International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT-V6I3P117
Eureka Vision Publication | Volume 6, Issue 3, 112-120, 2025

Original Article

Generalist Vision Models for Any-to-Any Image-to-Video Understanding

Sajud Hamza Elinjulliparambil
Pace University.

Received On: 27/06/2025 ***Revised On: 20/07/2025*** ***Accepted On: 01/08/2025*** ***Published On: 24/08/2025***

Abstract - Current developments in multimodal foundation models are driving computer vision beyond sets of task-specific systems to generalist vision models that can be used to do a large number of tasks and modalities in the same architecture. Simultaneously, video understanding has transitioned off of specialized backbones and to large models that collectively reason over images, videos and language. Any-to-any vision models strive to bring such a trend together: they take in heterogeneous visual and textual inputs (e.g. image to caption, video to action labels, of image+text to edited image) and deliver heterogeneous outputs via a common interface. The article examines the new space of generalist vision models of any-to-any image-to-video understanding. We introduce the concept of any-to-any modeling first and place it in the context of the rest of the literature on multitask and multimodal learning. We next outline typical models of such as Unified-IO 2, UnIVAL, PaLi-3, and 4M-21 that handle a broad input/output model of images, videos, audio, dense labels and free-form language [1] . We draw attention to general building blocks (unified tokenization, transformer backbones, diffusion or autoregressive heads) and training strategies (large-scale pretraining, instruction tuning, and multi-task curricula). We comment on the performance of these models on image and video benchmark tasks, including question answering, captioning, and spatiotemporal reasoning by using public benchmark results. Lastly, we provide practical and ethical advice in the deployment of generalist vision models in practice and provide open issues, such as unified evaluation in any-to-any context, efficient management of long videos sequences, and safety of open-ended visual interaction. We would like to offer a systematic and human-readable introduction that can assist the researcher and practitioner that is interested in constructing or using generalist image-to-video understanding systems.

Keywords - Generalist Vision Models, Any-To-Any Modeling, Multimodal Learning, Image Understanding, Video Understanding, Vision-Language Models, Unified-IO, Pali-3, 4M-21.

1. Introduction

The past decade in computer vision has been marked by two trends that are running in parallel Foundation models trained on large datasets, and multimodal models that bring vision and natural language into agreement [2]. Although the early systems were image-based, more recent applications have generalized these concepts to video, audio and even action spaces, and again to an increasing number of tasks depending upon a single back-bone. Meanwhile, the community has begun to stop using the notions of one-directedness (e.g., image to caption) and instead use more flexible any-to-any formulations: models can accept various combinations of inputs (images, videos, text, audio, sparse annotations) and give various types of output (caption, dense masks, temporal labels, edited images, next-frame prediction, etc.) via a common interface [3].

In the context of image-to-video understanding, this shift is especially important:

- Practical applications can be a combination of images and video: e.g. a system might be required to identify objects on a still image, answer questions about a short video, and then to be able to do reasoning on a long video with subtitles attached.
- It is costly to keep distinct models in each task and modality and to be able to transfer knowledge between tasks.
- Generalist models will be more sample efficient and can be deployed more easily, but require much more complex pretraining and system design.

1.1. From task-specific to generalist vision

Classical vision pipelines employed dedicated networks: there was a classification network, a detection network, a segmentation network, and dedicated video networks, including networks to perform action recognition. Even those which all used convolutional neural networks (CNNs) varied with respect to their heads, the input resolutions, and training data. This has been altered by transformer-based architectures and scale-to-large-pretraining. The identical visual backbone can now be used to support numerous tasks through the use of various heads or by being used to represent all things, images, videos, labels and text as sequences of tokens that can be run through a single transformer [4].

1.2. Any-to-any image-to-video understanding

In the literature, the term any-to-any is an informal one that has recently been attempted to be put into practice: a single model is supposed to be capable of many input and output modalities with only slight architectural modifications. The case of Apple, with 4M-21 model, is a good example, which is trained on tens of visual modalities and tasks, and is designed to map arbitrary combinations of input tokens to arbitrary combinations of output tokens in a unified space.

For image-to-video understanding, “any-to-any” means:

- Input can be in the form of one image, multiple images, a video clip or a combination of text and image.
- Production can take the form of a label (categorization), a caption or responses to questions, dense pixel-wise maps, time delimits, or even freshly generated frames.
- All these behaviors are often controlled by the same interface having an instruction-style text prompt and having visual tokens.

