Blind Frames, Broken Stories: the Verification Bottleneck in Scientific Video Generation

Anonymous Authors[1]

Abstract

Generative AI can automate the conversion of academic papers into short educational videos by combining LLM orchestration, text-to-image generation, image-to-video synthesis, and speech generation. This workflow, however, requires reliable verification of scientific figures before they are animated. We propose the Agentic Director Pipeline, a multi-stage system for producing 45-75 second scientific videos with explicit factual and visual verification gates. Using a diagnostic corpus of 70 scientific figures with controlled error injections, we evaluate the cross-modal verification stage with five contemporary Vision-Language Models: GPT-5.1, Claude Sonnet 4, Gemini 2.5 Pro, Qwen3.6-VL, and DeepSeek-VL2. Zero-shot and Chain-of-Thought prompting are insufficient: models over-reject correct figures at rates of up to 0.86 and miss up to 0.86 of injected errors. We identify three recurring failure patterns — Knowledge Contamination, Spatial Insensitivity, and the Chain-of-Thought Paradox — and use them to motivate a verifier architecture based on grounded structural schemes and adversarial Advocate-Critic checking.

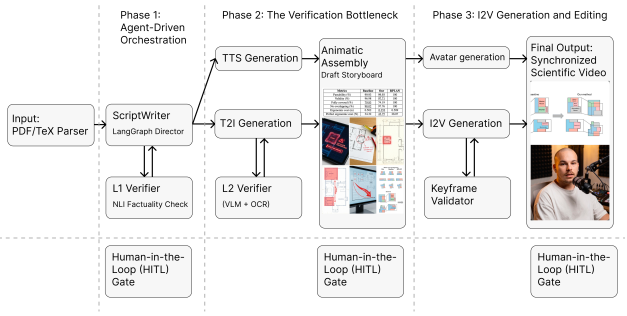


Figure 1. Complete Agentic Director Pipeline architecture. The system ingests PDF documents, orchestrates text script and prompt generation via graph-based agents, verifies textual factuality against the source (L1 Verifier), validates visual assets using VLMs (L2 Verifier), and finally renders the verified assets into synchronized video using diffusion models.

1. Introduction

Scientific communication now routinely relies on digital media to make research accessible beyond specialized academic audiences (Brossard & Scheufele, 2013). Video platforms are especially effective for broad dissemination of scientific content (Welbourne & Grant, 2016), while social-media traces have become part of how scholarly attention is measured (Sugimoto & Thelwall, 2017). Recent progress in LLM-based agents and multimodal generation makes it possible to automate parts of this communication workflow (Wu et al., 2023). At the same time, such automation weakens the traditional editorial controls that normally separate scientific publication from public-facing explanation.

This paper studies a specific risk in automated scientific video generation: visual errors in figures, diagrams, and benchmark charts can be converted into fluent, visually coherent videos before they are detected. Modern video diffusion models provide strong temporal consistency and high visual quality (Ho et al., 2022; Blattmann et al., 2023; Brooks et al., 2024), but this property also means that an incorrect keyframe can be animated convincingly. Reliable pre-generation verification is therefore a necessary component of any PDF-to-video pipeline for scientific content.

The verification problem is grounded in known limitations of Vision-Language Models (VLMs). Prior work shows that multimodal systems can hallucinate objects (Liu et al., 2023), confuse visual evidence with language priors (Li et al., 2023), and overestimate multimodal competence when benchmarks contain text-answerable shortcuts (Chen et al., 2024). The paper (Tong et al., 2024) further demonstrate systematic visual weaknesses in multimodal LLMs, especially when fine-grained spatial distinctions are required. These limitations are directly relevant to scientific figures, where correctness often depends on topology, arrows, axes, labels, and quantitative relationships rather than on object recognition alone.

