Large Language Models Can Follow Instructions, But Not Many at Once: Phase Transitions in Compositional Constraint Satisfaction

Mariya I. Vasileva [1]

Abstract

Large language models are increasingly deployed in settings that require simultaneous adherence to multiple explicit constraints. While individual constraints are often handled proficiently, the compositional regime—where many must hold jointly—remains poorly characterized: how rapidly does performance degrade, what governs the degradation, and is there a predictable ceiling? We introduce *Constraint Saturation Evaluation (CSE)*, a procedurally generated stress-testing benchmark that systematically varies the number of simultaneous constraints (k), each scored by fully deterministic, rule-based verifiers with zero LLM-judge involvement. We apply CSE to two structurally different tasks: constrained text generation, and constrained visual selection from procedurally generated scenes. In the text modality, three principal findings emerge. First, every model obeys *a two-regime decay pattern*: performance drops exponentially with constraint count before plateauing at an asymptotic floor. Second, constraint failures are *approximately independent*, implying that performance collapse arises from multiplicative accumulation of individual failures rather than pairwise interference. Third, a *depth-of-processing hierarchy* governs which constraints fail first: relational constraints degrade $2.4\times$ faster than lexical ones under compositional load—yet this ordering is uncorrelated with intrinsic single-constraint difficulty, pointing to a general capacity limit rather than constraint-specific bottlenecks. Across all models, reliable compositional instruction following breaks down beyond 5–6 simultaneous constraints. In the vision modality, the pattern is qualitatively different: the decay is *model-dependent*. Four of ten VLMs show compositional decay comparable to text, while six plateau or invert, and constraint failures are massively correlated rather than independent—a single perception factor explains 87.9% of co-failure variance. Whether a VLM degrades under compositional load depends on its perceptual accuracy: models that perceive the scene accurately experience near-zero compositional cost, while weaker perceivers show decay comparable to text.

1. Introduction

Language models deployed in complex applications must adhere to multiple explicit constraints simultaneously—reasoning structure, safety boundaries, domain-specific protocols, output schemas—within a single response. Models handle individual constraints well, but performance under composition is poorly understood: does it degrade gracefully, or is there a sharp transition? Existing benchmarks evaluate at low constraint counts ($k \leq 5$) or report aggregate difficulty without characterizing the trajectory, leaving the functional form of the degradation and the mechanisms driving it as open questions.

IFEval (Zhou et al., 2023) applies deterministic programmatic verification across 25 constraint types but composes at most three constraints per prompt ($k \leq 3$) and does not report performance as a function of k. COLLIE (Yao et al., 2023) permits arbitrary recursive composition and observes that composition increases difficulty, but compares specific task pairs rather than sweeping k. FollowBench (Jiang et al., 2024) adds one constraint per level up to $k=5$ and reports GPT-4's consistent-satisfaction ceiling at $\sim k=3$ (CSL $= 3.3$); however, its open-ended subset is scored by a GPT-4 judge, raising a potential self-evaluation bias — though the authors report 88% agreement with human annotators. InfoBench (Qin et al., 2024) decomposes instructions into fine-grained binary criteria and finds that the strongest model (GPT-4) satisfies only $\sim$89% of them overall while other models satisfy 78–87%, with a consistent constraint-type hierarchy: highest on content/style, lowest on number and linguistic constraints across all six models; we extend this hierarchy across k. RECAST (Guo et al., 2025) averages 13.4 constraints per instance in its training corpus

[1]Work done while at Meta Superintelligence Labs. Correspondence to: Mariya I. Vasileva <maria.i.vasileva@gmail.com>.

Accepted to the 2nd Workshop on Compositional Learning at ICML 2026, Seoul, South Korea. *Copyright 2026 by the author(s).*

but its benchmark reports at only four coarse difficulty tiers (k=5/10/15/all), so per-k decay cannot be measured from its results. In constrained tool use, CCTU (Ye et al., 2026) finds that no model exceeds 20% task completion under strict multi-constraint adherence; we observe a comparably low ceiling on text in CSE, suggesting that strict multi-constraint satisfaction is hard across modalities. Outside text, ConceptMix (ConceptMix Authors, 2024) reports that DALL·E 3 drops from 83% to 8% as concept count grows from k=1 to k=7, though evaluation relies on GPT-4o as judge rather than deterministic verification. Benchmarks that use LLM-based evaluation (Jiang et al., 2024; Guo et al., 2025; ConceptMix Authors, 2024) further face a circularity at high k: the judge may itself struggle with compositional constraint violations in exactly the regime where models fail most.

The difficulty of compositional generalization is well-established (Lake & Baroni, 2018; Dziri et al., 2023): Dziri et al. (2023) show that transformers reduce multi-step compositional reasoning to linearized subgraph matching, with performance decaying exponentially as composition depth increases. Press et al. (2023) quantify a "compositionality gap" — the fraction of multi-hop questions a model fails despite answering all sub-questions correctly — that persists at $\sim$40% regardless of model scale. What is left open is a characterization of the functional form for *instruction-following* specifically: whether the same exponential decay holds, and whether constraint failures interact or accumulate independently.

To answer these questions, we introduce **Constraint Saturation Evaluation (CSE)**. Every constraint is verified by a deterministic function rather than an LLM judge, a design choice that makes CSE's measurements invariant to the same compositional degradation it studies. We apply CSE to constrained text generation and extend it to vision, where the task becomes constrained *selection* over procedurally generated scenes with ground-truth scene graphs. This eliminates evaluator confounds, but it also changes the task structure: in generation, each added constraint shrinks the valid output space, while in selection, each constraint narrows the candidate set. The consequences of this structural difference — whether the decay profile reflects task structure, modality, or both — are examined in Section 4. **Text CSE** tests constrained free-form generation: 30 constraint types across 8 linguistic dimensions, 1,194 probes at k=1 to 12, and 14 models from 7 families (82,628 constraint checks). **Visual CSE** tests constrained selection from procedurally generated scenes: 18 visual constraints across 5 dimensions, 1,523 probes at k=1 to 8, and 10 VLMs from 4 families (70,410 constraint checks), verified deterministically against hidden scene graphs.