1.3. Scope and contributions of this review

The review will concentrate on generalist vision models that can process both images and videos as inputs, give various kinds of outputs (class labels, captions, dense predictions, etc.) and operate on a single architecture rather than training models individually with each task. We do not seek to discuss all of the specialized video models or all the image-only foundation models; rather, we highlight systems among which the design is clearly that of a generalist or any-to-any.

Our main contributions are:

- A clear conceptualization of any-to-any image-to-video understanding.
- A structured survey of representative models (Unified-IO 2, UnIVAL, PaLI-3, 4M-21, and others).
- A synthesis of common design patterns in architecture and training.
- A discussion of current capabilities, limitations, and open research directions.

2. Background: Multimodal and Generalist Vision Models

2.1. From multimodal to any-to-any

The earliest experiments on multimodal computing vision were mainly concerned with pairwise correlations between a small number of modalities. The first generation of multimodal systems, including Show and Tell along with Show, Attend and Tell, were classic image captioning models, which were usually limited to one direction image to text. The tasks involving vision-language tasks would later be extended to vision-language task-based visual question answering (VQA), grounded language understanding, cross-modal retrieval, referring expression comprehension and embodied navigation [5]. Nevertheless, all tasks typically needed specific architecture or education pipeline, which resulted in fragmentation across datasets and the research communities. The coming up with vision-language models (VLMs) was a shift to unified modeling whereby a single model was capable of undertaking multiple tasks without architectural modification. Contrastive training models (e.g., CLIP) supported zero-shot transfer to downstream tasks, and encoder-decoder models further supported text generation. With these improvements, even most systems were working on a narrow modality pair, like image ↔ text or video ↔ text. These limits are far much more than the modern paradigm of any-to-any multimodality. Rather than considering modalities as disjointed sets of two, any-to-any models strive to permit arbitrary input/output combinations using a common computational core. It can be accepted by a generalist system:

- Inputs: text, static images, video clips, audio waveforms, sensor inputs, or their combinations
- Outputs: captions, dense masks, bounding boxes, action tokens, rendered images, or synthesized video.

2.2. Vision-language foundation models

The earliest experiments on multimodal computing vision were mainly concerned with pairwise correlations between a small number of modalities. The first generation of multimodal systems, including Show and Tell along with Show, Attend and Tell, were classic image captioning models, which were usually limited to one direction image to text. The tasks involving vision-language tasks would later be extended to vision-language task-based visual question answering (VQA), grounded language understanding, cross-modal retrieval, referring expression comprehension and embodied navigation. Nevertheless, all tasks typically needed specific architecture or education pipeline, which resulted in fragmentation across datasets and the research communities. The coming up with vision-language models (VLMs) was a shift to unified modeling whereby a single model was capable of undertaking multiple tasks without architectural modification. Contrastive trained models (e.g., CLIP) made it possible to transfer zero-shot to downstream tasks, and encoder-decoder models were scaled up to text generation. With these improvements, even most systems were working on a narrow modality pair, like image ↔ text or video ↔ text [6].

These limits are far much more than the modern paradigm of any-to-any multimodality. Rather than using modalities as separate pairs, any-to-any models strive to allow arbitrary combinations of inputs and outputs via a common backbone of computations. It can be accepted by a generalist system:

- A shared encoder-decoder backbone that interprets heterogeneous signals in a consistent representation.
- Task generalization through instruction tuning, enabling flexible prompting rather than task-specific heads.
- Unified tokenization, mapping vision and language to a common symbol space.

  These characteristics form the conceptual bridge between traditional VLFMs and fully generalist image-video models.

2.3. Unified models for image, video, audio, and language

Parallel research is a line of research that addresses explicitly unified models in which multimodal processing is no longer considered an extension of language models, but rather a more general computational problem that cuts across a variety of data streams. These models consider the possibility of using a single backbone (typically based on transformers) with no domain variations. The influential example is UnIVAL (Unified Model for Image, Video, Audio and Language) which shows that single transformer could be competitive in image classification, video retrieval, audio understanding, and multimodal reasoning benchmarks. UnIVAL supports the fact that multimodal alignment can be best attained by unified modeling rather than expert modules trained on heterogeneous data by training on modality-specific tokenizers with a common core. The Unified-IO series widens this view to a great extent [7]. Unified-IO operates in an autoregressive style where all modality text, image patches, depth maps, audio spectrograms, video frames and segmentation masks are concatenated into a sequence of tokens. These sequences are then converted into structured outputs by the transformer backbone, effectively framing all tasks (e.g. video segmentation, image generation, VQA, audio captioning, robotics actions) as next-token prediction problems.