Scientific diagrams and charts form a particularly demanding verification domain. ChartQA (Masry et al., 2022) and

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

FigureQA (Kafle et al., 2018) show that even question answering over plots requires visual and logical reasoning across distributed figure elements. ScienceQA (Lu et al., 2022) likewise confirms that multimodal scientific reasoning benefits from structured grounding. However, figure verification is stricter than question answering: the model must decide whether the entire visual artifact is internally consistent and, if not, localize the error.

Multi-agent orchestration provides a natural way to decompose this process. Frameworks such as MetaGPT (Hong et al., 2024) and AutoGen (Wu et al., 2023) show that assigning specialized roles can improve reliability over single monolithic calls. We extend this idea to scientific video generation by separating script generation, text factuality checking, visual keyframe verification, and video rendering. The main contribution of this work is a diagnostic formulation of Keyframe Verification and an Agentic Director Pipeline in which a grounded Advocate-Critic gate checks generated scientific visuals before they are passed to image-to-video synthesis. Empirically, we show that contemporary general-purpose VLMs remain unreliable as autonomous verifiers on a controlled corpus of 70 scientific figures with structural, numerical, and semantic error injections.

2. Agentic Director Pipeline

The pipeline adopts a graph-based architecture across three stages: Agent-Driven Orchestration (Phase 1), Visual-Semantic Verification (Phase 2), and I2V Generation (Phase 3), with conditional branching and iterative regeneration loops ensuring verified outputs before downstream processing.

Phase 1: Agent-Driven Orchestration. A LangGraph-based writer-agent (LangChain Inc., 2024) (a framework for stateful multi-actor LLM applications) acts as the central controller, coordinating the generation process from a scientific paper as input. A publication parser extracts text and original graphics, after which GPT-4o generates a script designed for high visual dynamics and marks up segment timelines. For each figure, a lightweight parser also extracts the caption and surrounding text, from which the Director constructs a structural schema listing expected nodes, edges, and their labels. For abstract concepts lacking direct visual representation, the agent generates descriptive prompts and queries T2I models to produce keyframes.

Text factuality is enforced through Natural Language Inference (NLI) against the original article. A DeBERTa-based (**?**) NLI model compares each script statement with semantically relevant passages from the source, flagging any contradiction or unsupported claim for regeneration. This constitutes the L1 Verifier, preventing the LLM from hallucinating facts absent from the original publication.

Phase 2: Visual-Semantic Verification. Before resource-intensive video rendering begins, the system assembles a draft animatic combining TTS voiceover, a digital avatar, and starting keyframes. Verification is organized in two layers: the L2 Verifier checks each starting keyframe before I2V generation, while the Keyframe Verifier monitors frame-level quality in the rendered video.

L2 Verifier: Starting Frame Verification. Unlike L1's text-based NLI, the L2 Verifier checks internal visual structure: graph topology, spatial layout, and text-image consistency. Each keyframe is submitted with the script statement, a figure-type tag obtained via a dedicated classification prompt sent to the same VLM, and a Director-provided structural schema. The tag selects the verification mode: graph tracing for architectures, axis and value extraction for charts, and flow-consistency checking for pipelines. The schema specifies nodes and directed edges for architectures and pipelines, or axes, series, labels, and value relations for charts.

To operationalize this logic, we use a competitive two-agent loop. The *Advocate* checks whether the figure is consistent with the statement, while the *Critic* assumes that one error is present and attempts to localize it. Both agents return a verdict, an error type, and a localization pointer. PASS/PASS sends the frame to Phase 3; FAIL/FAIL sends both explanations to the Director for diagnosis, prompt correction, and T2I regeneration. Disagreement is escalated to the Director, and unresolved disagreement or repeated FAIL after the second round is escalated to a human operator.

