survives compression is precisely the evidence that drove the prediction. M2IB [11] operationalises this into a practical tool, identifying the specific image regions and text tokens that were most influential. This kind of grounded, theory-backed explainability is essential for deploying AI in medicine, law, and finance where trust and accountability are non-negotiable.

D. Computational Efficiency

Compressing information has practical computational benefits. BLIP-2's Q-Former architecture reduces the number of visual features passed to the language model from thousands to a small fixed number of query vectors, dramatically cutting memory and compute requirements while maintaining high accuracy. As AI moves to edge devices — smartphones, medical tablets, embedded sensors — this efficiency matters enormously.

E. Fair Multimodal Fusion

Without IB, information-dense modalities dominate fusion. OMIB [12] ensures equitable treatment: each modality is compressed to its task-relevant content before fusion, so neither vision nor language overwhelms the other. The result is more balanced and accurate multimodal understanding, particularly on tasks where the critical information is split across modalities.

VII. REAL-WORLD APPLICATIONS

A. Healthcare

Medical AI must integrate imaging data, written clinical notes, and laboratory results simultaneously. IB-based models improve generalisation across different hospitals and imaging equipment by focusing on clinically meaningful features rather than equipment-specific artefacts. M2IB provides the explainability that clinicians and regulators require — identifying exactly which regions of a scan and which words in a patient's notes drove a diagnostic recommendation.

B. Autonomous Vehicles

Self-driving cars receive simultaneous streams from cameras, LiDAR, radar, and GPS. IB compression focuses computation on safety-critical signals — nearby vehicles, road boundaries, traffic signals — and discards background noise. Research on IB-based multimodal reinforcement learning [16] showed that IB-trained agents with multiple sensors are more robust to sensor noise and environmental variation than standard approaches.

C. Content Moderation

Harmful content is often only detectable when image and caption are read together. Multimodal IB models that compress to cross-modal semantic meaning — rather than unimodal surface features — detect subtle, coordinated harmful content that image-only or text-only systems would miss. The improved generalisation of IB training is also critical as harmful content continuously evolves in form.

D. Education Technology

A tutoring system that reads a student's written answer, observes their facial expression, and hears their verbal explanation has far richer information available than a text-only system. IB-based multimodal fusion can identify the specific conceptual gap in the student's reasoning — ignoring irrelevant signals such as handwriting variation or background noise — and provide precisely targeted instruction.

VIII. OPEN CHALLENGES

A. Mutual Information Estimation at Scale

All IB methods rely on measuring mutual information in high-dimensional spaces — a computation that is infeasible to perform exactly. Current approximations (variational bounds, neural estimators such as MINE [17]) are tractable but may introduce systematic errors. As models grow to hundreds of billions of parameters, these approximation challenges become more severe. Developing fast, accurate, and scalable MI estimators is the most critical open problem in this area.

B. IB for Generative Models

Almost all IB research has focused on discriminative models that classify inputs. The most publicly impactful AI today — DALL-E, Stable Diffusion, ChatGPT — are generative models. How to apply IB principles to ensure that generative models retain relevant creative capacity while discarding harmful or irrelevant content patterns is almost entirely unexplored and represents a major research frontier.

C. Frontier-Scale Application

IB research has been validated mainly on small- to medium-scale models. Whether its benefits — improved generalisation, robustness, and explainability — translate reliably to billion-parameter frontier models such as GPT-4V and Gemini Ultra is an open empirical question. Developing computationally feasible IB training procedures for frontier-scale systems is essential for the field's practical impact.

D. Universal Evaluation Metrics

Current multimodal evaluation metrics are task-specific (CIDEr for captioning, CLIPScore for alignment) and fragmented. IB theory, grounded in information theory, suggests a path toward universal, task-agnostic metrics that measure the quality of a compressed multimodal representation directly. Such metrics would enable more principled model comparison and clearer design targets.

IX. CONCLUSION

This survey has examined the intersection of Information Bottleneck theory and multimodal foundation models. The core insight — retain what is relevant, discard what is not — is both theoretically elegant and practically powerful. Applied to multimodal AI, it addresses cross-modal information imbalance, overfitting to irrelevant detail, the demand for explainable decisions, and the computational cost of large-scale processing.

