Integrating Sensation into Neural Networks

Including real-world **sensation** (vision, hearing, touch, etc.) directly into AI models has become a major research focus in recent years. Researchers are working on neural networks that don't just rely on textual descriptions or separate pre-trained encoders (like CLIP) for perception, but can *truly* perceive – effectively giving AI systems something analogous to eyes, ears, and other senses. Below, we survey key developments in this area, from end-to-end sensory processing to multimodal and brain-inspired architectures. Throughout, the theme is leveraging the "knowledge of giants" – building on prior innovations so we don't have to reinvent the wheel.

End-to-End Perception and "Skipping Steps" of Sensation

A longstanding goal has been to feed raw sensor data directly into neural networks, bypassing hand-crafted preprocessing – much like feeding sensations straight to a brain. In humans, for example, the **Neurophone** invented by Patrick Flanagan (patented in 1968) demonstrated that one can **skip normal sensory organs**: it transmitted sound through ultrasonic waves applied to the skin, letting the brain perceive audio without using the ears [1] [2]. This idea of bypassing intermediate steps has its analogue in AI. Modern deep learning allows **end-to-end learning** from raw inputs, meaning a neural net can take in **raw pixels** from a camera or **raw waveforms** from a microphone and learn to make sense of them without manual feature extraction. Early successes of convolutional neural networks (CNNs) for vision and recurrent nets for audio showed that if you give the model *direct sensory input*, it can learn the necessary feature processing internally. This end-to-end approach "skips" traditional engineered features (just as Neurophone skipped the eardrum), allowing models to **"see" or "hear" directly from sensors**. The result is AI that handles sensation in a more brain-like way – for instance, today's speech recognition models often take raw audio and outperform older systems that relied on hand-crafted spectral features, and vision models learn directly from raw image pixels.

Multimodal AI Models That Can See and Hear

Research labs have realized that **multimodal perception** – combining vision, hearing, language, etc. – is crucial for more general intelligence. Microsoft's researchers argue that "multimodal perception is a necessity to achieve artificial general intelligence, in terms of knowledge acquisition and grounding to the real world" [3]. In 2023–2025 we've seen a wave of AI models that integrate multiple sensory modalities:

- **OpenAI GPT-4 (2023):** A large-scale multimodal model that *accepts image inputs in addition to text*, and produces text outputs [4]. GPT-4 can analyze images (e.g. identifying what's in a photo or reading a diagram) as part of its prompt, rather than relying on the user to describe the image in words. This was a big step beyond GPT-3.5, which only accepted text [5]. OpenAI demonstrated GPT-4 solving tasks like explaining what's funny in an image or answering questions about a diagram directly from the image content. This shows that a single neural net can *see* and reason jointly. (GPT-4's vision ability was initially announced in 2023, although broad access to the image-input feature rolled out later.)

- **Microsoft** Kosmos-1 **(2023):** A 1.6 billion-parameter **Multimodal Large Language Model** that was trained to handle images and text together [6]. Kosmos-1 can caption images, answer questions about visual content, read text in images (OCR), and even solve simple visual puzzles – all in natural language. It treats vision as another input to a language model. Notably, Kosmos-1 showed improved performance on certain tasks by having visual context, indicating that image understanding was feeding into its reasoning [6]. This model was an early demonstration that even relatively small models can learn to *perceive* and intermix modalities.

- **DeepMind** Flamingo **(2022):** An earlier multimodal model that also combines vision and language, used for tasks like image captioning and visual question answering. Flamingo introduced techniques for interleaving visual data into a language model's flow. Microsoft benchmarked Kosmos-1 against DeepMind's Flamingo on image Q&A and found comparable or better performance [7], showing the rapid progress in this area.

Illustration of a unified multimodal approach: Meta AI's ImageBind maps diverse sensory inputs – images, audio, text, depth (3D), thermal, and movement (IMU data) – into a single shared embedding space [8]. Models like this enable AI to link and integrate information across different senses.