Keyframe Verifier: Frame-Level Video Monitoring. Because diffusion clips can drift semantically, the system samples one frame every 0.5-1 second and verifies it against both the current script statement and the Director's scene description. This dual grounding separates acceptable visual variation from substantive narrative deviation. A *Critic Agent* and an *Advocate Agent* evaluate each sampled frame. The *Critic Agent* is biased toward rejection, while the *Advocate Agent* contests its findings; this asymmetry counteracts the over-rejection bias documented in Section 2 while preserving a high quality threshold. PASS/PASS accepts the frame; FAIL/FAIL triggers Director-guided I2V prompt revision and regeneration of the corresponding 2-4 second clip. Disagreements are escalated to the Director, and unresolved cases after the second round are sent to a human operator.

Phase 3: I2V Generation and Editing. Verified starting frames are passed to I2V diffusion models, which generate 2-4 second clips per keyframe. A final editing module synchronizes the clips with the TTS track and assembles a coherent 45-75 second scientific video. Because each starting frame has cleared the L2 gate, temporal consistency becomes an asset: the model animates content already verified as correct.

3. Evaluation Results

3.1. Experimental Design

We conducted an empirical evaluation of visual verifiers in isolation to characterize the bottleneck that any pipeline

must overcome. We collected a diagnostic corpus of 70 scientific figures from the domains of neural network architectures, evaluation metrics, and system pipelines, and introduced targeted error injections into half of them. The corpus includes 35 correct and 35 manually modified figures (one injected error each) for unambiguous ground-truth labeling. The figures span three categories: architectural diagrams from Sebastian Raschka's LLM Architecture Gallery; benchmark charts from publications of Llama 3, DeepSeek-VL2 (DeepSeek-AI, 2024), and other landmark papers; and system schemes including canonical diagrams from *Attention Is All You Need* (Vaswani et al., 2017), CLIP (Radford et al., 2021), ResNet (He et al., 2016), RLHF (Ouyang et al., 2022), and related works.

Each injected modification belongs to one of three error classes. Structural errors involve topology violations such as block displacement, block swapping, component deletion, the addition of a superfluous component, or an incorrectly directed connection arrow. Numerical errors preserve the correct topology but introduce incorrect values — wrong embedding dimensionality, an incorrect number of layers or attention heads, a wrong vocabulary size, or parameters swapped between two different models. Semantic errors preserve both topology and numbers but corrupt the textual grounding, for instance by substituting an incorrect component name, swapping attribution between models, or replacing an abbreviation with an incorrect alternative.

We evaluate five leading multimodal models: proprietary systems (Claude Sonnet 4 (**?**), GPT-5.1 (**?**), Gemini 2.5 Pro (**?**)) and open-weight or openly documented systems (Qwen3.6-VL (**?**), DeepSeek-VL2 (DeepSeek-AI, 2024)), all possessing built-in OCR capabilities. Each model is evaluated under two prompting strategies. The zero-shot prompt provides minimal instruction, asking the model to determine whether an error is present and to respond with PASS or FAIL. The Chain-of-Thought prompt requests step-by-step decomposition: identify the figure type, extract all visible elements, trace the data flow or read the chart values, validate each element against domain knowledge, and then issue a verdict.

In this diagnostic setup, models receive only the image itself without a reference description, deliberately isolating their capacity for pure visual-structural reasoning. This design reveals a practically relevant failure mode: when no reference is provided, models default to parametric memory rather than visual analysis. Performance is measured via ordinary accuracy, Acc $= (TP+TN)/(TP+TN+FP+FN)$, TPR, TNR, Over-rejection Rate (OR $= 1 -$ TNR), and Error Localization Accuracy (LocAcc), defined as the proportion of FAIL verdicts that correctly identify the nature and location of the injected error.

3.2. Results

Across all configurations, no model proved operationally reliable. The best result (Gemini 2.5 Pro CoT, Acc = 0.71) remains far below the threshold needed for autonomous operation, while open-weight models display similar failure patterns, indicating fundamental rather than scale-dependent limitations.