Models such as CLIP, BLIP-2, LLaVA, and Flamingo already implement IB principles implicitly. Explicit IB frameworks — MIB, OMIB, M2IB, CIBR — demonstrate that making these principles explicit and formal produces measurable improvements in generalisation, robustness, efficiency, and transparency. The advantages are felt in healthcare diagnostics, autonomous vehicles, content moderation, and personalised education.

Significant challenges remain in scaling MI estimation, extending IB to generative models, and developing universal evaluation metrics. These are also the most promising opportunities for researchers entering the field today. For MCA students and early-career researchers, this intersection of information theory and multimodal AI

is a frontier where theoretical insight and practical impact go hand in hand.

Looking ahead, the integration of IB into the training pipelines of frontier models represents perhaps the single most impactful direction for this research area. If the largest and most capable multimodal AI systems — those deployed in hospitals, embedded in vehicles, and integrated into educational platforms — could be trained with explicit IB objectives, the improvements in safety, reliability, and transparency would be felt at global scale. The groundwork laid by the papers surveyed here makes this future not only desirable but increasingly reachable.

Acknowledgement

The author thanks the faculty of the MCA Programme at Guru Nanak Institute of Management Studies, Mumbai for their guidance and support during the preparation of this survey. Special gratitude is extended to the research community whose work is surveyed here, and whose open publication of pre-prints on arXiv has made this rapidly evolving field accessible to students worldwide.

REFERENCES

1. A. Radford et al., "Learning transferable visual models from natural language supervision," ICML, 2021.
2. J. Li et al., "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," ICML, 2023.
3. H. Liu et al., "Visual instruction tuning (LLaVA)," NeurIPS, 2023.
4. J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," NeurIPS, 2022.
5. N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," IEEE ITW, 2015. arXiv:1503.02406.
6. A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," ICLR, 2017. arXiv:1612.00410.
7. N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," Allerton Conf., 1999.
8. R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," arXiv:1703.00810, 2017.
9. H. Hu et al., "CIBR: Cross-modal information bottleneck regularization for robust CLIP generalization," arXiv:2503.24182, 2025.
10. Y. Wang et al., "Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations," IEEE Trans. Multimedia, 2023.
11. Y. Wang, T. G. J. Rudner, and A. G. Wilson, "Visual explanations of image-text representations via multi-modal information bottleneck attribution," NeurIPS, 2024.
12. X. Chen et al., "Learning optimal multimodal information bottleneck representations (OMIB)," ICML, 2025.
13. Z. Zhu et al., "Narrowing information bottleneck theory for multimodal image-text representations interpretability," arXiv:2502.14889, 2025.
14. S. Schrodi et al., "Towards understanding cross-modal information imbalance in multimodal learning," ECCV, 2024.
15. J. Tian et al., "How vision becomes language: A layer-wise information-theoretic analysis of multimodal reasoning," arXiv:2602.15580, Feb. 2026.
16. H. Huang et al., "Multimodal information bottleneck for deep reinforcement learning with multiple sensors," arXiv:2410.17551, 2024.
17. M. I. Belghazi et al., "MINE: Mutual information neural estimation," ICML, 2018.
18. W. Wan et al., "A survey on information bottleneck," IEEE Trans. Pattern Anal. Mach. Intell., 2024.
19. L. Zhang et al., "Aligning multimodal representations through an information bottleneck," ICML, 2025.
20. R. Bommasani et al., "On the opportunities and risks of foundation models," arXiv:2108.07258, 2021.
21. C. E. Shannon, "A mathematical theory of communication," Bell Syst. Tech. J., vol. 27, pp. 379–423, 623–656, 1948.
22. A. Vaswani et al., "Attention is all you need," NeurIPS, 2017.

23. A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," ICLR, 2021.
24. D. P. Kingma and M. Welling, "Auto-encoding variational bayes," ICLR, 2014. arXiv:1312.6114.
25. R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," ICCV, 2017.
26. I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," ICLR, 2015. arXiv:1412.6572.