Unified-IO 2 enlarges the range of modality, and provides greater context length, higher tokenization fidelity, and a more efficient set of decoders. Notably, it demonstrates that big single-model training enables a single model to rival or surpass task-specific architectures in video reasoning, temporal segmentation and multimodal prediction. This convergence is an indication that specialisation models in domains can probably be overtaken by generalist systems that can model the world in any modality. The combined structures constitute the conceptual and technical foundation of the generalist vision models towards any-to-any image-to-video understanding. They show that multimodal learning is not just another way of involving more types of data, but of building general architectures that can allow arbitrary input-output transformations with high coherence in time and meaning.

3. Representative Generalist Vision Models

The generalist vision models have taken various architectural paradigms vision-language transformers, coherent multimodal systems, autoregressive token-based systems, and diffusion-driven perception models. This part takes a synthesis of representative systems that have influenced the landscape of any-to-any modeling with the attention to the contribution of each to coherent image-video understanding.

3.1. PaLI-3 and related vision-language models

PaLI-3 is a significant breakthrough in vision-language modeling as it closely integrates a strong ViT-based visual encoder with an mT5-like multi-lingual text decoder. In contrast to previous VLMs which were only trained on stationary images, PaLi-3 shows that large-scale image-text pre-training can surprisingly transfer to video reasoning tasks [8]. Even though the model is not specifically trained as an any-to-any system, its design already resembles a quasi-generalist system: one backbone is used to take in visual input, and a single textual decoder is used to work with a broad range of tasks. An interesting consequence of the design of PaLI-3 is its generalization to video question answering (VQA) benchmarks despite relatively small video specific finetuning. On MSR-VTT-QA, ActivityNet-QA, and NExT-QA, it has been demonstrated that the model is able to learn implicitly temporally grounded semantics and transient causal dynamics between frames without requiring dense video supervision as long as sufficiently large and multilingual training corpora are used. In addition, the language-neutral nature of PaLi-3 allows cross-linguistic reasoning, and tasks requiring the subject to answer questions about video content in previously unseen languages can be done. This is what allows PaLi-3 to play a key role in moving toward more general any-to-any systems and what is more to the point is to show that with concerted architectural decisions that gap between still image model and video based use can be truly narrowed.

3.2. UnIVAL

UnIVAL goes a step further with the concept of unification in that a model is explicitly constructed to be able to work across image, video and audio and language modalities all in a single common architecture. UnIVAL uses joint transformer backbone modality-specific encoders, rather than separate models targeting each task and each modality, which facilitates exchange of information across modalities. This guarantees that the system would be able to operate image captioning, video captioning, VQA, and audio-visual understanding with the same in the same sequence modeling approach. The simplicity of architecture to which this model devotes special attention is especially attractive. Even with a fairly small number of parameters relative to big foundation models, UnIVAL is able to cope with competitive performance on various benchmarks. This can be used to explain the fact that unification need not be restricted to billion parameter systems; all it needs is a uniform representation strategies and common reasoning layers [9]. UnIVAL shows that activities that have been conventionally considered distinct like time perception in video and space perception in image can be merged into a single platform without a substantial decline in performance. It is thus a good demonstration of how generalist design principles can be practiced even when there are limitations of model size and compute.

3.3. Unified-IO and Unified-IO 2

Unified-IO and its successor Unified-IO 2 are a couple of the most understandable operationalizations of the any-to-any paradigm. All inputs and outputs are represented as sequence of tokens of these models: text, images, video, audio, segmentation masks, bounding boxes, and even actions. In this tokenization approach, the system can do image generation, video captioning, segmentation, VQA, action recognition or multimodal tasks all with the same autoregressive mechanism. Everything, in other words, is understood as conditional sequence generation. Unified-IO 2 adds training on a larger scale, better modality embeddings, and more output representations to this framework. It is interesting to note that video is treated as also a tokenized modality, and frames are broken down into patch-sized or tubelet units, and passed into the model just like a tokenized text modality or a tokenized image modality. This allows the system to take on image-to-video tasks and video-to-video tasks without the need to have different branches in the architecture to take on the temporal modeling. The autoregressive interface is especially effective at unification: the model is taught to generate the appropriate output modality without any special heads or decoders to handle different tasks: rather, by simply being instructed to do a particular task and being conditioned to do it, the model will develop that behavior on its own. Consequently, Unified-IO 2 demonstrates the ability of a single transformer-based backbone to harmoniously address tasks that were traditionally unrelated into a single process (within a uniform token-generative process).

