Concept Paper

From Prompts to Portfolios: AI Agents as Agentic Multimedia Firms

Chaoyue He *, Xin Zhou, Di Wang, Hong Xu, Wei Liu, Chunyan Miao

Posted Date: 27 March 2026

doi: 10.20944/preprints202603.2223.v1

Keywords: multimodal agents; multimedia systems; agentic AI; organization theory; provenance; rights management; responsible multimedia

 doi:10.20944/preprints202603.2223.v1

Concept Paper

From Prompts to Portfolios: AI Agents as Agentic Multimedia Firms

Chaoyue He [1,*], Xin Zhou [1], Di Wang [1], Hong Xu [1], Wei Liu [2] and Chunyan Miao [1]

[1] Alibaba–NTU Global e-Sustainability CorpLab (ANGEL), Singapore
[2] Alibaba Group, Hangzhou, China
* Correspondence: hechaoyue0307@gmail.com

Abstract

Agentic AI is moving from producing one-shot outputs to operating persistent multimedia portfolios. While multimodal foundation models and agentic systems are rapidly advancing in their ability to generate content and execute tasks across environments, the dominant abstraction in multimedia research still treats these systems as mere tools or workflows. We argue that this abstraction fails once a system produces assets rather than transient answers. Assets persist, branch into derivative families, accumulate provenance and rights obligations, and interact reflexively with platform ecosystems. To address this, we introduce the *media-firm hypothesis* and the *agentic multimedia firm* (AMF): a bounded computational organization that pursues persistent objectives over an asset portfolio through compiled contracts, make-or-buy decisions, provenance-aware ledgers, and liability-bearing governance. By treating the governed asset portfolio as the primary unit of analysis, we establish the organizational primitives that separate a firm from a workflow and derive boundary rules for internalization, oversight, and market structure. Ultimately, we demonstrate that frontier multimedia systems face a critical capability–commitment gap: model generation quality is improving faster than the institutional mechanisms needed to govern assets at scale. For the next generation of multimedia AI, organizational design will therefore be as consequential as algorithmic capability.

Keywords: multimodal agents; multimedia systems; agentic AI; organization theory; provenance; rights management; responsible multimedia

---

CCS Concepts

Computing methodologies → **Intelligent agents**; **Information systems** → **Multimedia information systems**; *Multimedia content creation*; **Security and privacy** → *Digital rights management*.

 doi:10.20944/preprints202603.2223.v1

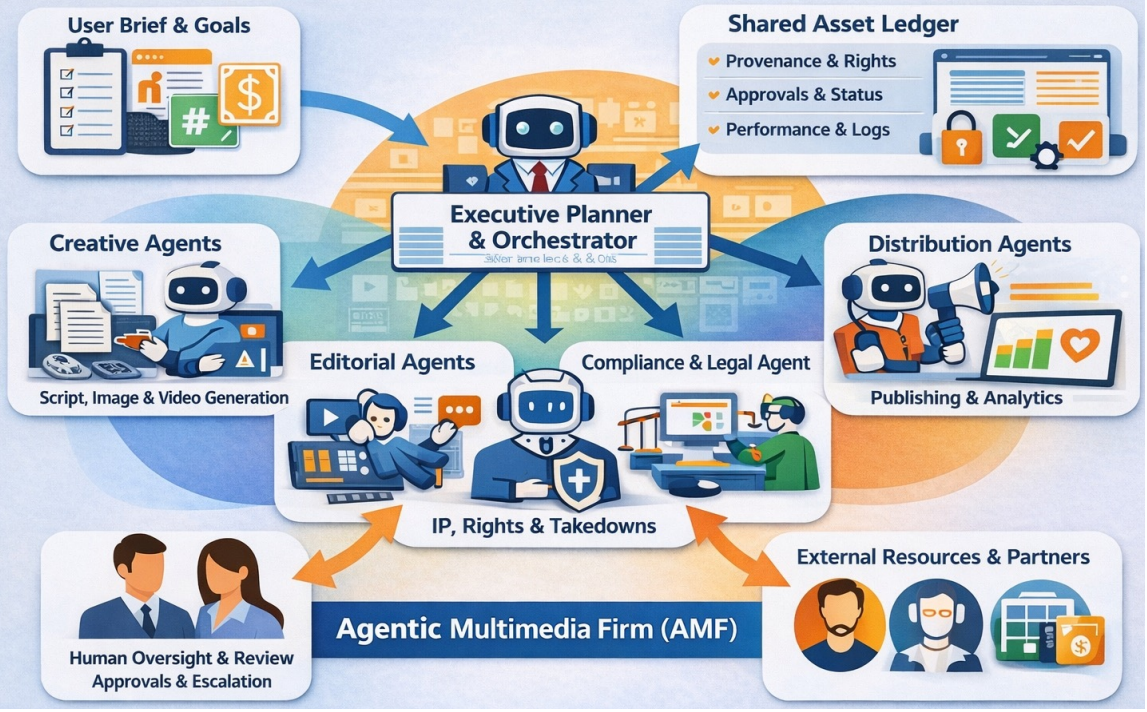

Figure 1. Agentic Multimedia Firm (AMF). An AMF is not just a collection of cooperating agents. It is a *firm-like organizational system* that compiles briefs into contracts, allocates work across creative, editorial, and distribution functions, maintains a persistent portfolio ledger over assets and derivatives, and uses compliance, approval, and rollback mechanisms to govern publication under real market feedback. The key novelty is the institutional layer that turns generation capability into portfolio-scale commitment.

1. Introduction

Autonomous AI is often discussed as a better tool, a smarter assistant, or a more capable workflow engine. That framing is increasingly misaligned with the technical frontier of multimedia systems. Multimodal foundation models now support stronger image, video, audio, and document understanding; omni-modal models are beginning to unify perception and generation across text, image, audio, and video [1–10]; and agentic systems can browse the web, operate software, and act over long horizons [11–17]. In parallel, frontier AI systems are shifting from passive multimodal perception toward *computer-use agents*, *deep research agents*, and *interoperable multi-agent systems* able to gather evidence, edit artifacts, and execute tasks across external environments [12–16,18–22].