- **Meta AI** ImageBind **(2023):** Rather than a traditional "language model," ImageBind is an **embedding model** that jointly learns representations for *six different modalities*: **visual** (images/video), **audio** (sound), **text** (natural language), **depth** (3D shape info), **thermal** (infrared heat vision), and **IMU sensor data** (motion/acceleration from devices) [8]. This is like giving the AI a sense of sight, hearing, touch (temperature/depth), and proprioception. Importantly, ImageBind learned to **align all these sensory inputs in one common feature space** without needing explicit pairwise data for every modality. For example, it was trained on image–audio pairs and image–depth pairs, and amazingly it learned implicit connections to other modalities as well [9]. In practical terms, ImageBind can take an input in one modality and retrieve or generate corresponding data in another. Meta's researchers demonstrated a striking case: feeding an **audio clip** (a spoken description) into the model and using its embedding to generate an image via a generative model – effectively producing an image **directly from sound** without any intermediate text transcription [10]. This cross-modal generation is akin to *skipping a step*: much like the Neurophone bypassed the normal hearing pathway, ImageBind can bypass text and go straight from sound to vision. The ability to bind multiple senses together in this way is pushing AI closer to a human-like holistic understanding of context.

- **Google** PaLM-E **(2023):** A 562 billion-parameter "**Embodied**" multimodal model that combines Google's PaLM language model with vision and sensor inputs for robotics [11]. PaLM-E is designed to interface with **robot sensors** – it can take **camera images** and other sensor readings as part of its input, and then output instructions or answers. For example, PaLM-E can answer a question like "What happened between image 1 and image 2?" by analyzing two input images (encoded through a Vision Transformer) alongside the text query [12]. Under the hood, all the different inputs (visual data, robot's state, etc.) are encoded into the same token embedding space that the language model uses, so the model is essentially reading a **"multimodal sentence"** that includes observations of the world [13]. This approach grounds the AI in the physical context. Impressively, PaLM-E can then output high-level *textual commands for the robot's actuators* or answer questions about the scene. It not only performed well on robotic tasks (like guiding a robot to grasp objects), but also **outperformed previous models on a visual question-answering benchmark (OK-VQA)**, showing robust multimodal reasoning [11]. In essence, PaLM-E *sees* through the robot's eyes and *thinks* in natural language. Google's team notes that PaLM-E points toward unifying what were separate tasks – vision, language understanding, and robotics – in one general model [14]. This is a clear example of leveraging sensors (camera, robotic state) in an AI model to give it embodied perception and reasoning.

Architecture of Google's PaLM-E, an embodied multi-modal model. A vision transformer (ViT) processes images from the robot's camera, and other sensors (like pose or environment state) are encoded, then all are inserted as input tokens alongside text into a language model (PaLM). The model's output is textual – for example, a description or an action plan for the robot [13] [12].

Beyond these, there are many other efforts: for instance, DeepMind's **Gato** (2022) was a single transformer model trained across images, text, and even robotic control data – aiming for a generalist agent. And **OpenAI's Whisper** (2022) is an end-to-end model that directly **hears** audio and transcribes it, without needing separate speech-recognition components. The clear trend is that **AI is moving toward multi-sensory integration**: models that *see*, *hear*, and *feel* (in a digital sense) rather than just process text. By encoding images, sounds, and other sensor readings directly, these systems don't have to rely on humans to describe every relevant detail – they can gather information from raw sensory sources.

Unified Sensory Representations and Encoding Over Description

A key theme in recent research is shifting from *describing* sensory input to AI (via text prompts) toward *encoding* those inputs directly for the AI to use. Earlier multimodal systems often worked by generating textual descriptions (captions) of an image which a language model could then read. But this is inefficient and lossy – it's like trying to explain a complex picture in words. Now, with approaches like those above, the image itself (or its feature encoding) can be fed into the model, preserving much more information than a caption. As an example, Microsoft's Kosmos-1 could read text on images and understand visual puzzles directly [6], rather than requiring the user to type out what they see. Likewise, PaLM-E's design of embedding visual tokens among text tokens means the model can reason on visual evidence without an intermediate narration [13]. Researchers have found that giving models direct perceptual inputs can even improve their purely textual reasoning, because it grounds them in reality – e.g. Kosmos-1 transferred its visual understanding to do better on some language tasks involving physical properties [15].