We make the following contributions:

1. We identify a **two-regime decay pattern** that governs compositional instruction following across all 14 evaluated models: constraint satisfaction rate (CSR) drops exponentially from 58% at k=1 before plateauing at an asymptotic floor ($\sim$16%), captured by a three-parameter model that predicts held-out performance within 1.4pp (Section 3).
2. We establish that constraint failures are **approximately independent**, demonstrating that compositional collapse arises from multiplicative accumulation of individual failures rather than pairwise interference (Section 3.1).
3. We discover a **depth-of-processing hierarchy** in which relational constraints degrade 2.4× faster than lexical ones under compositional load, yet this ordering is uncorrelated with intrinsic difficulty, pointing to a general capacity limit rather than constraint-specific bottlenecks (Section 3.1).
4. We extend CSE to vision and find that the decay is **model-dependent**: 4 of 10 VLMs decay comparably to text (-17 to -28pp) while 6 plateau or invert. Failures are massively correlated rather than independent, with a single perception factor explaining 87.9% of co-failure variance. Whether a model degrades depends on its perceptual accuracy: models above $\sim$60% experience near-zero compositional cost, while weaker perceivers ($< \sim$50%) show decay comparable to text (Section 3.2).

2. Constrained Saturation Evaluation (CSE) Framework

2.1. Text Constraint Saturation Evaluation (CSE)

Constraint taxonomy. CSE defines 30 constraint types across 8 dimensions (Table 6), selected to be deterministically verifiable and to span a range of processing demands. Constraints range from local pattern matching (lipogram: avoid a letter; forbidden word: exclude a token) through document-level structure (exact paragraph count, monotonic sentence length) to relational reasoning requiring structured output (scene graphs, logic grids) and self-referential reasoning (a sentence correctly reporting its own word count). We test whether this produces a measurable hierarchy in degradation rates in Section 3.1. All 30 verifiers are fully deterministic, each returning both a binary pass/fail judgment and a continuous score $s \in [0, 1]$ representing partial compliance.

The comprehension-maintenance gap. We define the *comprehension-maintenance gap* for each constraint as

$$\Delta_i = \bar{s}_i - \text{CSR}_i, \tag{1}$$

where $\bar{s}_i$ is the mean continuous score, and CSR_i the strict

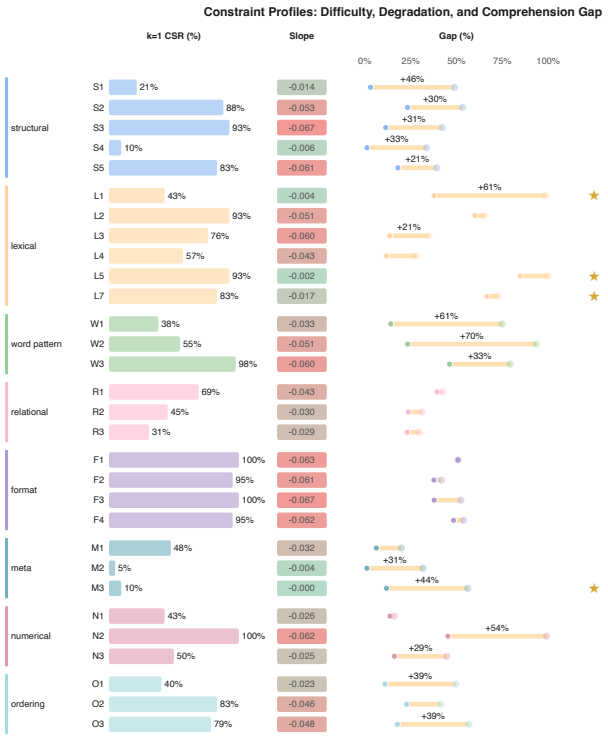


Figure 1. Constraint profiles for all 30 CSE constraints. **Left:** k=1 CSR. **Center:** degradation slope (warm = fast collapse). **Right:** comprehension-maintenance gap Δ (Equation (1)). Stars mark floor survivors (L1, L5, L7, M3). Slope is uncorrelated with k=1 difficulty ($\rho = -0.03$).

binary satisfaction rate. A large Δ indicates that a model partially complies, demonstrating comprehension of the constraint, but fails to maintain full constraint satisfaction. Per-constraint CSRs, gaps, and complete specifications appear in Table 6 and Section F.

Probe generation. CSE contains 1,194 procedurally generated probes. For each constraint count $k \in 1, 2, \ldots, 12$, we sample a probe by drawing k constraints from the 30-type pool without replacement, then instantiating their parameters from difficulty-calibrated ranges. Each probe is also given a topic prompt to anchor generation. Probe counts per k range from 90 (at k=1) to 18 (at k=12); the drop at high k comes from incompatibility filtering. An incompatibility checker rejects provably unsatisfiable combinations (e.g., embed numbers" + no digits") and parameter-dependent conflicts (e.g., a lipogram forbidding a letter that appears in a mandatory word). Rejection rates rise with k, from $\sim$5% at k=4 to $\sim$78% at k=12. As a result, surviving probes at high k are drawn from a narrower set of compatible combinations—a bias we control for in Section 3.1. Additionally, 21 deliberately impossible probes test whether models recognize contradictory constraint combinations rather than hallucinating compliance. All constraint parameters are sampled at `MEDIUM` difficulty uniformly across k values (e.g., word-count targets of 40–60, or lipograms banning letters of middling frequency in English (ranks 5–15)). Holding difficulty fixed across k ensures that the phase transition reflects compositional burden alone, rather than harder parameters.

Extraction and verification. Each model receives the full prompt: a topic and all k constraint instructions in natural language. We provide no chain-of-thought scaffolding, decomposition hints, or constraint-ordering cues. The full set of 30 verifiers is listed in Table 6, with specifications in Section F. A probe passes only if all k constraints pass simultaneously — strict conjunction, no partial credit.

2.2. Visual Constraint Saturation Evaluation (vCSE)

Visual CSE (vCSE) extends the constraint-saturation framework to vision-language models, using procedurally generated CLEVR-style scenes (Johnson et al., 2017) verified deterministically against hidden scene graphs. The task differs structurally from CSE: the model sees a rendered image of 12–15 objects (mean 13.4) and must identify which one(s) satisfy k visual constraints at once. This is constrained *selection* from a fixed candidate set rather than constrained *generation* over an unbounded output space — a difference that shapes the expected decay profile (Section 4).

Constraint taxonomy. vCSE uses 18 composable filter constraints across 5 dimensions: *attribute* (A1 color match, A2 shape match, A3 size match), *spatial* (S1 left/right, S2 above/below, S3 adjacency, S4 isolation, E1 extremal position), *relational* (R1 same/different attribute, R2 comparative size), *set/logical* (L1 threshold logic, L2 disjunction, L3 negation, L4 uniqueness, L5 attribute-count comparison), and *counting* (C1 counting filter, C2 group parity, C3 plurality). Full specifications are given in Section I.