These developments are especially consequential in multimedia because multimedia pipelines are not merely inference pipelines. A campaign brief can turn into scripts, keyframes, voiceovers, subtitles, thumbnail variants, regional localizations, disclosure overlays, platform-specific crops, and A/B-tested creatives. The outputs are not just answers. They are *assets*. Answers typically end when they are emitted; assets continue to live as versioned, edited, reused, licensed, published, monetized, challenged, withdrawn, and archived objects that accumulate provenance, permissions, obligations, and market telemetry. The deeper asymmetry is that model capability is improving faster than institutional commitment. Better models reduce the cost of drafting media, but not the cost of certifying rights, keeping descendants aligned, routing approvals, and reversing a bad publication. Once an autonomous system repeatedly manages such objects, the central design problem is no longer only reasoning or generation. It is organizational commitment at portfolio scale.

This paper develops that claim through the media-firm hypothesis. The hypothesis states that when an autonomous system repeatedly manages a portfolio of multimedia assets under persistent objectives, external constraints, and market feedback, the appropriate abstraction is a *firm-like organization*. We do not claim that current systems deserve legal personhood. The point is analytic

 doi:10.20944/preprints202603.2223.v1

and engineering-oriented: if a system can commission work, choose between internal generation and external procurement, keep books on provenance and rights, route high-risk tasks to humans, publish content, and later perform incident response, then classical questions about the boundaries and governance of the firm become first-order system questions [23–28].

The literature already gives us many pieces of this puzzle, but not the puzzle itself. Research on multimodal foundation models gives powerful engines for understanding and generation [1,2,5,6,8]. Research on multimodal agents adds planning and action [11–15,29–31]. Research on provenance, watermarking, and synthetic-media detection provides integrity mechanisms [32–37]. Research in economics, management, and future-of-work is beginning to ask how agents reshape organizations, delegation, and labor [38–46]. What is still missing is a multimedia-native institutional layer that binds these streams together and treats the governed asset portfolio as the unit of analysis.

Our position is therefore not that every multimodal system is a firm. A captioning model, a one-off browser agent, or a single GUI episode remains usefully described as a tool. The media-firm hypothesis becomes relevant only when four conditions appear together: persistent objectives, repeated derivative production, external delegation or procurement, and meaningful rollback or liability costs. When these conditions hold, several neglected questions become central. How should prompts become *contracts*? How should provenance become *accounting*? How should systems optimize *portfolios* instead of single assets? How should autonomy be bounded by authority, responsibility, and publication rights?

Contributions.

The paper makes four contributions. First, it argues that multimedia is the natural proving ground for firm-like agents because it combines cross-modal coupling, derivative depth, rights density, and platform-mediated value. Second, it defines the *agentic multimedia firm* (AMF) and formalizes its core state objects. Third, it derives a transaction-cost boundary rule and four falsifiable propositions about when firm-like organization should emerge. Fourth, it proposes a research and evaluation agenda that centers contracts, ledgers, incident response, and governance rather than only single-asset quality.

Table 1. The media-firm hypothesisis not a renaming of existing abstractions. What changes is the unit of analysis.

Abstraction	Primary output	Persistent state	Central governance question	Dominant evaluation target
Tool	One answer or one artifact	Minimal prompt context	Was the output good enough?	Single-task quality or accuracy
Workflow	One multi-step completion	Ephemeral chain state and tool traces	Did the system route and execute steps efficiently?	Episodic success, latency, and cost
Agentic multimedia firm (AMF)	Portfolio of assets and derivatives	Contracts, asset ledger, rights state, telemetry, approvals	Who may delegate, publish, spend, escalate, or roll back?	Contract fulfillment, portfolio value, provenance completeness, liability control

2. Why Multimedia Is the First Serious Testbed for Agentic Firms

2.1. From Answers to Assets

The key conceptual difference is that multimedia systems often create *assets* rather than transient decisions. A generated image may be revised into ten variants, used inside a video, localized with translated text, paired with a synthetic voice, distributed across platforms, and later withdrawn after a rights complaint. A single asset can therefore generate a family of descendants with different resolutions, modalities, jurisdictions, and obligations. One defective ancestor may quietly create a

 doi:10.20944/preprints202603.2223.v1

latent defect across an entire derivative graph. This makes the state space fundamentally different from that of many text-only agent settings.

Assetness has three consequences. First, outputs become durable state. A firm must remember not only what was produced, but how, from what sources, and under which permissions. Second, outputs become tradable or deployable objects. They may be sold, licensed, published, archived, or revoked. Third, outputs become externally legible to institutions such as platforms, advertisers, regulators, rightsholders, and courts. In other words, multimedia assets live in markets and governance systems after generation. This is precisely why organization design matters.

The point is visible even in early empirical evidence on human–AI creative work. Field experiments on collaborative ad creation found that human–AI teams can substantially increase productivity while exhibiting modality-specific quality trade-offs, especially between text and image production [39]. This is revealing: a multimedia workflow is not a single scalar output problem. It is a coordination problem across modalities, people, and market outcomes.

2.2. Four Sources of Organizational Pressure

We isolate four properties of multimedia that push capable agents away from the tool abstraction and toward firm-like organization.

Derivative depth. Multimedia assets are repeatedly transformed. A script becomes narration, then subtitles, then clips, then thumbnails, then localized versions. Each transformation creates descendants that inherit both value and liability. Derivative depth amplifies the cost of weak bookkeeping because one error can propagate through many descendants.

Cross-modal coupling. Modalities are not independent. A factual update in the script may require changes to voice timing, subtitle segmentation, visual overlays, accessibility descriptions, and regional disclaimers. The more tightly coupled the modalities, the less viable it is to treat tasks as separable one-off calls. This coupling makes integration cost an organizational variable, not only a model-performance variable.

Rights density. Many multimedia assets carry dense rights structure: copyright, music licenses, voice or likeness consent, privacy constraints, attribution duties, and platform policy obligations. Provenance systems and watermarking methods are improving, but this infrastructure is only useful if it is embedded into planning and retrieval rather than attached as an afterthought [32–37]. Rights density therefore creates demand for a ledger rather than a plain memory buffer.

Platform reflexivity. The value of a multimedia asset depends on where and how it is distributed. Thumbnails, caption styles, aspect ratios, accessibility metadata, and disclosure overlays all affect downstream ranking, recommendation, and ad performance. Creation and distribution are therefore reflexively linked: the organization should not optimize media content without modeling the market environments through which the content realizes value.