The **unified embedding spaces** (like in ImageBind) are a big innovation here. By mapping different senses to a common representation, an AI can associate, say, an image with the sound that typically accompanies it, or an action with the sight of its outcome. This goes beyond a simple input-output pairing – it's learning the *conceptual connections* between modalities. Meta's demo of generating an image from an audio clip (speech-to-image) with ImageBind is a vivid example: the model essentially "understands" the audio in an abstract way and finds the matching visual concept for image generation [10]. Such direct cross-modal encoding is fundamentally different from the old pipeline of "describe the audio with text, then feed to an image model." It skips those steps, using a *shared sensory language* of vectors learned from data.

In short, the field is moving toward **direct encoding of sensations**. Instead of forcing the AI (or user) to translate sensory data into words, we embed the data in a form the model can natively reason with. This is analogous to how the human brain doesn't narrate everything it sees; it processes the raw signals and integrates them into thought. Recent work is all about giving neural networks that same ability to directly take in sights, sounds, and other sensor readings.

Embodied AI and Sensor Fusion in Neural Networks

When you mention using a **cell phone's sensors** (camera, microphone, accelerometer, etc.) as an AI's input, you're essentially talking about **sensor fusion** in AI – combining multiple streams of sensory data to make decisions. This is a well-known problem in robotics and autonomous systems, and deep learning is increasingly being applied here too. A classic example is self-driving cars: they fuse cameras, LiDAR, radar, GPS, IMU (motion sensors) to perceive the environment. Initially, much of that fusion was done with separate systems (e.g. one network processes camera images, another processes LiDAR point clouds, then a rule-based system merges the outputs). But newer approaches use unified neural architectures to process all sensor inputs jointly or in a tightly integrated way.

The Penn State research on an **"artificial, multisensory integrated neuron"** is a fascinating bio-inspired instance of low-level sensor fusion [16] [17]. In that work, they literally built a *single device* (a memtransistor-based circuit) that takes in both visual input (light) and tactile input (pressure) at the same time [18]. The sensors influence each other directly, mimicking how in our brain, senses can modulate one another (e.g. a quick flash of light can prime our sense of touch in the dark) [19]. They found that when both the light and touch signals were weak, the combined artificial neuron could still produce a strong response, whereas each modality alone might not trigger detection [20]. In other words, **"the collective sum of biological inputs can be greater than their individual contributions"**, and their device captures that effect [21]. This is a hardware-level demonstration, but the principle carries to AI models: by fusing multiple sensors, an AI can detect and understand situations that single-sense AI might miss. It's also more efficient – the Penn State team noted that letting sensors talk to each other directly (like in the neuron) could save energy and speed, compared to each sensor reporting to a central processor separately [22] [23].

On a larger scale, **embodied AI** refers to AI agents that exist in and interact with the world (robots, drones, etc.). For such agents, processing various sensor modalities is crucial. Google's PaLM-E (discussed above) is one example of an embodied AI controller that uses sensor fusion at the neural network input level. Another example is work by researchers at Carnegie Mellon and others on **"sensorimotor networks"** where a single network takes in camera pixels and inertial measurement data and outputs navigation decisions for robots. These systems often leverage **recurrent neural networks or transformers** to handle time-series sensor data (like accelerometer streams) combined with images. The *World Models* research by David Ha et al. (2018) also explored an agent that builds an internal model of the world from a visual input (a simplified car racing game) and then uses it to plan actions – effectively learning a "sensory→cognitive→action" loop internally.

Crucially, integrating sensors into neural nets doesn't necessarily require reinventing the wheel; it often means using **encoders** for each sensor type to translate their readings into a common format that a central model can understand. This is exactly the strategy of PaLM-E (vision encoder + state encoder + language encoder feeding into one transformer) [13]. It's also seen in projects like **IBM's NEON** pathfinding robot, or the Allen Institute's work on embodied agents that use language models augmented with image inputs (e.g., an AI that can navigate by interpreting visual cues and following instructions). The trend is clear: **AI researchers are actively grafting senses onto neural nets** and training them in end-to-end fashion to handle the deluge of data from those senses.