Scene-first composition. Probes are built by a *scene-first* procedure that guarantees satisfiability by construction: we select a target object, compute its constraint signature, and sample k constraints from it. A small set of 60 deliberately impossible probes built by inverting the scene-first procedure, then sampling constraint combinations that *no* single object satisfies, tests whether models recognize infeasibility instead of forcing a selection (Section 3.2, Section P).

Extraction and verification. Each model is prompted with the image and a text instruction containing all k constraints, and asked to return its answer inside an `<answer>` tag for reliable extraction. All 1,523 probes across all 10 models produced a well-formed tag (100% compliance). The extracted object IDs are then checked against the scene graph by each constraint's deterministic verifier.

Table 1. Text CSE: 14 models ranked by overall CSR (82,628 constraint checks, temperature=0, greedy decode). **CSR@1** is the single-constraint satisfaction rate (k=1 probes; Table 4) — the no-composition ceiling. **CSR** is the overall constraint satisfaction rate across all k. **Score** is the mean partial-compliance score ($s \in [0, 1]$, as a percentage). **Gap** is the comprehension-maintenance gap (Score−CSR, Equation (1)): partial compliance the model attains but does not convert into strict satisfaction. **Bold** marks the best (higher is better) in each of CSR@1, CSR, and Score.

Model	CSR@1 ↑	CSR% ↑	Score% ↑	Gap
Gemini 3.1 Flash-Lite	78.9	**46.2**	**70.8**	24.6
Kimi K2.5	67.8	39.4	64.9	25.5
Llama 3.1 405B	60.0	31.2	57.3	26.1
Claude 4.6 Opus	70.0	29.0	52.8	23.8
GPT-5.4 Pro	**82.2**	28.5	51.5	23.0
Qwen 3 235B Inst.	55.6	27.8	55.5	27.7
GPT-5.2	68.9	26.2	49.6	23.4
Llama 3.1 70B	57.8	25.8	51.4	25.6
Claude 4.7 Opus	78.9	25.1	49.1	24.0
Gemini 2.5 Flash	66.7	25.0	49.5	24.5
Llama 3.1 8B	38.9	24.7	51.2	26.5
Gemini 3.1 Pro	73.3	22.7	45.6	22.9
Claude 4.5 Sonnet	48.9	17.6	43.0	25.4
Qwen 3 235B Base	50.0	15.9	40.2	24.3

3. Experiments

3.1. Text results.

Phase transition and decay. We evaluate 14 models spanning 7 families (Table 1), and find that every instruction-tuned model traces the same trajectory: high performance at k=1 (39–82%), a steep decline through k=4−7, and a floor at $k \geq 8$ (Table 4). This shape holds across all 7 families despite differences in architecture, scale, and training. The overall CSR across 82,628 constraint checks is 27.5%.

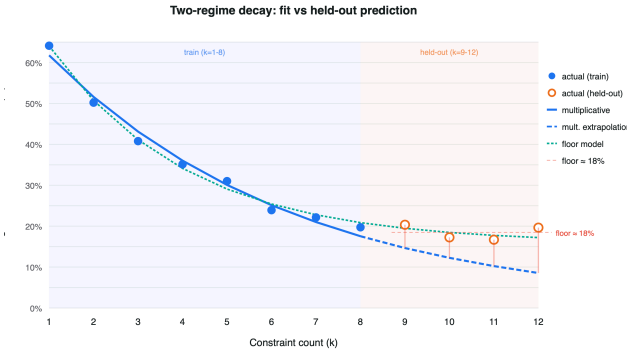
The transition is driven by k, not by difficulty. A natural objection is that higher-k probes might simply draw harder individual constraints, manufacturing the apparent collapse. We rule this out with an unconfounded difficulty measure — each constraint's k=1-only CSR, which isolates single-constraint capability from compositional load. Two checks confirm that intrinsic difficulty is balanced across k. First, a Kruskal–Wallis test on this measure finds no significant variation in mean per-constraint difficulty across k strata (H=5.20, p=0.92): the probe composer succeeded in holding difficulty constant. Second, stratifying probes within each k into Easy/Medium/Hard terciles yields the same qualitative collapse in every tercile—easy probes start higher but reach the floor by k=5−6 alongside hard ones. The k=1 measure also diverges sharply from the marginal CSR pooled over all k (Spearman ρ=0.714), and this divergence is itself diagnostic: it pinpoints which constraints are most composition-vulnerable. Paragraph structure (S3) is the extreme case, falling from 92.9% at k=1 to 11.3% marginal — an 81.6-point composition penalty despite being nearly trivial in isolation.

Two-regime decay model. If constraint failures are approximately independent (Section 3.1), the aggregate CSR should follow a multiplicative model: $P(k) = a \cdot r^{k-1}$. We test this prediction by fitting two models on k=1−8 and evaluating on held-out k=9−12 (Figure 2a). The pure multiplicative model (a=62%, r=0.84) accurately describes training data (MAE = 1.5pp) but systematically underpredicts at every held-out k value (MAE = 7.1pp, bias = −7.1pp). Rather than decaying to zero, models plateau. This systematic failure motivates a three-parameter floor model, $P(k) = a \cdot e^{-\beta k} + c$, which captures the asymptote (Figure 2a). The floor model (a=67%, β=0.32, c=15.8%) predicts held-out k=9−12 within 1.4pp MAE — a 5× improvement over the multiplicative baseline. The fitted floor of 15.8% matches observed rates of 16.7–20.3% at k=9−12. Per-k predictions versus actuals appear in Section D.

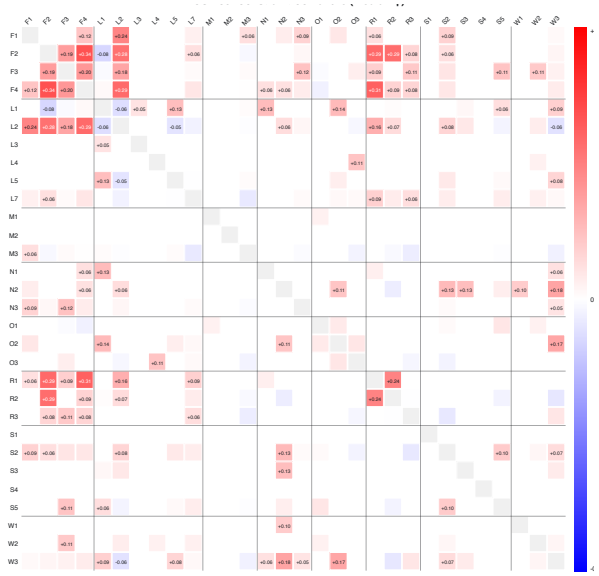
The floor is sustained by *compositionally immune* constraints: L5 (forbidden word, 79%), L7 (vowel ratio, 66%), L2 (mandatory words, 60%), N2 (no digits, 41%). These share a common characteristic: they are binary constraints that require no sustained maintenance during generation. A model either includes the mandatory word or does not; there is no partial compliance that degrades over the course of generation. This immunity to compositional load is what creates the asymptotic floor. Per-model decay factors range from $r = 0.75$ (Qwen Base, fastest decay) to $r = 0.94$ (Llama 8B, slowest), revealing two distinct profiles: *high-ceiling, fast-decay models* (e.g., GPT-5.4 Pro: a=80%, r=0.78) that start strong but collapse rapidly, and *low-ceiling, slow-decay models* (e.g., Llama 8B: a=34%, r=0.94) that start weak but barely degrade. Models that are best at k=1 are not necessarily best at k=12.