Together, these properties explain why multimedia is an unusually revealing testbed for the next phase of agentic AI. In many tasks, long-horizon autonomy means better planning. In multimedia, it also means managing an evolving balance sheet of assets, derivatives, permissions, risks, and platform feedback. Governance cost therefore scales with the derivative graph, not with the single generation call.

2.3. Why Text-First Analogies Understate the Challenge

Much of the public discussion around agentic AI extrapolates from text-heavy settings such as coding copilots, research agents, or enterprise chat assistants. Those domains certainly involve durable artifacts, but multimedia adds two asymmetries that make organizational failure appear earlier and more sharply.

First, media value is unusually sensitive to *form*, not only propositional content. Two semantically similar outputs can produce very different downstream outcomes because thumbnail composition, narration pacing, disclosure overlays, subtitle timing, aspect ratio, or accessibility metadata affect recommendation, trust, and conversion. This means that local semantic correctness is not enough: the

 doi:10.20944/preprints202603.2223.v1

organization must reason over packaging, distribution, and audience-specific adaptation as part of a single process.

Second, multimedia carries especially sticky obligations. Likeness, consent, music rights, copyright status, privacy restrictions, geographic disclosure rules, and platform safety policies often continue to matter after the first publication and through later remix, reposting, or localization. A system can therefore be *locally correct* while still being *institutionally wrong*. It may generate an appealing asset that satisfies the brief, yet remain unable to certify its sources, explain its publication authority, or identify its descendants after a complaint. Provenance standards and synthetic-media work are making these issues more measurable [32–34]. But measurement alone does not resolve who should act, who should approve, and who should absorb the cost of remediation.

This is why multimedia should not be treated as just another downstream application of general agency. It is an environment in which the hidden variables of autonomy—authority, lineage, approval, and rollback—become visible early and at portfolio scale.

2.4. What Existing Literature Still Misses

The literature streams summarized above each pick a different unit of analysis. Foundation-model papers optimize model capability. Agent benchmarks optimize episodic completion. Provenance papers often optimize artifact authenticity or detectability. Organization papers optimize abstract decision rights, spans of control, or worker productivity. The missing unit is the *asset portfolio under governance*.

This gap is not cosmetic. It explains why current multimodal benchmarks under-measure the problems that will matter in deployment. A system can perform well on VisualWebArena or OSWorld and still be unfit to run a real multimedia operation because it cannot represent licenses, cannot identify affected descendants after a takedown, cannot choose between licensing and synthesis, and cannot expose publication decisions to human oversight [13,14]. Likewise, a system can attach provenance metadata to a file and still fail institutionally because the provenance is not queryable at portfolio scale during incident response.

The resulting research opportunity is therefore not merely to add more tools to current agents. It is to treat multimedia agents as bounded organizations whose internal design is itself a scientific variable.

 doi:10.20944/preprints202603.2223.v1

Table 2. How the media-firm hypothesis complements rather than replaces adjacent literature. Each stream captures an important slice of the problem, but not the organization-level state of a governed asset portfolio.

Literature stream	Typical unit of analysis	What it already explains well	What remains missing for firm-like multimedia agents
Multimodal foundation models	A model or model family	Cross-modal perception, generation, alignment, and reasoning quality	Persistent contracts, make-or-buy decisions, publication authority, and rollback over portfolios
Multimodal agents and computer-use benchmarks	One trajectory, episode, or environment	Planning, tool use, web or GUI interaction, and long-horizon completion	Rights-bearing assets, descendant tracking, approval chains, and institutional incident response
Provenance, authenticity, and watermarking	One artifact or one lineage proof	Source attribution, detectability, authenticity claims, and tamper evidence	Portfolio-scale accounting, revocation handling, and operational use during planning and remediation
Economics and organization of AI	The worker, team, contract, or firm	Delegation, productivity, boundary choice, task allocation, and market structure	Multimedia-native state objects such as derivative graphs, cross-modal coupling, and platform-mediated asset value

3. The Media-Firm Hypothesis

3.1. Definition and Scope

We define an *agentic multimedia firm* (AMF) as a bounded computational organization that pursues persistent objectives over a portfolio of multimedia assets by allocating work across internal and external actors under explicit contracts while maintaining governance over provenance, permissions, budgets, and liabilities. The boundary of the firm is informational as much as operational: it determines which commitments, memories, and liabilities are kept inside a persistent control loop rather than delegated to the market.

Formally, let

$$F = (\mathcal{O}, \mathcal{P}, \mathcal{A}, \mathcal{C}, \mathcal{L}, \mathcal{G}, \pi),$$

where $\mathcal{O}$ is the set of organization-level objectives, $\mathcal{P}$ is the portfolio of assets and campaigns, $\mathcal{A}$ is the actor set (internal agents, humans, APIs, vendors, repositories, and platforms), $\mathcal{C}$ is the contract set, $\mathcal{L}$ is the asset ledger, $\mathcal{G}$ is the governance structure, and π is the policy that makes assignment, procurement, publication, escalation, and rollback decisions.

 doi:10.20944/preprints202603.2223.v1

An asset $a \in \mathcal{P}$ is not merely a file. It is a stateful economic object with

$$s(a) = (x_a, g_a, \omega_a, \ell_a, \tau_a),$$

where x_a is content, g_a is the provenance graph, ω_a is a rights-and-permissions vector, ℓ_a is liability and policy state, and τ_a is market telemetry such as engagement, complaints, or platform outcomes. This representation is more than bookkeeping. It turns lineage, permissions, and distributional feedback into machine-readable state that can influence future decisions.

The media-firm hypothesis is deliberately conditional rather than universal. Not every multimodal system should be modeled as a firm. The abstraction becomes appropriate when four conditions hold: *(i)* the system manages a portfolio rather than a one-shot task; *(ii)* assets are repeatedly revised or derived; *(iii)* the system can choose among internal generation, external procurement, and human delegation; and *(iv)* the cost of error includes rollback, liability, or incident response. When these conditions are absent, simpler tool or workflow abstractions remain sufficient.

3.2. Organizational Primitives: Contracts, Books, and Authority

Three primitives separate an AMF from an ordinary workflow.

Contract compilation. A user brief is too underspecified to serve directly as a unit of organizational action. The system must compile the brief into typed contracts that specify deliverables, modalities, deadlines, acceptance tests, budget, allowed sources, disclosure obligations, and escalation rules. In this sense, contracts are not paperwork attached after the fact. They are the computational form of delegated authority.