Brain-Inspired Architectures and Cognitive Integration

Your mention of Sapient's **Hierarchical Reasoning Model (HRM)** is a great example of another frontier: injecting cognitive *structure* (inspired by neuroscience and psychology) into AI models. HRM was introduced in mid-2025 as a novel architecture that achieves remarkable reasoning performance with only ~27 million parameters [24]. Its secret is taking inspiration from the brain's **hierarchical, multi-timescale processing**: it has a two-level recurrent design, where a **high-level module** operates slowly to do abstract planning, and a **low-level module** works fast to fill in details [25]. During a single forward-pass, the model iteratively uses these modules (without needing step-by-step supervision, unlike typical chain-of-thought prompting). The result is that HRM can solve complex problems – e.g. it achieved nearly perfect scores on hard tasks like Sudoku puzzles and finding optimal paths in mazes, and it beat models 10x larger on the Abstraction and Reasoning Corpus (ARC) challenge [26]. This was achieved with very limited training data (only 1,000 examples in some cases) [27]. Such efficiency hints that built-in cognitive biases (here, hierarchical reasoning and temporal abstraction) let it generalize much better from few examples, much like human cognitive architecture does.

Why is HRM relevant to "sensation in neural nets"? While HRM itself was demonstrated mostly on symbolic or textual reasoning tasks, it represents a *brain-inspired framework* that could, in principle, be extended to incorporate sensory inputs. The human brain doesn't process vision or sound in one flat sequence; it has hierarchical layers and feedback loops. Models like HRM suggest we might not need billions of parameters *if* we architect the network to reason in a human-like way. In the future, one could imagine a Hierarchical Reasoning Model that takes in multi-modal sensory data: a high-level planner that reasons about what it sees/hears ("slow thinking"), combined with low-level neural circuits that handle the immediate sensorimotor processing ("fast reactive thinking"). The **cognition theory baked in** here – that intelligent behavior emerges from interacting fast and slow processes – could synergize with rich sensory inputs to yield very capable AI on a modest compute budget.

Another area of brain-inspired work is **neuromorphic computing** and **spiking neural networks** for sensory processing. For instance, researchers are studying the olfactory (smell) system's circuitry to design better AI olfaction – the Cornell study from 2025 showed how the early olfactory brain creates a "firewall" to organize chaotic chemical signals into stable representations, and they drew parallels to quantization in machine learning [28] [29]. The motivation is to achieve the brain's efficiency – the human brain uses very little power – by copying how biological neurons encode and compress sensory data. Neuromorphic chips like Intel's **Loihi** have been used to directly interface with sensors (like event-based cameras or touch sensors) and run spiking neural nets that process the data with high efficiency. This is somewhat parallel to the question of "sensation in neural nets" – it's about *how* the raw sensation is represented and processed (spikes rather than continuous activations, in this case) for efficiency and robustness.

In summary, **cognitive and neuromorphic inspirations are influencing AI design**. HRM's success indicates we can integrate principles of human cognition (hierarchy, multiple timescales) to get better reasoning without brute force. And on the sensory side, mimicking biological sensor processing (like the multisensory neuron, or the olfactory pre-processing) can make AI more adept at handling raw, noisy inputs from the world. The long-term vision is an AI that *perceives* and *reasons* in a unified, human-like loop – seeing, hearing, planning, and acting all with brain-level efficiency.

Conclusions: Standing on the Shoulders of Giants

As you suspected, many **giants in AI research have laid groundwork** that you can build upon for an AI model that can truly "sense" the world. From the examples above, a few clear takeaways emerge:

- **Use integrated sensors:** Even a basic setup like a smartphone (camera, mic, accelerometer) can feed a rich stream of data to a neural network. Research in multimodal models shows that it's feasible to train models on combined inputs – e.g. an AI that takes an image and an audio clip together and outputs a decision or description. You don't have to start from scratch: architectures like transformers can be adapted to multi-sensor input (as done in Kosmos-1 and PaLM-E) by adding encoding layers for each sensor type [13].

- **Encode, don't just describe:** Instead of converting every sensory input to text descriptions, leverage models that accept non-text embeddings. CLIP was an initial bridge (image to text embeddings), but newer models go beyond, letting the AI directly ingest the *encoded sensation*. For instance, you might use a vision transformer to encode camera frames and feed those embeddings into your main model alongside textual data – essentially what PaLM-E and others do [12]. This way the AI's "internal language" can include the patterns of vision or sound. Meta's ImageBind further suggests you can have a single latent space where different sensors' data meets [8], enabling creative applications (like sound → image) that weren't possible before.