Constraint failures are approximately independent. A natural hypothesis is that compositional difficulty stems from constraint *interactions* — specific pairs that conflict, compete for shared capacity, or interfere with one another's satisfaction. Whether two constraints fail together is measured by the *phi coefficient* (φ), the correlation between their binary pass/fail outcomes (φ=0: independent; φ>0: co-failing); we write $\bar{\varphi}$ for the mean φ over a group of pairs. We evaluate the hypothesis directly by computing pairwise φ across all $\binom{30}{2} = 435$ constraint pairs conditioned on k, since otherwise the near-universal failure at high k would inflate apparent co-failure. To prevent weak models from distorting co-failure rates, each φ is computed per-model first, then aggregated via median across models.

Across the 432 analyzable pairs (3 excluded for insufficient co-occurrence), pairwise correlations are negligible: the mean phi coefficient is $\varphi = +0.017$ (s.d. 0.051), and 376 pairs (87%) fall within $|\varphi| \leq 0.05$. Testing 432 pairs at once inflates the chance of a spurious hit, so we apply the



(a) Two-regime decay. Blue: training CSRs (k=1−8). Red: held-out (k=9−12). The floor model (solid) predicts held-out within 1.4pp; the multiplicative model (dashed) underpredicts by 7.1pp.



(b) Pairwise co-failure (φ). Nearly uniform gray (mean φ=0.017), confirming approximate independence. The only elevated block is the format–relational cluster (F1–F4, R1–R3).

Figure 2. Core text CSE results.

Bonferroni correction — dividing the significance threshold by the number of tests — and find that no pair reaches significance: no pairwise correlation is distinguishable from chance. Compositional collapse is therefore not driven by pairwise constraint conflict.

Structured-output coupling. The only elevated φ values cluster among format and relational constraints: format–format ($\bar{\varphi} = 0.142$), format–relational ($\bar{\varphi} = 0.116$), relational–relational ($\bar{\varphi} = 0.080$). These reflect shared dependency on structured generation capability: a model that fails to produce valid JSON fails all format and relational constraints simultaneously, not constraint interference per se. L2 (mandatory words) is largely independent of the format constraints, coupling with only a few ($\bar{\varphi} = 0.046$ on average, but $\varphi = 0.18{-}0.29$ for individual pairs). The reason is shared output structure: when the model must place the mandatory words inside a JSON object or table and that structure collapses, the words have nowhere left to go and fail along with it. Outside the format–relational cluster, mean $|\varphi| < 0.02$, suggesting that *constraints are approximately independent*. No synergistic pairs exist ($\varphi < -0.05$) — so reordering or hand-picking constraints cannot help, leaving per-constraint reliability as the only lever for mitigation. Further analysis of within- vs. across-dimension structure can be found in Section L.

Degradation hierarchy. If constraints fail independently and the transition is purely combinatorial, does every constraint degrade at the same rate? To address this question, we compute per-constraint degradation slopes (linear fit of CSR vs. k) and aggregate by dimension (Table 2). Relational and word-level constraints degrade $2.4\times$ faster than lexical ones — an ordering that points to *depth of processing* as the driver. Shallow surface constraints (avoid a letter, include a word) survive because they reduce to simple pattern matching, while deep constraints (reason about object relationships, maintain vocabulary discipline across the full output) require sustained attention that gets overwhelmed under compositional load.

Four constraints (L1, L5, L7, M3) have slopes > -0.02 — essentially flat across k — forming the asymptotic floor identified in Section 3.1. Degradation rate is uncorrelated with intrinsic difficulty ($\rho = -0.030$, $p = 0.88$) and with the comprehension-maintenance gap Δ ($\rho = +0.196$, $p = 0.30$): the per-constraint pass rates accumulate multiplicatively, and what drives the transition is only *how often* each constraint is satisfied, not how hard it is or how it collapses. The appendix (Section M) reports three follow-ups: control analyses confirming that the degradation hierarchy (the ordering of dimensions by decay rate) is not an artifact of difficulty; a paradoxical effect in which a few constraints (e.g., M3, L1) actually improve as k grows due to longer compositional prompts eliciting longer outputs that happen to satisfy them; and a caveat that the fast degradation of relational constraints may be partly inflated by their tendency to fail together with format constraints (the shared-structured-

Table 2. Degradation rate by dimension. Each value is the slope of a linear fit of constraint satisfaction rate (CSR) against the constraint count k, averaged over the constraints in that dimension. The slope is expressed in CSR units per added constraint (here CSR is on a 0–1 scale, so -0.077 means a constraint loses about 7.7 percentage points of satisfaction for each additional constraint in the probe); more negative values indicate faster collapse under compositional load. Relational and word-level constraints degrade more than twice as fast as lexical ones. **Bold** marks the fastest- and slowest-degrading dimensions.

Dimension	Mean slope	Interpretation
Relational	$\mathbf{-0.077}$	Fastest
Word-level	-0.075	Fast
Format	-0.056	Medium
Structural	-0.045	Medium
Numerical	-0.041	Medium
Meta	-0.040	Slow
Ordering	-0.038	Slow
Lexical	$\mathbf{-0.032}$	Slowest

output coupling of Sections 3.1 and L), rather than reflecting relational difficulty alone.

Failure mode decomposition. The L1 (lipogram) constraint exhibits *the largest comprehension-maintenance gap in the benchmark* (+61.0pp; Table 6) and serves as a case study for failure mode analysis. Position-of-first-violation analysis on all n=1,420 L1 failures reveals three distinct failure modes — refusal (39%), meta-preamble self-narration (8%), and genuine content failure (53%). Refusal rates are driven by alignment training (Qwen Base: 12% vs. Instruct: 58%), and genuine failures cluster early in the response (median position 17.7%), indicating engagement failure rather than resource depletion.