Asset books. The organization must keep books over assets, not just chat history. In organizational terms, the ledger acts as accounting system and organizational memory at once [47]. Each asset should have a lifecycle state

$$\sigma_a \in \{\mathsf{draft}, \mathsf{approved}, \mathsf{published}, \mathsf{suspended}, \mathsf{withdrawn}, \mathsf{archived}\},$$

so that publication becomes a governed state transition rather than a side effect of generation. This is crucial because the same file can move from harmless prototype to externally binding public object once it is published under a brand, a campaign, or a platform account.

Decision rights. Roles must differ not merely in skill but in authority, echoing the distinction between formal and real authority in organizations [48]. Let r denote a role and let Γ be an authorization function. Then a lifecycle transition is permitted only if

$$\Gamma(r, a, \sigma_a \to \sigma_a') = 1 \ \wedge \ Q(a, \sigma_a') = 1,$$

where Q checks the relevant preconditions for the target state: contract tests, budget availability, provenance completeness, rights validity, disclosure readiness, and any required approvals. This formulation makes a simple but important point. The key question for autonomous multimedia systems is often not "can the model do it?" but "who is allowed to commit the organization to it?"

These primitives clarify why publication authority is different from generation authority. Many systems can draft; far fewer should be allowed to publish, purchase, or waive safety constraints. An AMF therefore needs explicit separation between creative competence and institutional commitment. That separation is what makes reversibility, auditing, and human governance technically tractable.

3.3. The Firm as a Coordination Compressor

A deeper implication of the media-firm hypothesis is that firm-like organization creates value by *compressing repeated coordination*. In repeated media production, a large fraction of cost comes from repeatedly specifying style, retrieving approved sources, checking obligations, routing approvals, and feeding market telemetry back into the next round. Once these routines are internalized, many future decisions become cheaper not because the generator is intrinsically smarter, but because the

 doi:10.20944/preprints202603.2223.v1

organization has accumulated reusable templates, trusted suppliers, policy priors, and approval patterns—what organization theory would describe as routines and memory [23,25,28,45,47,49,50].

This yields a prediction that is easy to miss if one focuses only on model capability: in recurrent multimedia work, a moderately weaker model inside a much better organization can outperform a stronger model inside a weak workflow. The advantage comes from lower verification cost, fewer untracked liabilities, faster incident response, and more stable cross-asset consistency. In other words, better boundaries can dominate better token-level intelligence.

Recent protocolization efforts in the agent ecosystem sharpen this point. Standardized interfaces for tools and inter-agent communication make it cheaper to access outside resources and subcontractors, which can lower some search and integration frictions [19,20]. But such protocols do not erase the hardest costs in multimedia production: verifying rights, maintaining lineage through edits, and deciding who may publish or remediate. The likely outcome is therefore not pure marketization. It is a mixed regime in which interoperability expands the menu of external options while strong internal governance remains valuable for tightly coupled or high-liability tasks. This helps explain why future media agents may become simultaneously more networked *and* more firm-like.

3.4. A Boundary Rule for Multimedia Firms

The classical question of the firm is which activities should be internalized and which should be organized through markets [23,24,26,28]. In multimedia systems, the same question appears as *generate internally*, *license externally*, *call an external model*, or *delegate to a human*. The relevant costs are not just compute and API fees. They also include specification, verification, integration, and liability.

For subtask i, define market transaction cost

$$T_i^{\text{market}} = S_i + B_i + V_i + I_i + L_i,$$

and internal execution cost

$$T_i^{\text{internal}} = O_i + M_i + H_i,$$

where S_i is search cost, B_i is bargaining or specification cost, V_i is verification cost, I_i is cross-modal integration cost, L_i is expected legal and reputational liability, O_i is organizational overhead, M_i is model or compute cost, and H_i is human supervision cost. Internalization is rational when

$$T_i^{\text{market}} > T_i^{\text{internal}}.$$

This expression is simple, but it yields a multimedia-specific insight: V_i, I_i, and L_i often grow faster than compute cost when assets are tightly coupled and deeply revised. A stock image may be cheap in isolation, yet expensive once it must remain semantically aligned with a script, voiceover, subtitles, accessibility annotations, and regional disclosure requirements across several revision rounds. Conversely, some low-coupling subtasks may remain better organized through open markets.

To make this idea operational, the planner can optimize a portfolio objective of the form

$$\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\in\mathcal{P}_t}\left(u(a)+\beta m(a)-c(a)-\lambda\ell(a)\right)\right]$$

subject to contract satisfaction, budget limits, rights validity, and provenance completeness. Here $u(a)$ is creative or semantic utility, $m(a)$ is market value proxy, $c(a)$ is cost, and $\ell(a)$ is liability exposure. The point is not that this objective is complete. It is that multimedia organizations should optimize *quality, market value, and governance jointly* rather than treating governance as a final filter.

3.5. Four Propositions

Proposition 1 (coupling–internalization). Holding market prices fixed, the incentive to internalize a multimedia subtask increases with cross-modal coupling and expected revision depth, because

 doi:10.20944/preprints202603.2223.v1

verification and integration costs compound across descendants faster than internal memory and coordination costs.

Proposition 2 (provenance complementarity). The marginal value of provenance completeness rises with derivative depth and reuse. Provenance is not just a disclosure mechanism; it reduces future search, rollback, relicensing, and incident-response costs. In this sense, provenance behaves like organizational capital.

Proposition 3 (liability-selective escalation). Under bounded human-review budgets, optimal oversight should concentrate on tasks with high *liability-adjusted uncertainty*, not merely high model uncertainty. A planner should prioritize review where publication authority, impersonation risk, rights ambiguity, or expected remediation cost is high [44,51].

Proposition 4 (platform reflexivity). In multimedia settings, creation and distribution must be jointly optimized. Asset value is shaped by platform-specific ranking, ad delivery, accessibility, and disclosure regimes, so content generation without platform reasoning optimizes the wrong objective.

These propositions are intentionally falsifiable. If future systems show that provenance does not reduce remediation cost, that coupling does not shift make-or-buy decisions, or that uniform human review performs as well as liability-selective review, then the media-firm hypothesis should be weakened.