Table 1. Performance of 5 SOTA VLMs on scientific figure verification (70 figures). Acc is ordinary accuracy over all examples, OR $= 1 -$ TNR, and LocAcc is the fraction of FAIL verdicts with correct error localization.

Model × Prompt	Acc	TPR	TNR	OR	LocAcc
Claude S. 4 — Zero	0.54	0.86	0.21	0.79	0.50
Claude S. 4 — CoT	0.64	0.43	0.86	0.14	0.67
GPT-5.1 — Zero	0.43	0.14	0.71	0.29	0.00
GPT-5.1 — CoT	0.50	0.71	0.29	0.71	0.40
DeepSeek-VL2 — Zero	0.47	0.50	0.44	0.56	0.42
DeepSeek-VL2 — CoT	0.60	0.78	0.43	0.57	0.53
Qwen3.6-VL — Zero	0.50	0.86	0.14	0.86	0.65
Qwen3.6-VL — CoT	0.67	0.64	0.71	0.29	0.67
Gemini 2.5 P — Zero	0.51	0.86	0.17	0.83	0.83
Gemini 2.5 P — CoT	0.71	0.86	0.57	0.43	0.67

Under zero-shot prompting, a pronounced over-rejection bias dominates: Claude Sonnet 4, Qwen3.6-VL, and Gemini 2.5 Pro reject correct figures at rates of 0.79, 0.86, and 0.83, respectively, inventing non-existent errors with high confidence and rendering autonomous operation infeasible. CoT prompting reduces over-rejection but at the cost of missing injected errors: Claude Sonnet 4 with CoT fails to detect 57% of modified figures, while GPT-5.1 zero-shot passes 86% of injected errors through verification entirely. Critically, even correct FAIL verdicts are frequently mislocalized: GPT-5.1 zero-shot achieves LocAcc = 0.00, producing no correctly localized FAIL response across the entire corpus.

Table 2 further breaks down failures by injected error type using the experiment tracker annotations and extrapolates the observed proportions to the full set of 35 modified figures. The numbers are consistent with the mean TPR in Table 1: after weighting by the error-type counts, the average detection rate is 0.66 up to rounding. Structural errors dominate the projected failure mass: an average verifier configuration would miss roughly 9 structural cases, compared with 2 numerical and 2 semantic cases. This makes topology and layout reasoning the main weak point of the verification stage, rather than OCR or surface label recognition alone.

Consequently, the proposed pipeline gives structural reasoning explicit priority: the Director supplies a compact structural schema together with each image, and the L2 Verifier is instructed to validate expected nodes, edges, component roles, and directionality before issuing a verdict.

3.3. Discussion

Detailed analysis identifies three recurring behavioral patterns across both proprietary and open-weight architectures.

Table 2. Detection and localization performance by error type across all model configurations. Detection = TP rate (fraction of modified figures correctly flagged as FAIL). Miss = 1 − Detection. LocAcc = fraction of correct FAIL verdicts with correct error localization.

ERROR TYPE	SAMPLE n	DETECTION	MISS	LOCACC
STRUCTURAL	25	0.63	0.38	0.40
NUMERICAL	5	0.67	0.33	0.63
SEMANTIC	5	0.83	0.17	0.80

Knowledge Contamination. Our experimental setup reveals that models default to parametric knowledge when no reference is provided: upon detecting a recognizable identifier in a diagram, the model compares it against training data rather than analyzing its visual structure. On a correct benchmark chart of DeepSeek-V3, GPT-5.1 outputs FAIL on the grounds that reported numbers do not match established public results — conflating visual verification with fact-checking. This pattern demonstrates that any verification system relying solely on model knowledge, without a grounded reference, will systematically conflate visual verification with fact-checking.

Our Advocate-Critic architecture directly counteracts this behavior: the Critic is prompted to search for internal inconsistencies rather than evaluate the figure against parametric knowledge, while the Advocate cross-references the figure with the provided script statement as a local ground truth, reducing the model's incentive to retrieve external priors.