- **Learn from cognitive science:** The Hierarchical Reasoning Model shows that we can bake theories of cognition into our model architecture for big gains in performance and efficiency [25] [26]. If your goal is an AI that perceives and also **understands** and reasons about what it perceives, consider hierarchical or modular designs. Classic "perception-action" loops in robotics (perceive → plan → act) are now being folded into end-to-end neural models; with inspiration from the brain, one can design networks that naturally handle these phases. For example, a two-module system (vision module feeds into a reasoning module) could be trained together – somewhat analogous to how our occipital lobe processes visuals and then frontal cortex plans with that information.

- **Leverage existing models and data:** Standing on giants' shoulders also means using pre-trained components when appropriate. If you don't want to rely on CLIP, you might still take advantage of, say, an open-source vision backbone (like a ResNet or ViT pretrained on ImageNet or LAION images) to give your model a head start in "seeing." The same goes for audio (there are models pretrained on AudioSet or waveforms) and other senses. The current trend is **foundation models** – large pretrained models for each modality – and then **connecting or unifying** them. Projects like ImageBind show you don't need a labeled dataset for every combination of modalities; pairing a few (like images with audio, images with depth) was enough to align all six modalities [9]. So you can use the knowledge encoded in such foundation models to avoid collecting massive new datasets yourself.

In conclusion, **yes – a lot of research has already tackled making AI models that can "feel" and "perceive."** From the Neurophone's legacy of bypassing sensory organs [1] to today's multimodal transformers that *see and hear*, the path is laid out to create AI with genuine sensory integration. The exciting part is combining these advances: imagine taking a model like HRM (with its human-like reasoning prowess) and hooking it up to ImageBind's multi-sensory input streams – you'd get a compact, brain-inspired system that can intake real-world sights and sounds and *think* about them. While we're not all the way to human-level AI, the work done by these “giants” provides a rich toolkit. By learning from their approaches – end-to-end sensor learning, unified embeddings, embodied integration, and cognitive architectures – you can develop a new model that truly **sees** (and hears, and more) from scratch, without wasting the hard-earned knowledge already out there. The consensus in the field is that integrating all the senses and cognition is key to next-generation AI, so it’s a great direction to pursue, armed with the methods pioneered in recent research. Good luck on building on these foundations!

Sources: The answer incorporates findings and quotes from recent AI research and news, including Meta’s ImageBind model for multi-sensory embeddings [8] [10], Microsoft’s Kosmos-1 multimodal LLM [6] [3], Google’s PaLM-E embodied model [13] [11], the Sapient Hierarchical Reasoning Model paper [25] [26], the Neurophone patent description [1] [2], and a Penn State study on multisensory neurons [17] [20], among others. Each of these demonstrates progress in giving neural networks sensory capabilities and cognitive structures analogous to human perception.

---

[1] [2] US3393279A - Nervous system excitation device - Google Patents
https://patents.google.com/patent/US3393279A/en

[3] [6] [7] [15] Microsoft's Kosmos-1 is a multimodal step toward more general AI
https://the-decoder.com/microsofts-kosmos-1-is-a-multimodal-step-toward-more-general-ai/

[4] [2303.08774] GPT-4 Technical Report
https://arxiv.org/abs/2303.08774

[5] OpenAI releases GPT-4, a multimodal AI that it claims is state-of-the-art
https://techcrunch.com/2023/03/14/openai-releases-gpt-4-ai-that-it-claims-is-state-of-the-art/

[8] [9] [10] What is ImageBind? A Deep Dive
https://blog.roboflow.com/what-is-imagebind/

[11] [12] [13] [14] Google's PaLM-E Combines Vision and Language AI for Robot Control - InfoQ
https://www.infoq.com/news/2023/06/google-palm-e-robot/

[16] [17] [18] [19] [20] [21] [22] [23] Making AI smarter with an artificial, multisensory integrated neuron | Penn State University
https://www.psu.edu/news/engineering/story/making-ai-smarter-artificial-multisensory-integrated-neuron

[24] [25] [26] [27] [2506.21734] Hierarchical Reasoning Model
https://arxiv.org/abs/2506.21734

[28] [29] Brain-inspired AI model learns sensory data efficiently
https://as.cornell.edu/news/brain-inspired-ai-model-learns-sensory-data-efficiently