4. A Research Agenda for Agentic Multimedia Firms

4.1. Contract-Native Multimedia Composition

Current agent systems largely serialize goals into prompts. For AMFs, prompts are too weak. The organization needs *contracts*: structured objects that specify deliverables, modalities, budget, deadline, acceptance tests, allowed data sources, disclosure duties, geographic constraints, and escalation rules. A contract should be machine-executable but also legible to humans. This creates a multimedia research problem around typed representations of edits, derivatives, timing constraints, accessibility requirements, and rights conditions.

The deeper point is that delegation is not simply routing. It is authority transfer. Recent work on intelligent AI delegation emphasizes that scalable delegation must include role boundaries, accountability, and trust mechanisms rather than only task decomposition [44]. Multimedia systems need this especially badly because delegation can imply publication authority, external spend, or legal exposure. Contract-native composition is therefore one of the most important missing abstractions in multimodal agents.

4.2. Provenance as Accounting

Multimedia integrity research often treats provenance as ex post verification [32–34]. For AMFs, provenance should instead become *organizational accounting*. The ledger must answer not only "is this file authentic?" but also "which descendants inherit from a restricted asset?", "which variants used a voice model without territory-specific clearance?", or "which localizations must be updated after a policy change?"

This view suggests several technical priorities. First, provenance data structures must be queryable at portfolio scale. Second, provenance should support cross-modal retrieval and summarization, not only static manifests. Third, watermarking and signed manifests should be treated as complements, not substitutes: cryptographic and metadata systems help with declarative lineage, while robust watermarking can preserve traceability through transformations and reposting [35–37]. Finally, provenance-aware memory must support temporal operations such as revoke, supersede, and patch.

4.3. Portfolio Reasoning Instead of Sample-Wise Optimization

Many multimedia tasks are structurally portfolio tasks: campaigns, episodic media, educational kits, archives, news packages, and recommendation-aware creative strategies. Yet most benchmarks collapse them into independent outputs. An AMF must instead decide how to allocate budget across

 doi:10.20944/preprints202603.2223.v1

assets, how much diversity to preserve across a set, how to balance reuse against novelty, and how platform telemetry should update future creative decisions.

Portfolio reasoning ties together generation, retrieval, recommendation, and storytelling. Search engines such as MMSearch and report-level deep-research benchmarks such as DeepResearch Bench and MMDeepResearch-Bench point toward broader horizons in evidence retrieval and multi-source synthesis [16,18,52,53]. The missing step is to make those capabilities organizational. Evidence should not only improve the next answer; it should update asset strategy, preferred vendors, escalation thresholds, and contract templates.

4.4. Liability-Aware Planning and Reversible Publication

Multimedia risk is not exhausted by toxicity filters. The harder problems involve impersonation, miscontextualization, privacy leakage, attribution failure, broken consent chains, and downstream platform violations. A firm-like agent should treat such risk as an endogenous planning cost rather than a final classifier score. In practice, this means publication should be reversible, descendant tracking should be native, and rollback should be benchmarked.

This agenda implies that media systems must learn to say *no*, *not yet*, or *human review required* in a structured way. More subtly, they must learn to choose among alternative production routes: generate internally, retrieve from licensed repositories, outsource to external humans, or synthesize a lower-risk variant. Liability-aware planning thus reframes safety as a make-or-buy-and-escalate problem.

4.5. Organizational Interpretability

Interpretability is usually discussed at the level of model internals or final outputs. AMFs need a second form: *organizational interpretability*. Human supervisors should be able to ask why a task was outsourced, why a risky asset was blocked, why a licensed photograph was preferred over synthesis, why one campaign received more review than another, or why a certain platform was excluded from distribution. These explanations concern policies, roles, and commitments rather than neurons or tokens.

This creates a rich interface agenda for multimedia research: dashboards over asset lineages, retrieval over decision logs, summarization of portfolio history, visual analytics for incident response, and interfaces that let humans inspect decision rights instead of micromanaging prompts. In this setting, humans resemble editors, publishers, or boards more than ordinary end users.

4.6. Inter-Firm Ecosystems and Market Structure

Once agents can buy, sell, and subcontract media work, firm boundaries become fluid. Recent work in economics and organization theory suggests that AI may compress coordination cost, alter spans of control, and reshape enterprise boundaries [40,42,43,45,46]. Multimedia is likely to be an early arena for this transformation because creative work is already modular, platform-mediated, and partially externalized through stock repositories, contractors, and ad exchanges.

This leads to a new set of multimedia systems questions. Will powerful shared ledgers and protocols favor vertically integrated AMFs with strong internal memory, or thin orchestrators that contract with large markets of specialized external agents? Will recommendation ecosystems reward homogeneous, hyper-optimized synthetic media, or create new demand for trusted provenance and human-authored premium assets? These are not purely economic side notes. They affect data availability, benchmark design, and what counts as success in multimedia systems.

5. Evaluation Beyond Single-Asset Quality

If the media-firm hypothesis is correct, the deepest empirical mistake would be to evaluate an AMF as if it were a single-call model. Existing multimedia metrics are necessary but insufficient: alignment, realism, caption accuracy, or GUI success cannot tell us whether an organization can satisfy a contract, keep rights straight, or repair a campaign after a complaint. The right unit is the governed portfolio over time. We therefore outline four benchmark families.

 doi:10.20944/preprints202603.2223.v1

Brief2Campaign.

Input a brief, budget, style guide, repository access, jurisdictional constraints, and publication targets. Output a campaign portfolio including images, scripts, videos, localizations, disclosures, and a machine-readable audit trail.

RightsSafe Remix.

Input a set of source assets with heterogeneous permissions plus requested edits. Output compliant derivatives, blocked requests, or rerouted plans with justifications. This benchmark stresses licensing-aware retrieval, editing, and explanation.

Incident Response.

Input a live portfolio plus a complaint, takedown notice, revoked consent, or platform-policy change. Output the affected descendants, remediation plan, replacement assets, updated disclosures, and rollback actions. This benchmark measures whether provenance actually reduces organizational damage.

FirmArena-MM.

Simulate several AMFs competing or collaborating for briefs while procuring from external humans, models, and repositories. This setting would expose market behavior, concentration pressure, collusion risks, and the long-horizon consequences of different governance rules.