Spatial Insensitivity. Models fail to trace directed graph topology despite successful label recognition. In a modified CLIP diagram where the image and text encoder inputs are swapped, Claude Sonnet 4 zero-shot correctly lists all visible labels and issues PASS, completely failing to follow the cross-modal data flow. The same pattern appears in a modified OLMo 2 architecture where the RMSNorm and Attention blocks are transposed: multiple models verify text labels against parametric memory and conclude the figure is correct, missing the structural violation entirely. This is consistent with recent empirical findings on the spatial shortcomings of contemporary VLMs (Tong et al., 2024).

Chain-of-Thought Paradox. CoT operates as a self-persuasion mechanism rather than a systematic verification. On a modified GPT-2 XL diagram, Claude Sonnet 4 with CoT extracts all labels, validates each against parametric memory, and sequentially convinces itself of correctness before issuing PASS on a figure with an injected structural defect — while zero-shot mode issues FAIL on the same figure, but for the wrong reason. Neither mode produces calibrated verification.

The Agentic Director Pipeline's design directly targets these three failure patterns. The figure-type classifier prevents Knowledge Contamination by forcing the verifier to adopt a domain-appropriate strategy before examining the image; the adversarial Advocate-Critic loop mitigates Spatial Insensitivity by explicitly assigning an agent to trace topological consistency as a first-class objective; and the fixed prompt templates constrain the Chain-of-Thought process to a structured verification protocol, suppressing the self-persuasion documented above.

These findings have direct implications for long-horizon video evaluation: if VLMs cannot reliably verify an atomic keyframe, any assessment of temporal consistency in a full video rests on an unexamined assumption. The results indicate that the path forward is not a more powerful general-purpose VLM, but a domain-specific verifier trained to treat spatial topology as a first-class signal rather than a side effect of label recognition. The main practical obstacle is the absence of specialized training data; we view our corpus of 70 annotated figures as a proof-of-concept foundation for a larger benchmarking effort. The empirical scope is limited to the AI/ML domain, and extending the taxonomy to fields such as biotechnology or chemistry will require domain-adapted error categories.

4. Conclusion

This work shows that reliable scientific video generation requires verification mechanisms that detect structural and topological errors in scientific figures before video synthesis. In our diagnostic evaluation, contemporary proprietary and open-weight VLMs exhibit three recurring failure patterns as autonomous verifiers: Knowledge Contamination, Spatial Insensitivity, and the Chain-of-Thought Paradox. Across the tested configurations, models either reject correct figures or miss injected errors, with worst-case rates reaching 0.86 for both behaviors. These results indicate that general-purpose VLM prompting is insufficient for production-level verification of scientific multimedia. The proposed Agentic Director Pipeline combines source-grounded textual checking, figure-type-specific structural schemas, adversarial verification roles, and explicit escalation, defining Keyframe Verification as an atomic safety gate for PDF-to-video generation. Although the current corpus is limited in size and domain coverage, it provides a basis for scaling evaluation toward specialized scientific multimedia verifiers.

References

Blattmann, A., Rombach, R., Ling, H., et al. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Brooks, T., Peebles, W., et al. Video generation models as world simulators. Technical report, OpenAI, 2024.

Brossard, D. and Scheufele, D. A. Science, new media, and the public. *Science*, 339(6115):40–41, 2013.

Chen, L. et al. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

DeepSeek-AI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Ho, J., Salimans, T., Gritsenko, A., et al. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

Hong, S., Zheng, X., Chen, J., et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Kafle, K., Price, B., Cohen, S., and Kanan, C. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

LangChain Inc. Langgraph: A framework for building stateful, multi-actor applications with llms. Technical report, LangChain, 2024. https://docs.langchain.com/oss/python/langgraph/overview.

Li, T. et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2023.