The failure space is two-dimensional, and the two axes are independent. First, a model decides *whether to attempt* the constraint at all or refuse it – a choice set by alignment training (base models almost always try; heavily RLHF'd models usually decline). Second, among the attempts, a model fails at a family-characteristic position, i.e. where in the response the first violation appears (Claude slips almost immediately, Gemini holds out past the midpoint). Because the two axes are independent — a model's refusal rate does not predict where it slips when it *does* try — failure cannot be reduced to a single spectrum like "better models fail later." The full decomposition is in Section N. Per-model failure profiles — each model's refusal rate and where in the response it first slips — and the rule-based classifier that sorts each failed response into one of the three modes (refusal, meta-preamble self-narration, or genuine content failure) using regular-expression heuristics, are detailed in Section E.

Table 3. Visual CSE: 10 VLMs ranked by overall constraint satisfaction rate (CSR) on the 1,114 solvable probes (44,930 constraint checks, temperature=0). **Score** is the mean partial-compliance score ($s \in [0, 1]$, shown as a percentage). $\Delta_{1\to 8}$ is the change in CSR from k=1 to k=8 (negative=decay, positive=improvement). **Behavior** classifies each model's phase-transition profile: *steep* decay, *plateau*, or *inverted* (improves with k). **Bold** indicates best in column.

Model	CSR% ↑	Score% ↑	$\mathbf{\Delta_{1\to 8}}$	Behavior
Gemini 3.1 Pro	**79.3**	**83.6**	-1.7	Plateau
Gemini 3.1 Flash-Lite	70.9	73.6	-1.4	Plateau
GPT-5.4 Pro	70.8	73.7	$+7.4$	Inverted
Claude 4.6 Opus	67.0	70.1	$+15.2$	Inverted
Claude 4.7 Opus	62.2	64.3	-11.6	Plateau
Llama 3.2 90B Vision	56.5	61.2	$+17.0$	Inverted
Gemini 2.5 Flash	51.7	55.4	-21.4	Steep
Claude 4.5 Sonnet	44.4	47.5	-26.3	Steep
Llama 4 Maverick 17B	41.1	44.8	-16.8	Steep
GPT-5.2	37.8	40.7	-27.8	Steep

3.2. Visual results.

We evaluate on 1,114 probes with verified ground-truth solutions. Per-model results are summarized in Table 3.

Phase transition. Unlike text, where all 14 models decay universally, the visual phase transition is model-dependent. The cluster boundary is perceptual ability: weak perceivers decay, strong perceivers plateau. Three behavioral clusters emerge (full per-k values in Table 5).

Steep decayers (4 models) lose 17–28pp from k=1 to k=8: GPT-5.2 drops from 57.1% to 29.3%, Claude 4.5 from 62.6% to 36.3%, Gemini 2.5 Flash from 62.6% to 41.2%, and Maverick from 51.5% to 34.7%. These models have overall CSRs of 37.8–51.7%, meaning they misperceive the scene often enough that accuracy erodes with each added constraint: every constraint is another attribute or relation the model must read correctly, so per-constraint perception errors accumulate as k grows.

Plateau models (3 models) show minimal change: Gemini Pro drops just 1.7pp, Flash-Lite 1.4pp, and Claude 4.7 11.6pp. With CSRs of 62.2–79.3%, these models perceive the scene accurately enough that additional constraints add negligible failure probability.

Inverted models (3 models) actually *improve* with k: GPT-5.4 gains 7.4pp, Claude 4.6 gains 15.2pp, and Llama 90B gains 17.0pp. This occurs because additional constraints narrow the candidate set from ∼10 objects to 1–2, reducing *over-inclusion errors* — when many objects qualify, models include some that do not; as k grows the valid set shrinks toward one object, so there is little room left to over-include.

Co-failure structure. In stark contrast to text ($\varphi = 0.017$), visual constraint failures are massively correlated:

Table 4. Probe-level CSR (%) by model and constraint count k. A probe passes only if *all* k constraints are satisfied. **Bold** marks the best model at each k. All instruction-tuned models exhibit the same qualitative pattern: high k=1 (39–82%), steep decline, then a floor at $k\geq 8$. Gemini 3.1 Flash-Lite leads at $k\geq 2$ (GPT-5.4 Pro leads at k=1). Llama 8B shows remarkably flat degradation (k=1: 39%, k=12: 23%), resisting collapse more than its size would suggest. The k=12 uptick visible in several models reflects the incompatibility filter bias: surviving probes at k=12 exclude the hardest constraint combinations.

Model	k=1	2	3	4	5	6	7	8	9	10	11	12
Claude 4.5 Sonnet	48.9	41.0	29.5	28.6	20.0	14.8	12.6	7.8	9.2	7.9	7.6	11.6
Claude 4.6 Opus	70.0	61.5	47.1	44.9	36.4	24.1	21.5	21.1	17.9	11.9	7.6	14.8
Claude 4.7 Opus	78.9	63.4	45.9	38.1	27.9	20.8	18.6	13.9	11.0	10.0	7.3	12.0
GPT-5.2	68.9	53.9	44.9	36.6	34.5	22.0	17.1	11.9	17.1	13.5	13.0	16.7
GPT-5.4 Pro	**82.2**	60.0	47.3	38.6	29.2	23.5	17.1	16.8	20.5	17.1	17.0	21.3
Gemini 2.5 Flash	66.7	50.7	40.8	32.0	27.2	16.9	18.2	17.2	19.0	15.4	13.0	18.5
Gemini 3.1 Flash-Lite	78.9	**72.2**	**63.0**	**57.6**	**50.6**	**44.7**	**40.9**	**35.4**	**37.6**	**32.1**	**34.8**	**31.9**
Gemini 3.1 Pro	73.3	55.1	42.1	29.0	26.2	14.8	16.1	11.9	11.8	12.6	9.7	14.8
Kimi K2.5	67.8	54.9	51.0	47.7	43.6	37.7	34.5	34.4	33.2	27.8	33.0	32.4
Llama 3.1 405B	60.0	47.6	41.9	39.1	37.3	29.0	28.4	24.8	23.5	21.0	22.1	23.1
Llama 3.1 70B	57.8	40.2	32.3	30.0	28.9	25.6	22.8	21.1	22.4	19.0	17.3	14.8
Llama 3.1 8B	38.9	27.8	28.3	25.5	27.2	22.9	25.0	24.2	23.9	19.7	20.9	23.1
Qwen 3 235B Base	50.0	35.6	23.0	15.4	13.5	11.2	13.6	12.1	12.2	11.6	9.1	16.2
Qwen 3 235B Inst.	55.6	39.0	34.2	29.0	31.8	27.1	23.6	23.6	25.6	21.9	21.2	23.6

mean $\varphi = +0.887$ across 150 pairs, with all pairs positive and 9 reaching perfect co-failure ($\varphi = 1.0$) — a $52\times$ difference. Spectral decomposition reveals a single dominant factor ($\lambda_1 = 87.9\%$ of variance), with no dimension-specific structure beneath it. The multiplicative accumulation model does not hold for vision: independence predicts a 3.6% success rate at k=8, but the observed rate is 49.7%, pointing to a shared perceptual input as the driver. Whether this reflects modality or task structure is examined in Section 4. Per-constraint CSR heatmaps are shown in Section A.