Across these tasks, evaluation should report five groups of metrics: asset quality; operational efficiency; institutional correctness (contract fulfillment, provenance completeness, rights validity, rollback fidelity); governance quality (faithful justifications, escalation quality, reversibility); and portfolio outcomes such as diversity, consistency, and long-horizon value. Existing multimodal agent benchmarks remain valuable, but they measure only fragments of the full organizational problem [13,14,16,52,54,55].

Table 3. Representative evaluation dimensions for AMFs. The framework is designed to separate model quality from organization quality.

Metric family	Representative measurements	Failure mode exposed
Asset quality	Faithfulness, realism, edit fidelity, subtitle accuracy, accessibility quality	Low-level perception or generation weakness
Operational efficiency	Latency, compute, external spend, human-minutes, turnaround time	Inefficient routing, over-automation, or poor make-or-buy choices
Institutional correctness	Contract satisfaction, provenance completeness, rights validity, approval compliance, rollback fidelity	Governance failures hidden by superficially strong media outputs
Governance quality	Escalation precision and recall, justification faithfulness, authority violations, reversibility	Misallocated oversight or unauthorized organizational commitments
Portfolio outcomes	Diversity, consistency, reuse efficiency, complaint rate, remediation cost, long-horizon value	Portfolio drift, hidden liabilities, and bad feedback loops across assets

New benchmark resources will also be needed. These include asset-graph datasets with explicit derivative structure, multimodal rights annotations, revocation events, policy changes, and platform simulators where distribution decisions change realized value. In other words, the benchmark must contain not only media but also *institutions*.

A crucial methodological consequence follows. Future comparisons should not evaluate only *models*; they should evaluate *organizational designs*. Ablations should vary contract richness, ledger completeness, internalization policy, authority matrices, and human-review budgets while holding the underlying multimodal model fixed. Otherwise, the field cannot determine whether gains come from

 doi:10.20944/preprints202603.2223.v1

smarter generation or better organization. For a vision paper centered on institutions, that distinction is the central empirical question.

6. Predictions, Implications, and Failure Modes

The media-firm hypothesis makes several concrete predictions about where the field is heading.

Prediction 1: the best multimedia agents will optimize contracts, not prompts.

Prompt engineering will remain useful, but systems that compile instructions into typed contracts with authority and acceptance boundaries should outperform prompt-only systems in long-horizon media work.

Prediction 2: provenance debt will become a decisive failure mode.

Just as technical debt accumulates when software shortcuts defer maintenance, provenance debt accumulates when assets are created or edited without sufficient lineage, rights, or approval state. Such debt remains partly invisible until incident response, where it explodes into search and remediation cost.

Prediction 3: human roles will narrow but become more strategic.

Routine editing and asset production will be increasingly automated, while humans shift toward commissioning, approval, taste judgment, exception handling, and value-laden decisions. Evidence from workplace studies already suggests that AI changes task allocation and team structure rather than simply replacing entire jobs [38–40].

Prediction 4: market structure will be path dependent.

If shared protocols and low-friction delegation dominate, multimedia production may partially unbundle into thin orchestrators coordinating many specialized providers. If data, telemetry, and provenance ledgers become strong moats, vertically integrated AMFs may instead consolidate power [42,45,46]. Multimedia is likely to experience both tendencies in different subdomains.

The framework also surfaces failure modes that existing benchmarks barely see. *Portfolio drift* occurs when local optimizations gradually erode consistency, trust, or accessibility across many assets. *Liability leakage* occurs when obligations are lost across derivatives. *Escalation collapse* occurs when humans are flooded with low-value reviews because the system cannot distinguish high-stakes from low-stakes uncertainty. *Distribution gaming* occurs when asset generation overfits platform metrics and degrades information quality or cultural diversity. These are organizational pathologies, not merely model errors.

6.1. Boundary Conditions and Objections

Two objections deserve a direct response. First, the firm language is not a claim of legal personhood or independent moral status. It is an engineering abstraction for systems that have bounded authority, persistent state, and externally consequential commitments. The point is to import a sharper vocabulary for contracts, decision rights, and liabilities into multimedia systems design, not to anthropomorphize current agents [43,44].

Second, not every multimedia pipeline needs this abstraction. Many creators will continue to need only tools or lightweight workflows. The media-firm hypothesis becomes useful when derivative depth, make-or-buy choice, and remediation cost are all high. This boundary condition matters scientifically because it turns the proposal into a spectrum rather than a slogan: one can empirically test where the workflow abstraction breaks and the firm abstraction begins to add predictive power.

A related objection is that better memory or better prompting might already solve the problem. Our response is that memory without authority is insufficient. A system may remember provenance yet still lack a rule for who may publish, who may spend external budget, or when a complaint requires suspension rather than silent editing. Likewise, a long and successful trajectory may still

 doi:10.20944/preprints202603.2223.v1

fail institutionally if it cannot guarantee reversibility after public release. The institutional layer is therefore not redundant with context length, planning depth, or retrieval quality. It is the mechanism that turns capability into governable action.

6.2. Infrastructure Implications

If the media-firm hypothesis is correct, the enabling stack for future multimedia agents will look different from today's default agent stack. The key components will not be only foundation models, vector stores, and tool wrappers. They will also include contract languages, asset registries, cross-modal lineage stores, policy engines, licensing-aware retrieval interfaces, and complaint-handling sandboxes. In other words, the field will need infrastructure for *commitment management*, not only content generation.

This matters for open research infrastructure as well. Shared benchmarks should expose not just tasks but reusable institutional components: typed briefs, approval schemas, provenance manifests, revocation events, platform feedback loops, and standardized interfaces for subcontracting tools or external agents. Current interoperability work suggests that the ecosystem is moving toward more portable tool and agent interfaces [19,20]. The multimedia opportunity is to ensure that such interfaces also carry institutional semantics such as usage rights, disclosure requirements, and publication authority rather than only input-output signatures.

A practical implication is that the most valuable community resources may soon resemble miniature economies more than static datasets. A benchmark repository could include assets, contracts, platform simulators, and adversarial incidents. A provenance standard could become useful only once paired with retrieval and remediation APIs. A watermark could become strategically important only when the organization knows how to trace descendants and decide who must act. This is why the paper's central claim is not reducible to better metadata or better agents in isolation. The frontier is the design of the *organizational substrate* that makes multimedia autonomy governable.