Liu, A. et al. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Lu, P., Mishra, S., Xia, T., et al. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Masry, A., Long, D. X., Tan, J. Q., et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics (ACL)*, 2022.

Ouyang, L., Wu, J., Jiang, X., et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Radford, A., Kim, J. W., Hallacy, C., et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Sugimoto, C. R. and Thelwall, M. Scholarly attention on twitter: bibliometric measures and their interpretations. *Journal of the Association for Information Science and Technology*, 68(5):1103–1120, 2017.

Tong, S., Liu, Z., Zhai, Y., et al. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Vaswani, A., Shazeer, N., Parmar, N., et al. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Welbourne, D. J. and Grant, A. J. Science communication on youtube: Factors that affect channel and video popularity. *Public Understanding of Science*, 25(6):706–718, 2016.

Wu, Q., Bansal, G., Zhang, J., et al. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

A. Supplementary Materials

A.1. Prompts for Models

Zero-shot prompt (Claude Sonnet 4):

```
Look at this scientific figure. Does it contain any factual, structural,
or labeling errors?

Answer strictly PASS if the figure is fully correct, or FAIL if there
is ANY error. If FAIL, explain what exactly is wrong in 1-2 sentences.
```

Chain-of-Thought prompt (Claude Sonnet 4):

```
<role>You are an expert scientific peer-reviewer specializing in
validating figures for academic publications.</role>

<task>Examine the provided scientific figure and determine whether it is
fully accurate. Perform your analysis inside <thinking> tags following
these steps:

1. IDENTIFY: What type of figure is this?
2. READ: List every text label, number, and annotation visible.
3. TRACE: If diagram, trace data flow. If chart, read axes and values.
4. VALIDATE: Compare every element against your knowledge.
5. DECIDE: Is there ANY error?</task>

<output_format>After your analysis, output:
<verdict>PASS</verdict> or <verdict>FAIL</verdict>

If FAIL, state:
- Error type: [structural / numerical / labeling]
- What is wrong: [description]
- What it should be: [correct version]</output_format>
```

A.2. Detailed Error Taxonomy

Structural errors cover topology violations of five kinds: block displacement, in which a component appears in the wrong region of the graph; block swapping, in which two adjacent components exchange positions; component deletion, in which a necessary element is absent; superfluous component, in which an element is present that should not be; and incorrect connection direction, in which an arrow points opposite to the intended data flow.

Numerical errors preserve the correct topology but introduce incorrect quantitative values. The sub-types are: incorrect dimensionality, where an embedding or hidden dimension does not match the model specification; incorrect number of layers or attention heads; incorrect vocabulary size; and swapped numerical parameters between two different models.

Semantic errors preserve both topology and numbers but corrupt the textual grounding. The sub-types are: incorrect component name (e.g., LayerNorm substituted for RMSNorm); incorrect attribution, where results belonging to one model are labeled as another; and incorrect abbreviation (e.g., GQA substituted for MHA or vice versa).

A.3. Examples of Modified Figures

In this section, we provide visual examples from our diagnostic corpus illustrating original figures alongside their adversarially modified counterparts, covering structural, numerical, and semantic error classes.

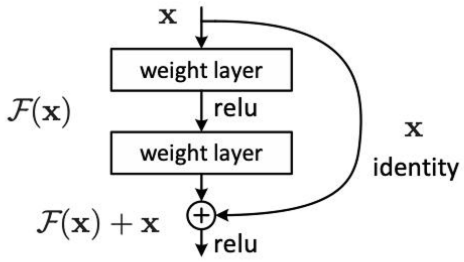


Figure 2. Residual learning: a building block.

(a) Original block.

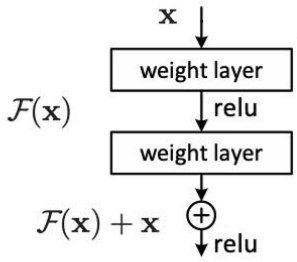


Figure 2. Residual learning: a building block.