3.4. 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities

The model of the 4M-21 of Apple is among the most explicitly any-anywithin vision domain. In contrast to models that emphasize image-text or image video interactions, 4M-21 uses a wide variety of visual modalities RGB, depth, optical flow, surface normals, and others and works with a wide variety of prediction targets which include classification labels, segmentation masks, depth maps and generative outputs. Its training policy focuses on massive co-training on multitasks to produce the greatest number of cross-tasks transfer to the minimum negative transfer influence. Despite the fact that 4M-21 is vision oriented (without integrating the audio or long-form language features in deep) it proves that one single integrated model may lead to the state of the art outcomes in terms of many benchmarks [10]. One of the features that make 4M-21 stand out is that the system is scalable in its operation, it can accommodate tens of tasks and modalities at once, which confirms the feasibility of extensive unification even in limited, domain-specific contexts. This renders it especially applicable to generalist vision research, as it points out the practical engineering considerations dataset balancing, representation normalization, and shared losses which are required in large-scale multimodal co-training.

3.5. Diffusion-based generalist models

Most any-to-any models are constructed using autoregressive transformers, but diffusion models have recently become a new successful competitor to generalist perception and editing. Diffusion models have a very natural implementation of controlled generation and multistep refinement, making them an ideal choice when it comes to image editing, segmentation, keypoint detection, inpainting, and conditional synthesis. The InstructDiffusion is an example of the way diffusion systems can be reconfigured to do instruction-based generalist vision tasks. The model can also effectively implement a great number of low-level and mid-level vision tasks using one unified backbone, by conditioning the diffusion process on natural language instructions [11]. On the same note, DICEPTION is an architecture that employs a diffusion-based architecture but assigns it to solve perceptual tasks, utilizing pre-trained text-to-image diffusion models as general visual encoders. It attains competitive performance through specialized systems at the same time having a single architecture that can be used to deal with various tasks. Although currently several diffusion-based generalist models operate on images, they are inherently generalized to video by using temporal noise schedules and spatiotemporal UNets. This renders diffusion a viable strategy in any-to-any video understanding in the future, particularly in those tasks that need finer detail in spatial consistency across frames.

3.6. Emerging generalist reasoning models

In addition to the models that are perception-centric, there is a novel group of generalist systems that are task-based high-level visual reasoning. Large language models (LLMs) like OneThinker can be used to define models that are a central controller tasked with running on shared visual encoders to allow more flexible task specification, multi-step reasoning, and chain of thoughts-style visual processing [12]. Such systems usually have the capability of receiving both image and video input so that reasoning can be done on the time sequence, causality, and multi-frame visual clues all in a single framework. These models emphasize a move to the vision reasoning agent, in which multiple-task-performance is not only a concern of architectural convergence but also the capability to use shared knowledge in multitask. These reasoning-oriented models are able to do complex tasks in the long-range video setting with the use of LLM, including multi-frame inference, counterfactual reasoning, and instruction-following. This new type is therefore an intermediate between the prototypical VLMs and the ultimate cognitive any-to-any agents.

Table 1. Representative Generalist Vision Models Relevant to Any-to-Any Image-to-Video Understanding

Model	Modalities (I=Image, V=Video, A=Audio, T=Text)	Example Tasks (incl. video)	Interface Type	Any-to-Any? (qualitative)
PaLI-3	I, limited V, T	Image captioning, VQA, video QA, video captioning	Encoder–decoder (ViT + LLM)	Partially (vision to text)
UnIVAL	I, V, A, T	Image/video captioning, AV tasks, VQA	Unified transformer	Multi-input, text-centric
Unified-IO 2	I, V, A, T, actions	Generation, segmentation, action prediction	Autoregressive token model	Strong any-to-any
4M-21	I, depth, flow, other vision modalities	Classification, detection, depth, segmentation	Unified vision backbone	Vision-focused any-to-any

Instruct Diffusion	I, T	Editing, segmentation, keypoints, low-level tasks	Diffusion with instructions	Image-centric any-to-any
DICEPTION	I, T	Multiple perception tasks	Diffusion-based	Image any-to-any
One Thinker	I, (emerging V), T	Multiple vision tasks + reasoning	LLM-based	Reasoning-oriented

4. Architectural Design Patterns for Any-to-Any Image-to-Video Understanding

Despite differences in details, generalist models tend to share a few core design principles.