Set/logical constraints degrade fastest (mean slope -0.047) and attribute constraints slowest (-0.028), a $1.7\times$ ratio (smaller than text CSE's $2.4\times$), with some constraints exhibiting *inverted* degradation (improving with k) due to answer-set cardinality narrowing. Directional interference analysis identifies L5 (attribute count comparison) as the dominant "difficulty amplifier": -16 to -24pp on co-occurring constraints, and E1 (extremal filter) as the dominant "synergist": $+12$ to $+19$pp. Full details appear in Sections O and P.

Impossibility detection. Following the impossible-probe design introduced in Section 2.2, we test whether models recognize unsatisfiable constraint combinations rather than forcing an answer. The evaluation draws on 409 impossible probes from two sources: 60 *deliberately* constructed contradictions — combinations the scene-first composer guarantees no object can jointly satisfy — and 349 probes the post-hoc satisfiability audit reclassified as secretly unsatisfiable (Section J). A model is scored as passing an impossible probe correctly when it reports that no valid object exists. Nine of 10 models detect impossibility above 84%; Llama 3.2 90B is the outlier at 64.5%. Detection improves with k: from 18% at k=2 to 91% at k=8, implying that more constraints provide more opportunities to identify conflicts. Critically, *solving ability and impossibility detection are inversely correlated*: GPT-5.2 is the weakest solver (37.8%) but tied for best detector (90.2%), while Gemini Pro is the strongest solver (79.3%) but an average detector (84.4%). This dissociation suggests that constraint satisfaction and conflict detection draw on partially independent capabilities.

4. Discussion

A combinatorial consequence. Under independence, the probability of satisfying all k constraints equals the product of individual pass rates. Each rate sits slightly below 1, so this product shrinks toward zero as k grows — a purely multiplicative decay driven by the per-constraint rates alone. In contrast, the k-SAT phase transition (Mertens et al., 2006; Mézard & Montanari, 2009) works differently: its threshold emerges from interaction structure (clause density and shared variables), whereas ours is affected only by each constraint's marginal pass rate. Since constraint failures accumulate independently (Section 3.1), the primary lever that shifts the transition rightward is per-constraint reliability: the joint pass rate improves only as individual constraints become more reliable. Constrained decoding (Lu et al., 2021; 2022) and attention steering (Lv et al., 2026) are promising candidate directions to this end, and an open question is whether they delay the transition's onset or raise its floor.

Two distinct phenomena. CSE reveals two phenomena in compositional instruction following that coexist but are mechanistically independent. First, a phase transition driven by combinatorial accumulation of approximately indepen-

Table 5. Visual CSE: constraint-level CSR (%) by model and constraint count k, computed on the 1,114 solvable probes (44,930 constraint checks). **Bold** marks the best model at each k. Unlike text, the visual decay is model-dependent: strong perceivers (Gemini 3.1 Pro, Flash-Lite) stay roughly flat, while weak perceivers (GPT-5.2, Maverick, Claude 4.5) decay steeply and three models (GPT-5.4 Pro, Claude 4.6 Opus, Llama 3.2 90B) *improve* with k as added constraints narrow the answer set (Section 3.2). Rows are ordered by overall solvable CSR (Table 3).

Model	k=1	2	3	4	5	6	7	8
Gemini 3.1 Pro	**74.7**	**83.6**	**81.3**	**85.3**	**79.5**	**80.4**	**77.5**	73.0
Gemini 3.1 Flash-Lite	73.7	78.7	76.6	69.7	72.4	62.4	69.3	72.3
GPT-5.4 Pro	65.7	79.1	70.9	67.7	75.3	69.0	67.5	73.1
Claude 4.6 Opus	61.1	69.4	73.1	63.7	68.2	60.1	62.5	**76.3**
Claude 4.7 Opus	71.7	73.9	70.5	60.8	64.0	58.0	55.4	60.1
Llama 3.2 90B Vision	42.4	53.7	57.7	56.6	59.5	57.7	53.5	59.4
Gemini 2.5 Flash	62.6	66.8	63.0	59.5	53.3	46.3	43.2	41.2
Claude 4.5 Sonnet	62.6	58.6	57.5	52.1	48.5	35.8	32.2	36.3
Llama 4 Maverick 17B	51.5	56.3	53.9	46.8	40.8	38.2	29.3	34.7
GPT-5.2	57.1	53.4	60.0	45.5	46.4	21.6	21.9	29.3

dent constraint failures, with a dimension-level depth-of-processing hierarchy governing degradation rates. Second, a comprehension-maintenance dissociation visible at the per-constraint level, where models achieve high partial compliance but fail strict pass criteria. The gap between these measures does not predict compositional vulnerability ($\rho = +0.196$, $p = 0.30$)—the transition would be equally sharp if all constraints had zero comprehension-maintenance gap, because the mechanism depends only on per-constraint CSRs, not on the nature of per-constraint failure.

Deployment implications. The two-regime decay has direct consequences for any system that stacks instructions into a single prompt. A common assumption is that reliability falls off linearly with the number of constraints — that a system handling N requirements will handle (N+3) only slightly less well. The multiplicative decay shows the opposite: per-constraint pass rates compound, so a configuration that is dependable at N constraints can degrade sharply by (N+3), and the drop arrives sooner than a linear extrapolation would predict. The fitted decay factors make this quantitative: each model's r (ranging from 0.75 to 0.94) sets how quickly its compliance drops with k, which lets a practitioner estimate the largest k at which a model still meets a target threshold of reliability. At a 90% target, most models sustain *only one or two simultaneous constraints*, and the strongest, Gemini 3.1 Flash-Lite, reaches roughly three. Applications that genuinely require many requirements to hold at once should plan to decompose the task, verify constraints with external checks, or operate well below the apparent single-constraint ceiling.

Furthermore, how compliance is measured matters just as much as which model is chosen. Continuous, partial-credit metrics *overstate true satisfaction by as much as 61 percentage points near the transition boundary* (Equation (1)), which is the regime where reliability is most fragile. Production monitoring should therefore score each constraint with strict binary verification, so that a system's reported reliability reflects what it actually delivers under compositional load.