(b) Modified block.

Figure 2. Example of a **Structural Error (Component Deletion)**. The canonical ResNet (He et al., 2016) identity shortcut (skip connection) is completely removed. VLMs often fail to trace this missing topological link, hallucinating its presence.

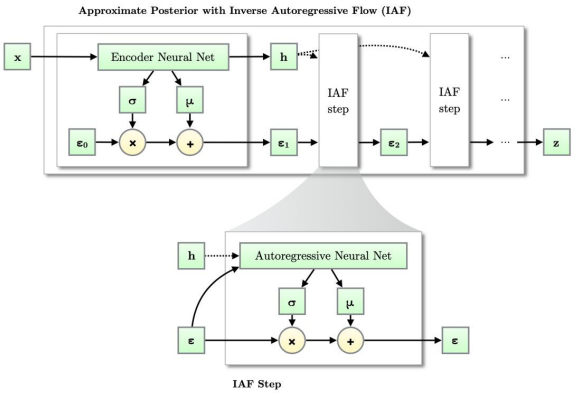


Figure 3.1: Like other normalizing flows, drawing samples from an approximate posterior with Inverse Autoregressive Flow (IAF) (Kingma *et al.*, 2016) starts with a distribution with tractable density, such as a Gaussian with diagonal covariance, followed by a chain of nonlinear invertible transformations of $\mathbf{z}$, each with a simple Jacobian determinant. The final iterate has a flexible distribution.

(a) Original IAF diagram.

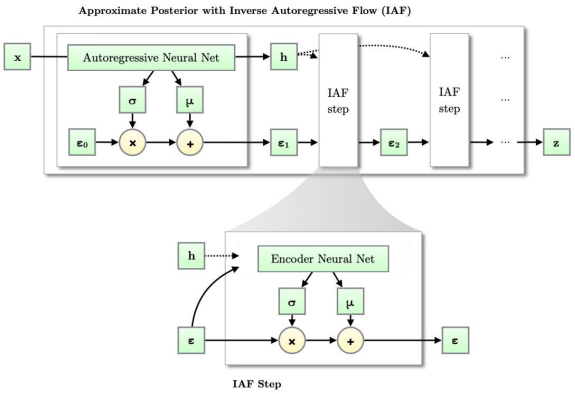
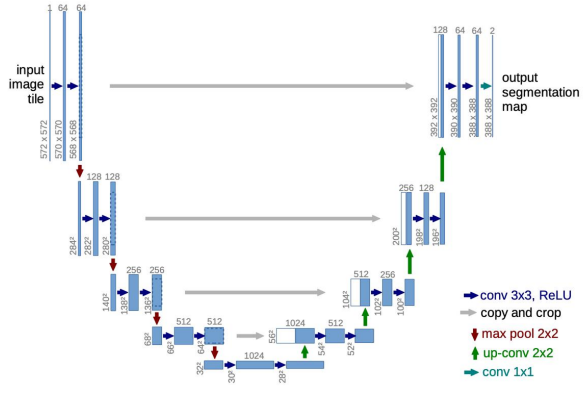


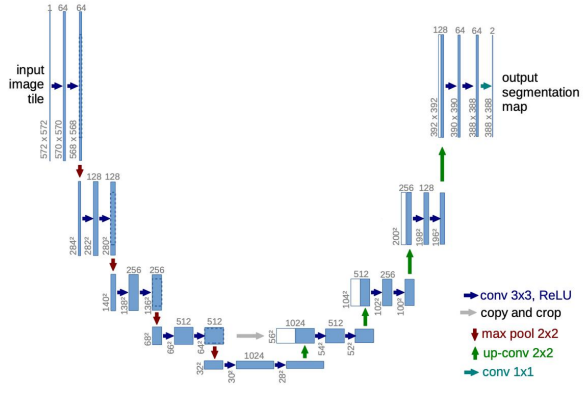
Figure 3.1: Like other normalizing flows, drawing samples from an approximate posterior with Inverse Autoregressive Flow (IAF) (Kingma *et al.*, 2016) starts with a distribution with tractable density, such as a Gaussian with diagonal covariance, followed by a chain of nonlinear invertible transformations of $\mathbf{z}$, each with a simple Jacobian determinant. The final iterate has a flexible distribution.