4.1. Unified tokenization and representation

Unified tokenization- making all modalities available to a single model in a form that they can be jointly processed by a single model is a basic requirement of any-to-any modeling. In the case of images and video, a patch-based tokenization based on ViT-like embeddings is typical, and in videos, a so-called tubelet, i.e., pixels in space and time, is used. Latent codes of generative systems are often discrete (such as those of VQ-VAE or VQ-GAN), where images or videos are encoded into a sequence of small tokens which can be decoded by a transformer model. Unified tokenization, in order to support mixed-modes sequences (e.g., interleaving text instructions with visual tokens) and to accept outcomes of any modality, is important. Any-to-any vision systems rely on this abstraction, since there is no need to have specialized branches of architecture that are specific to tasks [13]. The model instead acquires a common, task-blind representation space, in which it can project any input modality to any output. This positively contributes to flexibility as well as promoting robust cross-modal generalization.

4.2. Backbone architectures

Most generalist models rely on transformers as their main workhorse due to their capacity to handle the variable length sequences and a variety of token types. Images are normally inputted into vision transformers or hybrid CNN-ViT encoders, and videos can be supported by adding temporal attention layers or tubelet embeddings. Unified-IO 2 uses autoregressive transformers in its Unified-IO, which gives generative interface of Unified modeling and uses outputs as token sequences which the model produces step-by-step. Generalist models that are based on diffusion make use of UNet-like backbones with cross-attention to text tokens or instruction. Certain hybrid frameworks even change the transformers into diffusion stages in order to bring reasoning even closer to pixel-level operations. In any architecture, the main principle is the same: the central backbone does the job of representing all the tasks, and specialization is obtained by means of prompts, conditioning vectors or lightweight adapters.

4.3. Task interfaces: prompts and instructions

Instructional-based interfaces have been significant in the generalist vision models. These models do not follow task-specific training cues only but interpret natural-language instructions which define what task is to be carried out. An example of Unified-IO 2 is prompting that caption this video or segment the object in this frame, which is simply added to the input sequence. Equally, the diffusion-based systems such as InstructDiffusion incorporate textual prompts to control the diffusion path to a particular task, like editing or segmentation [14]. The vision systems based on LLM also come in with this idea enhanced with further abilities to chain commands in complex ways, multi-step reasoning and even dynamic task inferences.

There are two important benefits to prompt-based interfaces: they are easier to use since they enable the user to formulate tasks in natural language, and they enable the model to be generalized to unseen tasks on inference by decoding new prompts. There are two common strategies to interface used by generalist models:

- A single generic generative head, which outputs token sequences corresponding to any modality.
- Multiple specialized heads appended to a shared backbone, used for tasks that require structured outputs such as dense segmentation maps.

While the first approach offers maximal flexibility and aligns more closely with true any-to-any modeling, hybrid architectures remain common for efficiency reasons, particularly in dense prediction scenarios.

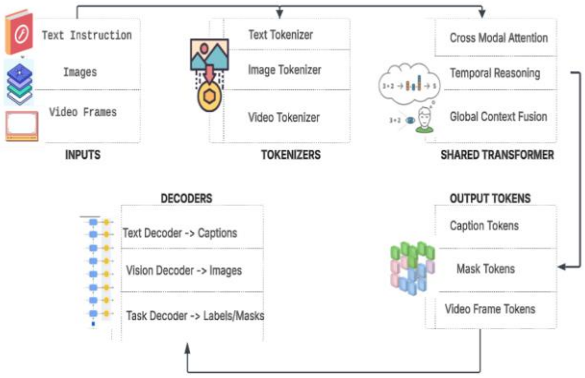

Figure 1. Conceptual Architecture of an Any-To-Any Vision Model

5. Training Paradigms for Generalist Image-To-Video Models

5.1. Large-scale pretraining corpora

Generalist models are data-hungry. They rely on:

- Web-scale image–text datasets for general visual understanding and grounding.
- Curated video datasets for temporal reasoning, video QA, and captioning.

- Synthetic data or task-specific datasets (e.g., segmentation, depth estimation, optical flow) for dense prediction tasks.

The main challenge is *balancing* these disparate sources, so that the model learns each task reasonably well without overfitting to any single distribution [15].

