Multimodal AI
Governance Framework

Type: Framework

Target Audience: AI Governance Teams, Product Managers, Risk Officers

This framework addresses the unique risks of systems that process and generate multiple modalities (text, images, audio, video) through targeted policies and technical controls.

1. Multimodal AI Overview

Multimodal AI systems combine multiple input/output types, creating compound risks not present in single-modality systems. Each modality brings unique governance challenges.

Modality	Examples	Key Risk Categories
Text	LLMs, chatbots, translation	Hallucination, bias, misinformation
Image	Image generation, classification	Deepfakes, copyright, harmful content
Audio	Speech synthesis, transcription	Voice cloning, impersonation
Video	Video generation, analysis	Deepfakes, surveillance, consent
Code	Code generation, completion	Vulnerabilities, IP, supply chain

2. Image & Video Governance Controls

Visual content generation requires specific safeguards against misuse.

Content Generation Policies

☐ Prohibit generation of real individuals without consent
☐ Block generation of minors in any potentially harmful context
☐ Implement NSFW/harmful content filters
☐ Maintain prompt blocklists for harmful generation requests

Authenticity & Provenance

☐ Apply invisible watermarks to all AI-generated images/video
☐ Embed C2PA metadata for content authenticity
☐ Maintain generation logs for audit and takedown requests
☐ Implement deepfake detection for uploaded content

3. Audio & Voice Governance Controls

Voice synthesis creates impersonation and fraud risks.

□ Require explicit consent for voice cloning of real individuals
□ Prohibit voice synthesis for fraud, impersonation, or deception
□ Apply audio watermarking to synthetic speech
□ Implement liveness detection for voice authentication systems
□ Maintain voice model registry with consent documentation

4. Cross-Modal Risk Assessment

When modalities combine, risks compound. Assess interactions between modalities.

4. Cross-Modal Risk Assessment

When modalities combine, risks compound. Assess interactions between modalities.

Combination	Compound Risk	Mitigation
Text + Image	Generated misinformation with fake "evidence"	Provenance tracking; fact-check integration
Text + Audio	Fake audio statements attributed to real people	Voice consent registry; watermarking
Image + Video	Realistic deepfakes from single photo	Identity verification; detection tools
Text + Code	Plausible-looking vulnerable code	Security scanning; human review
All Modalities	Fully synthetic media indistinguishable from real	Content authenticity infrastructure

5. Technical Control Requirements

Input Controls

- □ Implement content moderation on all user-uploaded media
- □ Scan uploads for CSAM and illegal content
- □ Validate file types and reject malformed inputs
- □ Rate-limit generation requests to prevent abuse

Output Controls

- □ Apply content safety classifiers to all generated output
- □ Block output matching known harmful content signatures
- □ Log all generations with user attribution for audit
- □ Implement automated takedown for policy violations

6. Policy Requirements

- □ Define acceptable use policy for each modality
- □ Establish consent requirements for likeness/voice use
- □ Create incident response procedures for synthetic media misuse
- □ Define retention and deletion policies for generated content
- □ Establish copyright/IP guidelines for AI-generated content
- □ Document human oversight requirements by modality

7. Regulatory Compliance Mapping

Regulation	Multimodal Relevance	Key Requirement
EU AI Act	Deepfakes, biometrics	Transparency obligations; prohibited uses
GDPR	Biometric processing	Consent for facial/voice data
US State Laws	Deepfake disclosure	Labeling requirements (CA, TX)
Copyright Law	Training data, outputs	Fair use assessment; licensing
CSAM Laws	Image generation	Detection and reporting obligations