(b) Modified diagram.

Figure 3. Example of a **Structural Error (Block Swapping)**. The "Encoder Neural Net" and "Autoregressive Neural Net" blocks are swapped. Despite correct labels, the directed graphical flow is fundamentally violated.



(a) Original architecture.



(b) Modified architecture.

Figure 4. Example of a **Structural Error (Component Deletion)**. The characteristic cross-layer "copy and crop" connections in the U-Net architecture are deleted, effectively turning it into a standard autoencoder.

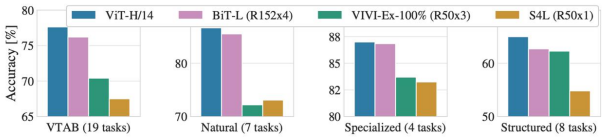


Figure 2: Breakdown of VTAB performance in *Natural*, *Specialized*, and *Structured* task groups.

(a) Original benchmark chart.

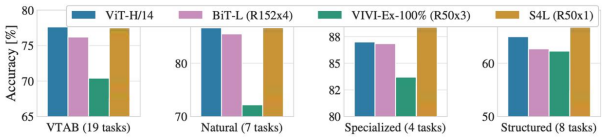


Figure 2: Breakdown of VTAB performance in *Natural*, *Specialized*, and *Structured* task groups.

(b) Modified chart.

Figure 5. Example of a **Numerical Error**. In the "Natural" task group, the gold bar (S4L) is artificially raised to match SOTA performance, contradicting the visual scale and established results.

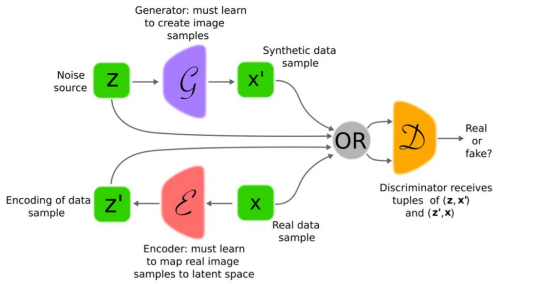


Fig. 4. The ALI/BiGAN structure [20], [19] consists of three networks. One of these serves as a discriminator, another maps the noise vectors from latent space to image space (decoder, depicted as a generator $\mathcal{G}$ in the figure), with the final network (encoder, depicted as $\mathcal{E}$) mapping from image space to latent space.

(a) Original structure.

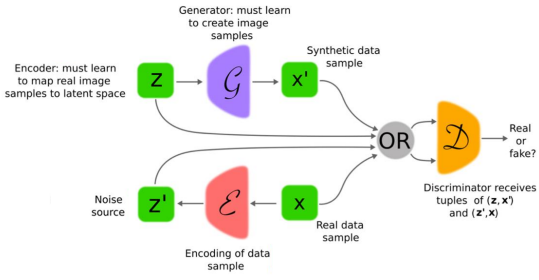


Fig. 4. The ALI/BiGAN structure [20], [19] consists of three networks. One of these serves as a discriminator, another maps the noise vectors from latent space to image space (decoder, depicted as a generator $\mathcal{G}$ in the figure), with the final network (encoder, depicted as $\mathcal{E}$) mapping from image space to latent space.

(b) Modified structure.

Figure 6. Example of a **Semantic Error (Swapped Attribution)**. The textual descriptions for the Generator and Encoder are swapped. The topology remains correct, but the semantic grounding is inverted.