5.2. Multi-task learning and curriculum design

Any-to-any training is a form of large-scale multi-task learning:

- Tasks can be mixed uniformly, proportional to dataset size, or using more sophisticated sampling strategies.
- Some works use curricula, starting from easier tasks (e.g., image classification) and progressively adding harder ones (e.g., dense predictions, long video reasoning).
- 4M-21, for example, focuses on co-training across tens of vision tasks and modalities while controlling negative transfer through careful task weighting and architecture choices.

5.3. Instruction tuning and alignment

For models that interact through natural language, instruction tuning is critical:

- The model is fine-tuned on supervised pairs of (instruction, input, output), covering many tasks and modalities.
- This tuning helps the model interpret task descriptions, follow instructions, and generalize to novel combinations of tasks and inputs.

When combined with video data, instruction tuning can enable behaviors like [16]:

- “Describe what happens between 5 and 10 seconds in this clip.”
- “Identify anomalies in the surveillance video.”
- “Track the red car over time and output its trajectory.”

5.4. Efficient adaptation: LoRA, adapters, and task tokens

Given the cost of training such models from scratch, many works explore efficient adaptation:

- Low-Rank Adaptation (LoRA) and adapters allow finetuning on new tasks or domains with a small number of additional parameters [17].
- Task tokens are learned embeddings that specialize the shared backbone to particular tasks without changing most weights.

These techniques are especially useful for video, where domain shifts (e.g., from movies to medical videos) can be large.

6. Capabilities and Evaluation

6.1. Image and video understanding tasks

Generalist models are evaluated on a broad range of tasks:

- Image tasks: classification (ImageNet variants), detection, segmentation, keypoint estimation, referring expressions, image captioning, and VQA.
- Video tasks: action recognition, temporal localization, video captioning, and video QA.

PaLI-3, for example, achieves state-of-the-art or near state-of-the-art performance on MSR-VTT-QA and ActivityNet-QA while also performing well on image benchmarks. Unified-IO 2 and UnIVAL similarly show competitive performance across diverse benchmarks, demonstrating that generalist models can match specialized systems on many tasks.

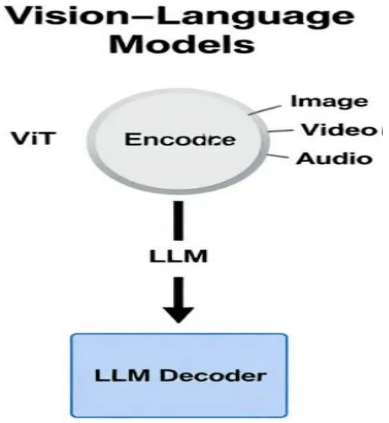

Figure 2 .Vision Language Models

6.2. Zero-shot and few-shot generalization

A central promise of generalist models is zero-shot or few-shot transfer:

- Because tasks share a backbone and representation, models can exploit common structure. For instance, training on image captioning and VQA can improve performance on video QA.
- DeepMind’s recent work on video models (e.g., Veo) suggests that video models trained at scale become strong zero-shot learners and reasoners, indicating a path toward unified video foundation models [18].

6.3. Temporal and compositional reasoning

Video understanding requires more than per-frame recognition:

- Temporal reasoning: tracking objects, understanding actions, and linking events across time.
- Compositional reasoning: answering multi-step questions, reasoning about causality (“why did the glass fall?”), or counting events.

Generalist models vary widely in how well they address these aspects. Some rely on relatively short clips and simple temporal pooling, while others introduce explicit temporal transformers. Benchmarks like ActivityNet-QA and NExT-QA are commonly used to measure these abilities.

6.4. Robustness, fairness, and biases

Training on large web-scale data brings known challenges:

- Biases in the data can lead to representational harms, especially when models are applied to surveillance, hiring, or safety-critical monitoring.
- Distribution shifts (e.g., from short curated clips to real-world continuous video streams) may cause failure modes that are not well captured by standard benchmarks.

Generalist models need systematic evaluation on robustness (adversarial perturbations, occlusions, lighting changes) and fairness (across demographic groups and contexts), especially when deployed at scale.