Preamble discipline as a dominant factor. A post-hoc analysis reveals that 60–85% of constraint failures co-occur with meta-preamble text in the first 200 characters — e.g., models writing "Constraint Analysis: 1. [N2] requires no digits…" before producing actual content. The two highest-ranked models (Flash-Lite: 7.9% preamble rate; Kimi: 3.7%) almost never produce preambles; the lowest-ranked (Llama 70B: 93.7%; Claude 4.5 Sonnet: 87.3%) almost always do. This does not invalidate the measurement — a model that narrates its constraint-solving process *is* failing at instruction following — but preamble discipline correlates strongly with model ranking, and deserves investigation as a training signal. It also means the benchmark measures a more specific capability than "compositional constraint satisfaction" in the abstract: whether a model can produce clean, constraint-compliant output without self-referential meta-commentary under compositional load.

Scale alone does not determine compositional robustness. In the one comparison where size is known and isolated — Llama 3.1 8B versus 70B — the smaller model matches or exceeds the larger in CSR at high k (24.2 vs. 21.1 at k=8). Among the Gemini models, the lighter Flash-Lite leads the Pro variant by 23.5pp overall, though here scale and training differ together, so the gap reflects more than size. These inversions indicate that compositional robustness is shaped by training choices at least as much as by scale, rather than emerging automatically with it.

Cross-modal synthesis. The divergence between text and visual CSE —universal decay versus model-dependent decay, $\varphi \approx 0.02$ versus 0.89 — requires careful interpretation. Text CSE is constrained *generation*, where the output space is unbounded and each added constraint shrinks the region

of valid responses. Visual CSE, on the other hand, is constrained *selection*, where the candidate set is finite and each constraint narrows it down further. These impose opposite pressures: an added constraint makes a valid generation harder to produce, but a valid selection easier to isolate, so the divergence could reflect task structure, modality, or both. Despite this confound, the CSE framework produces informative measurements in both settings: IRT-calibrated ability rankings ($\rho \geq 0.98$), within-constraint degradation profiles, and the φ matrix that diagnoses the independence structure in each modality. The visual three-way decomposition (solvable/impossible/combined) further shows that compositional decay and impossibility detection are partially independent capabilities: the weakest solver (GPT-5.2, 37.8%) is tied for best impossibility detector (90.2%).

5. Related Work

Prior instruction-following benchmarks established deterministic, programmatic verification, yet evaluate at low or fixed constraint counts, leaving the shape of compositional degradation uncharacterized. IFEval (Zhou et al., 2023) composes at most three constraints ($k \leq 3$); FollowBench (Jiang et al., 2024) reaches $k=5$ and scores its open-ended subset with a GPT-4 judge. COLLIE (Yao et al., 2023) is the closest prior work: it pioneered CFG-based constraint specification and showed that composition is hard, though it tests fixed compositions rather than sweeping k continuously, and reports only binary pass/fail. CSE adds the systematic k-sweep that recovers the decay trajectory itself, together with dual binary/continuous scoring.

Our finding that constraint failures are approximately independent ($\varphi \approx 0.02$) connects to the k-SAT phase-transition literature (Mertens et al., 2006; Mézard & Montanari, 2009); the CSE transition is mechanistically simpler, arising from multiplicative accumulation of independent failures. The same decay recurs across domains: ConceptMix (ConceptMix Authors, 2024) reports a comparable rate in text-to-image generation ($r \approx 0.65$ per concept, versus our $r=0.84$), and CCTU (Ye et al., 2026) finds a 20% ceiling in constrained tool use, close to our 16% text floor — evidence that the transition may be a general property of LLM compositional processing. ATA (Lv et al., 2026) offers a mechanistic account via attention distraction and shows that inference-time steering partially recovers compositional performance. ImpossibleBench (Zhong et al., 2025) tests detection of infeasible constraint combinations in code; CSE includes deliberate impossible probes in both modalities and finds that solving and impossibility detection are partially independent capabilities (Section 3.2). A comprehensive discussion of all related benchmarks, training methods, cross-modal parallels, and theoretical connections appears in Section K.

6. Conclusion

CSE reveals that compositional instruction following degrades via a predictable two-regime pattern: exponential decay followed by an asymptotic floor at $\sim$16%, driven by multiplicative accumulation of approximately independent constraint failures. Reliable performance breaks down beyond 5–6 simultaneous constraints across all 14 models tested. A depth-of-processing hierarchy governs which constraints fail first, and this ordering reflects processing demands rather than intrinsic difficulty. In vision, the pattern diverges: decay is model-dependent and governed by perceptual accuracy, with failures massively correlated rather than independent — revealing that task structure (generation vs. selection) fundamentally shapes the compositional decay profile. At a 90% compliance target, even the best model handles only about three constraints at once, and partial-credit scores can overstate true constraint satisfaction rate by up to 61 points. Systems that depend on a model obeying many rules at once, such as safety guardrails, output schemas, or policy constraints, degrade sharply once more than a few constraints are combined, so each should be verified independently rather than presumed to hold based on a deceitful aggregate compliance score.

Three directions would advance this work: (1) constrained image *generation* with deterministic pixel-level verification, the true visual analog of text CSE; (2) constrained text *selection* to test whether selection tasks universally plateau; and (3) measuring how constraint load degrades primary-task accuracy when models must simultaneously answer questions and satisfy format constraints. Models follow instructions well one at a time; following many at once is where they still break.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning evaluation. The benchmarks reveal limitations of current models—both text-only and vision-language—in compositional constraint satisfaction, with implications for safe deployment in applications where multiple requirements must be satisfied simultaneously. The finding that continuous compliance metrics can overestimate true CSRs by up to 61 percentage points near the transition boundary is directly relevant to production monitoring systems. The visual CSE results reveal a perception-composition tradeoff: VLMs that perceive accurately show no compositional cost, while weaker perceivers decay under constraint load, suggesting that perceptual capability is a prerequisite for compositional robustness in grounded settings. We release all deterministic verifiers, probes, scenes, and evaluation code for reproducible evaluation in both modalities.

References

An, K., Sheng, L., Cui, G., Si, S., Ding, N., Cheng, Y., and Chang, B. ULTRA-IF: Advancing instruction following from the wild. *arXiv preprint arXiv:2502.04153*, 2025.

Balkır, E., Pernthaller, A., Basaldella, M., Hernández-Orallo, J., and Collier, N. Confident rankings with fewer items: Adaptive LLM evaluation with continuous scores. *arXiv preprint arXiv:2601.13885*, 2026.

BYO-Eval Authors. BYO-Eval: Procedurally generated synthetic diagnostics for VLM perception. *arXiv preprint arXiv:2506.05440*, 2025.