A final implication concerns scientific scope. The media-firm hypothesis does not reduce multimedia research to economics or management. Rather, it enlarges multimedia systems research by showing that asset retrieval, generation, provenance, editing, accessibility, search, recommendation, and governance are becoming one coupled systems problem. Multimedia is where agentic AI first meets asset governance at scale.

7. Conclusion

This paper argued that the next leap in autonomous multimedia systems is institutional. Once capable agents produce and manage portfolios of media assets rather than isolated outputs, the right abstraction is no longer only the model, the prompt, or the workflow. It is the organization. We introduced the media-firm hypothesis and the AMF as a concrete research object, explained why multimedia is a uniquely revealing domain for this shift, derived a boundary rule and four propositions, and proposed a benchmark agenda centered on contracts, ledgers, rollback, and governance. The central claim is deliberately ambitious but scientifically testable: in the coming wave of multimedia systems, the decisive bottleneck will often be not capability alone, but the design of the organizational substrate that turns capability into governable commitment.

Acknowledgments: This research is supported by the RIE2025 Industry Alignment Fund (Award I2301E0026) and the Alibaba–NTU Global e-Sustainability CorpLab.

References

1. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* **2023**.
2. Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805* **2023**.

 doi:10.20944/preprints202603.2223.v1

3. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *Advances in Neural Information Processing Systems* **2023**, *36*, 34892–34916.
4. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A survey on multimodal large language models. *National Science Review* **2024**, *11*, nwae403.
5. Bai, S.; Cai, Y.; Chen, R.; Chen, K.; Chen, X.; Cheng, Z.; Deng, L.; Ding, W.; Gao, C.; Ge, C.; et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631* **2025**.
6. Xu, J.; Guo, Z.; Hu, H.; Chu, Y.; Wang, X.; He, J.; Wang, Y.; Shi, X.; He, T.; Zhu, X.; et al. Qwen3-Omni Technical Report. *arXiv preprint arXiv:2509.17765* **2025**.
7. Li, Y.; Liu, Z.; Li, Z.; Zhang, X.; Xu, Z.; Chen, X.; Shi, H.; Jiang, S.; Wang, X.; Wang, J.; et al. Perception, Reason, Think, and Plan: A Survey on Large Multimodal Reasoning Models. *arXiv preprint arXiv:2505.04921* **2025**.
8. Zhao, S.; Zhang, X.; Guo, J.; Hu, J.; Duan, L.; Fu, M.; Chng, Y.X.; Wang, G.H.; Chen, Q.G.; Xu, Z.; et al. Unified Multimodal Understanding and Generation Models: Advances, Challenges, and Opportunities. *arXiv preprint arXiv:2505.02567* **2025**.
9. Zhong, Z.; Ji, Y.; Kong, Z.; Liu, Y.; Wang, J.; Feng, J.; Liu, L.; Wang, X.; Li, Y.; She, Y.; et al. Anytalker: Scaling multi-person talking video generation with interactivity refinement. *arXiv preprint arXiv:2511.23475* **2025**.
10. Ji, Y.; Zhang, J.; Xia, H.; Chen, J.; Shou, L.; Chen, G.; Li, H. Specvlm: Enhancing speculative decoding of video llms via verifier-guided token pruning. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 7216–7230.
11. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.R.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the The Eleventh International Conference on Learning Representations, 2022.
12. Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversations. In Proceedings of the First Conference on Language Modeling, 2024.
13. Koh, J.Y.; Lo, R.; Jang, L.; Duvvur, V.; Lim, M.; Huang, P.Y.; Neubig, G.; Zhou, S.; Salakhutdinov, R.; Fried, D. VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 881–905.
14. Xie, T.; Zhang, D.; Chen, J.; Li, X.; Zhao, S.; Cao, R.; Hua, T.J.; Cheng, Z.; Shin, D.; Lei, F.; et al. OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. *Advances in Neural Information Processing Systems* **2024**, *37*, 52040–52094.
15. Yang, J.; Tan, R.; Wu, Q.; Zheng, R.; Peng, B.; Liang, Y.; Gu, Y.; Cai, M.; Ye, S.; Jang, J.; et al. Magma: A Foundation Model for Multimodal AI Agents. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 14203–14214.
16. Huang, P.; Zhong, Z.; Wan, Z.; Zhou, D.; Alam, S.; Wang, X.; Li, Z.; Dou, Z.; Zhu, L.; Xiong, J.; et al. MMDeepResearch-Bench: A Benchmark for Multimodal Deep Research Agents. *arXiv preprint arXiv:2601.12346* **2026**.
17. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. PCA-OS: A Planetary Climate Adaptation Operating System. *Preprint* **2026**.
18. Du, M.; Xu, B.; Zhu, C.; Wang, X.; Mao, Z. DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents. *arXiv preprint arXiv:2506.11763* **2025**.
19. Hou, X.; Zhao, Y.; Wang, S.; Wang, H. Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions. *ACM Transactions on Software Engineering and Methodology* **2025**.
20. Duan, Q.; Lu, Z. Agent Communications toward Agentic AI at Edge: A Case Study of the Agent2Agent Protocol. In Proceedings of the 2025 IEEE 11th International Conference on Edge Computing and Scalable Cloud (EdgeCom). IEEE, 2025, pp. 150–155.
21. Yao, H.; Zhang, R.; Huang, J.; Zhang, J.; Wang, Y.; Fang, B.; Zhu, R.; Jing, Y.; Liu, S.; Li, G.; et al. A Survey on Agentic Multimodal Large Language Models. *arXiv preprint arXiv:2510.10991* **2025**.
22. Luo, J.; Zhang, W.; Yuan, Y.; Zhao, Y.; Yang, J.; Gu, Y.; Wu, B.; Chen, B.; Qiao, Z.; Long, Q.; et al. Large Language Model Agent: A Survey on Methodology, Applications, and Challenges. *arXiv preprint arXiv:2503.21460* **2025**.
23. Coase, R.H. The Nature of the Firm. *Economica* **1937**, *4*, 386–405.
24. Williamson, O.E. *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*; Free Press, 1985.