Table 2. Typical Benchmarks for Image-to-Video Understanding and Their Use in Generalist Models

Benchmark	Modality	Task Type	Used by (examples)
ImageNet, COCO	Image	Classification, detection, captioning	PaLI-3, Unified-IO 2, DICEPTION, 4M-21
NYUv2, ADE20K	Image	Depth estimation, segmentation	Unified-IO 2, 4M-21
MSR-VTT-QA	Video	Video question answering	PaLI-3, UnIVAL
ActivityNet-QA	Video	Video question answering	PaLI-3, UnIVAL
NExT-QA	Video	Video reasoning / QA	PaLI-3 and other VLMs
Kinetics, SSv2	Video	Action recognition	Unified-style video models (e.g., UnIVAL)

7. System-Level and Application Considerations

7.1. Deployment in real-time systems

Deploying any-to-any models for real-time image-to-video understanding raises practical issues:

- Latency: Video processing is computationally expensive. Even with efficient architectures, running a huge transformer on every frame may be infeasible.
- Memory and bandwidth: Long sequences of video tokens can exceed typical GPU memory limits, forcing temporal subsampling or frame selection.
- Edge vs. cloud: Some applications (e.g., mobile AR, robotics) may require on-device inference or tight latency bounds, pushing toward lighter generalist models or hierarchical designs (e.g., small local model + large cloud model).

Techniques such as model distillation, quantization, and caching of visual features across frames become important for practical deployments.

7.2. Safety, privacy, and responsible use

Any-to-any models can be deployed in sensitive contexts, such as surveillance, healthcare, or autonomous driving. This raises ethical concerns:

- Privacy: Video streams often contain identifiable individuals and sensitive contexts. Models should be combined with strong data governance and anonymization where possible.
- Misuse: Generalist models can be repurposed for harmful applications (e.g., unauthorized tracking, deepfake generation).
- Transparency: Users and stakeholders need clear documentation of model capabilities and limitations, including known failure modes.

7.3. Compute and energy considerations

Training and running generalist models has a non-trivial environmental and financial cost:

- Multi-modal pretraining on video is significantly more expensive than on images alone.
- Any-to-any architectures often require larger models to handle the increased variability in data and tasks.

Research into efficient generalist models through better architectures, tokenization, and training strategies is therefore not only scientifically interesting but also practically important [19].

8. Open Challenges and Future Directions

Despite impressive progress, current generalist vision models are still early steps toward fully flexible any-to-any image-to-video understanding. We highlight several open challenges.

8.1. Unified benchmarks for any-to-any evaluation

Most current evaluations treat tasks separately, even when the model itself is generalist. There is a need for:

- Benchmark suites that jointly cover images, videos, and multiple modalities, with unified metrics for both performance and efficiency.
- Scenario-driven evaluations where models must solve sequences of heterogeneous tasks (e.g., “caption this clip, then detect anomalies, then answer questions about a frame”), closer to how they would be used in real applications.

8.2. Long-range temporal modeling

Many systems still work with short clips (e.g., 8–32 frames), which limits their ability to reason about:

- Long-term dependencies (minutes or hours).
- Multi-episode narratives, such as surveillance over an entire day.

Efficient architectures for long-range video (sparse attention, memory mechanisms, hierarchical temporal modeling) remain an active research area.

8.3. Fine-grained control and editing

Diffusion-based generalist models show that multi-task editing is possible, but:

- Control over complex video edits (e.g., “replace the sky in all outdoor scenes with sunset” while preserving motion coherence) remains difficult.
- Integrating structured constraints (physical consistency, safety rules) into any-to-any models is still an open problem.

8.4. Embodied and interactive systems

Any-to-any models are natural candidates for embodied AI, where agents must perceive images and videos, understand language instructions, and output actions in the physical world or simulations.

Future research will likely explore:

- Combining any-to-any visual backbones with reinforcement learning or model-based control.
- Interactive learning, where the model can ask for clarifications or additional observations to resolve ambiguities.

8.5. Safety-aligned generalist models

Generalist models will increasingly be deployed in high-stakes environments. There is a growing demand for:

- Safety-aligned training objectives that penalize unsafe outputs and encourage conservative behavior in ambiguous situations.
- Explainability tools that help users understand why a model produced a particular decision from complex visual evidence.

These concerns overlap with broader efforts in responsible AI but take on new forms when dealing with continuous video and any-to-any outputs.

9. Conclusion

Any-to-any image-to-video understanding Generalist vision models are quickly developing as an ambitious concept into practical systems on research and industry. PaLi-3, UnIVAL, Unified-IO 2 and 4M-21 models have shown that one architecture can perform a very diverse set of visual tasks involving images and videos, and in many cases, it can perform well or even better than special purpose models on standard benchmarks. Nonetheless, the benefits of this development involve serious problems to deal with: the diversity of data sources, the lack of the ability to negatively affect the task, the ability to adapt to long video sequences, the use of powerful models in a safe and responsible manner. Even the methodological gap between the flexible any-to-any capabilities these models can provide, and the comparatively limited means by which we today test them, is ever-present. This research area is full of opportunities to researchers: new architectures, training regimes, benchmarks and areas of application. To practitioners, generalist models are offering a less challenging and more integrated way of developing real-world systems that need to be knowledgeable of both images and videos in dynamic settings. With the ecosystem coming of age, we will probably be able to observe a more intimate binding between generalist vision backbones and generalist language or action models where we are finally able to see multimodal agents that are capable of experiencing any-to-any interactions in a complex visual world.