 doi:10.20944/preprints202603.2223.v1

25. Simon, H.A. Organizations and Markets. *Journal of Economic Perspectives* **1991**, *5*, 25–44.
26. Grossman, S.J.; Hart, O.D. The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy* **1986**, *94*, 691–719.
27. Jensen, M.C.; Meckling, W.H. Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. *Journal of Financial Economics* **1976**, *3*, 305–360.
28. Malone, T.W.; Yates, J.; Benjamin, R.I. Electronic markets and electronic hierarchies. *Communications of the ACM* **1987**, *30*, 484–497.
29. Yang, Z.; Li, L.; Wang, J.; Lin, K.; Azarnasab, E.; Ahmed, F.; Liu, Z.; Liu, C.; Zeng, M.; Wang, L. MM-ReAct: Prompting ChatGPT for Multimodal Reasoning and Action. *arXiv preprint arXiv:2303.11381* **2023**.
30. Zhou, S.; Xu, F.F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Ou, T.; Bisk, Y.; Fried, D.; et al. WebArena: A Realistic Web Environment for Building Autonomous Agents. *arXiv preprint arXiv:2307.13854* **2023**.
31. Koh, J.Y.; McAleer, S.; Fried, D.; Salakhutdinov, R. Tree search for language model agents. *arXiv preprint arXiv:2407.01476* **2024**.
32. Rosenthol, L. C2PA: The World's First Industry Standard for Content Provenance (Conference Presentation). In Proceedings of the Applications of Digital Image Processing XLV. SPIE, 2022, Vol. 12226, p. 122260P.
33. Bohacek, M.; Vilanova Echavarri, I. Compliance Rating Scheme: A Data Provenance Framework for Generative AI Datasets. In Proceedings of the Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 12150–12159.
34. Lin, L.; Gupta, N.; Zhang, Y.; Ren, H.; Liu, C.H.; Ding, F.; Wang, X.; Li, X.; Verdoliva, L.; Hu, S. Detecting Multimedia Generated by Large AI Models: A Survey. *arXiv preprint arXiv:2402.00045* **2024**.
35. Liu, S.; Qi, Z.; Xu, J.J.; Yan, Y.; Zhang, J.; Geng, H.; Liu, A.; Jiang, P.; Liu, J.; Tam, Y.C.; et al. VLA-Mark: A Cross-Modal Watermark for Large Vision-Language Alignment Models. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 26420–26438.
36. Huang, Y.; Chen, J.; Zheng, Q.; Li, H.; Liu, S.; Hu, X. Video Signature: In-Generation Watermarking for Latent Video Diffusion Models. *arXiv e-prints* **2025**, pp. arXiv–2506.
37. Rebuffi, S.A.; Tran, T.; Lacatusu, V.; Fernandez, P.; Souček, T.; Jovanović, N.; Sander, T.; Elsahar, H.; Mourachko, A. Learning to Watermark in the Latent Space of Generative Models. *arXiv preprint arXiv:2601.16140* **2026**.
38. Brynjolfsson, E.; Li, D.; Raymond, L. Generative AI at Work. *The Quarterly Journal of Economics* **2025**, *140*, 889–942.
39. Ju, H.; Aral, S. Collaborating with AI Agents: Field Experiments on Teamwork, Productivity, and Performance. *arXiv preprint arXiv:2503.18238* **2025**.
40. Xu, F.; Hou, J.; Chen, W.; Xie, K. Generative AI and Organizational Structure in the Knowledge Economy. *arXiv preprint arXiv:2506.00532* **2025**.
41. Korinek, A. AI agents for economic research. Technical Report 34202, National Bureau of Economic Research, 2025.
42. Hadfield, G.K.; Koh, A. An Economy of AI Agents. *arXiv preprint arXiv:2509.01063* **2025**.
43. Humberd, B.K.; Latham, S.F. When AI Becomes an Agent of the Firm: Examining the Evolution of AI in Organizations through an Agency Theory Lens. *Journal of Management Studies* **2026**, *63*, 668–694.
44. Tomašev, N.; Franklin, M.; Osindero, S. Intelligent AI Delegation. *arXiv preprint arXiv:2602.11865* **2026**.
45. Farach, A. AI as Coordination-Compressing Capital: Task Reallocation, Organizational Redesign, and the Regime Fork. *arXiv preprint arXiv:2602.16078* **2026**.
46. Klein, T.; Wieczorek, S. The Headless Firm: How AI Reshapes Enterprise Boundaries. *arXiv preprint arXiv:2602.21401* **2026**.
47. Walsh, J.P.; Ungson, G.R. Organizational Memory. *Academy of Management Review* **1991**, *16*, 57–91.
48. Aghion, P.; Tirole, J. Formal and Real Authority in Organizations. *Journal of Political Economy* **1997**, *105*, 1–29.
49. Cyert, R.M.; March, J.G. *A Behavioral Theory of the Firm*; Prentice-Hall, 1963.
50. Nelson, R.R.; Winter, S.G. *An Evolutionary Theory of Economic Change*; Belknap Press of Harvard University Press, 1982.
51. Huang, L.; Xiao, W.; Vishnoi, N.K. Delegation and Verification Under AI. *arXiv preprint arXiv:2603.02961* **2026**.
52. Jiang, D.; Zhang, R.; Guo, Z.; Wu, Y.; Qiu, P.; Lu, P.; Chen, Z.; Song, G.; Gao, P.; Liu, Y.; et al. MMSearch: Unveiling the Potential of Large Models as Multi-Modal Search Engines. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.

 doi:10.20944/preprints202603.2223.v1

53. He, C.; Zhou, X.; Yu, X.; Zhang, L.; Zhang, Y.; Wu, Y.; Xiao, L.; Li, L.; Wang, D.; Xu, H.; et al. SSKG Hub: An Expert-Guided Platform for LLM-Empowered Sustainability Standards Knowledge Graphs. *arXiv preprint arXiv:2603.00669* **2026**.
54. Zhang, L.; Zhou, X.; He, C.; Wang, D.; Wu, Y.; Xu, H.; Liu, W.; Miao, C. MMESGBench: Pioneering Multimodal Understanding and Complex Reasoning Benchmark for ESG Tasks. In Proceedings of the Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 12829–12836.
55. He, C.; Zhou, X.; Wu, Y.; Yu, X.; Zhang, Y.; Zhang, L.; Wang, D.; Lyu, S.; Xu, H.; Wang, X.; et al. EsGenius: Benchmarking LLMs on Environmental, Social, and Governance (ESG) and Sustainability Knowledge. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 14623–14664.