References

[1] Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., ... & Kembhavi, A. (2024). Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 26439-26455).

[2] Wang, Z., Wang, J., & Jiang, C. (2022, October). Unified multimodal model with unlikelihood training for visual dialog. In Proceedings of the 30th ACM International Conference on Multimedia (pp. 4625-4634).

[3] Tang, Z., Yang, Z., Zhu, C., Zeng, M., & Bansal, M. (2023). Any-to-any generation via composable diffusion. Advances in Neural Information Processing Systems, 36, 16083-16099.

[4] Bai, Y., Zhou, Y., Zhou, J., Goh, R. S. M., Ting, D. S. W., & Liu, Y. (2024). From generalist to specialist: Adapting vision language models via task-specific visual instruction tuning. arXiv preprint arXiv:2410.06456.

[5] Wu, S., Fei, H., Qu, L., Ji, W., & Chua, T. S. (2024, July). Next-gpt: Any-to-any multimodal llm. In Forty-first International Conference on Machine Learning.

[6] Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., ... & Chandra, V. (2024). An introduction to vision-language modeling. arXiv preprint arXiv:2405.17247.

[7] Lu, J., Clark, C., Zellers, R., Mottaghi, R., & Kembhavi, A. (2022). Unified-io: A unified model for vision, language, and multi-modal tasks. arXiv preprint arXiv:2206.08916.

[8] Chen, X., Wang, X., Beyer, L., Kolesnikov, A., Wu, J., Voigtlaender, P., ... & Soricut, R. (2023). Pali-3 vision language models: Smaller, faster, stronger. arXiv preprint arXiv:2310.09199.

[9] Shukor, M., Dancette, C., Rame, A., & Cord, M. (2023). Unival: Unified model for image, video, audio and language tasks. arXiv preprint arXiv:2307.16184.

[10] Bachmann, R., Kar, O. F., Mizrahi, D., Garjani, A., Gao, M., Griffiths, D., ... & Zamir, A. (2024). 4m-21: An any-to-any vision model for tens of tasks and modalities. Advances in Neural Information Processing Systems, 37, 61872-61911.

[11] Fan, Y., Xian, Y., Zhai, X., Kolesnikov, A., Naeem, M. F., Schiele, B., & Tombari, F. (2024). Toward a diffusion-based generalist for dense vision tasks. arXiv preprint arXiv:2407.00503.

[12] Feng, K., Zhang, M., Li, H., Fan, K., Chen, S., Jiang, Y., ... & Yue, X. (2025). OneThinker: All-in-one Reasoning Model for Image and Video. arXiv preprint arXiv:2512.03043.

[13] Lu, J., Song, L., Xu, M., Ahn, B., Wang, Y., Chen, C., ... & Yang, Y. (2025). Atoken: A unified tokenizer for vision. arXiv preprint arXiv:2509.14476.

[14] Xu, X., Guo, J., Wang, Z., Huang, G., Essa, I., & Shi, H. (2024). Prompt-free diffusion: Taking" text" out of text-

to-image diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8682-8692).

[15] Qian, R., Ding, S., & Lin, D. (2024, September). Rethinking image-to-video adaptation: An object-centric perspective. In European Conference on Computer Vision (pp. 329-348). Cham: Springer Nature Switzerland.

[16] Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., ... & Wu, F. (2023). Instruction tuning for large language models: A survey. ACM Computing Surveys.

[17] Sun, Z., Yang, H., Liu, K., Yin, Z., Li, Z., & Xu, W. (2022). Recent advances in LoRa: A comprehensive survey. ACM Transactions on Sensor Networks, 18(4), 1-44.

[18] Rahman, S., Khan, S., & Porikli, F. (2018). A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning. IEEE Transactions on Image Processing, 27(11), 5652-5667.

[19] Wang, X., Chen, G., Qian, G., Gao, P., Wei, X. Y., Wang, Y., ... & Gao, W. (2023). Large-scale multi-modal pre-trained models: A comprehensive survey. Machine Intelligence Research, 20(4), 447-482.