Cohen, V., Nangue Tasse, G., Gopalan, N., James, S., Gombolay, M., Mooney, R., and Rosman, B. Compositional instruction following with language models and reinforcement learning. *Transactions on Machine Learning Research*, 2024. arXiv:2501.12539.

ConceptMix Authors. ConceptMix: A compositional image generation benchmark with controllable difficulty. *arXiv preprint arXiv:2408.14339*, 2024.

Dong, G., Lu, K., Li, C., Xia, T., Yu, B., Zhou, C., and Zhou, J. Self-play with execution feedback: Improving instruction-following capabilities of large language models. 2025.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., West, P., Bhagavatula, C., Le Bras, R., Hwang, J. D., Sanyal, S., Welleck, S., Ren, X., Ettinger, A., Harchaoui, Z., and Choi, Y. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36, 2023.

Favero, A., Sclocchi, A., Cagnetta, F., Frossard, P., and Wyart, M. How compositional generalization and creativity improve as diffusion models are trained. In *International Conference on Machine Learning (ICML)*, 2025.

Gan, H., Tinaz, B., Sepehri, M. S., Fabian, Z., and Soltanolkotabi, M. ConceptMix++: Leveling the playing field in text-to-image benchmarking via iterative prompt optimization. *arXiv preprint arXiv:2507.03275*, 2025.

GenAI-Bench Authors. GenAI-Bench: Evaluating and improving compositional text-to-visual generation with VQAScore. *arXiv preprint arXiv:2406.13743*, 2024.

Guo, Z., Liu, W., Xie, M., Xu, J., Huang, Z., Tian, M., Xu, J., Shen, Y., Qian, Q., Wu, M., Wang, X., Lv, C., Wang, H., Yao, H., Zheng, X., and Huang, X. RECAST: Expanding the boundaries of LLMs' complex instruction following with multi-constraint data. *arXiv preprint arXiv:2505.19030*, 2025.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022.

Hsieh, C.-Y. et al. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. *arXiv preprint arXiv:2306.14610*, 2023.

Jiang, Y., Wang, Y., Zeng, X., Zhong, W., Li, L., Mi, F., Shang, L., Jiang, X., Liu, Q., and Wang, W. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Lake, B. M. and Baroni, M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *International Conference on Machine Learning (ICML)*, 2018.

LexInstructEval Authors. LexInstructEval: Formal grammar-based evaluation of lexical instruction following. *arXiv preprint arXiv:2511.17561*, 2025.

Li, B., Corona, R., Mangalam, K., Chen, C., Flaherty, D., Belongie, S., Weinberger, K. Q., Malik, J., Darrell, T., and Klein, D. Does unsupervised grammar induction need pixels? *arXiv preprint arXiv:2212.10564*, 2022.

Li, P., Tang, X., Chen, S., Cheng, Y., Metoyer, R., Hua, T., and Chawla, N. V. Adaptive testing for LLM evaluation: A psychometric alternative to static benchmarks. *arXiv preprint arXiv:2511.04689*, 2025.

Lu, X., West, P., Zellers, R., Le Bras, R., Bhagavatula, C., and Choi, Y. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of NAACL*, 2021.

Lu, X., Welleck, S., West, P., Jiang, L., Kasai, J., Khashabi, D., Le Bras, R., Qin, L., Yu, Y., Zellers, R., Smith, N. A., and Choi, Y. NeuroLogic A*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of NAACL*, 2022.

Lv, F., Qin, Y., Hong, R., Liang, J., Wu, J., Li, K., Sun, X., and Zhang, C. Attend to the active: Structure-aware dynamic attention in LLMs for compositional instruction following. In *International Conference on Learning Representations (ICLR)*, 2026.

Mertens, S., Mézard, M., and Zecchina, R. Threshold values of random K-SAT from the cavity method. *Random Structures & Algorithms*, 28(3):340–373, 2006.

Mézard, M. and Montanari, A. *Information, Physics, and Computation*. Oxford University Press, 2009.

Order Matters Authors. Order matters: Investigating position bias in multi-constraint instruction following. *arXiv preprint arXiv:2502.17204*, 2025.

Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N. A., and Lewis, M. Measuring and narrowing the compositionality gap in language models. *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Pyatkin, V., Malik, S., Graf, V., Ivison, H., Huang, S., Dasigi, P., Lambert, N., and Hajishirzi, H. Generalizing verifiable instruction following. *arXiv preprint arXiv:2507.02833*, 2025.

Qin, Y., Song, K., Hu, Y., Yao, W., Cho, S., Wang, X., Wu, X., Liu, F., Liu, P., and Yu, D. InFoBench: Evaluating instruction following ability in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Sakai, Y., Kamigaito, H., and Watanabe, T. Revisiting compositional generalization capability of large language models considering instruction following ability. *arXiv preprint arXiv:2506.15629*, 2025.

Sun, H., Liu, L., Li, J., Wang, F., Dong, B., Lin, R., and Huang, R. Conifer: Improving complex constrained instruction-following ability of large language models. *arXiv preprint arXiv:2404.02823*, 2024.

T2I-CompBench Authors. T2I-CompBench++: An enhanced open-world compositional T2I generation benchmark. *arXiv preprint arXiv:2307.06350*, 2023.

T2V-CompBench Authors. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. *arXiv preprint arXiv:2407.14505*, 2024.

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of CVPR*, 2022.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Wen, B., Ke, P., Gu, X., Wu, L., Huang, H., Zhou, J., Li, W., Hu, B., Gao, W., Xu, J., et al. ComplexBench: Benchmarking complex instruction following with multi-constraint composition. *Advances in Neural Information Processing Systems*, 2024.

Yao, S., Peng, H., Guo, Q., Chen, J., Li, Y., Liang, J., Wu, Z., et al. COLLIE: Systematic construction of constrained text generation tasks. In *Proceedings of ICLR*, 2023.

Ye, J., Huang, C., Chen, Z., Fu, W., Yang, C., Yang, L., Wu, Y., Wang, P., Zhou, M., Yang, X., Gui, T., Zhang, Q., Shi, Z., Fan, J., and Huang, X. A multi-dimensional constraint framework for evaluating and improving instruction following in large language models. *arXiv preprint arXiv:2505.07591*, 2025.

Ye, J., Zhang, G., Fu, W., Gui, T., Zhang, Q., and Huang, X. CCTU: A benchmark for tool use under complex constraints. *arXiv preprint arXiv:2603.15309*, 2026.

Zhong, Z., Raghunathan, A., and Carlini, N. ImpossibleBench: Measuring LLMs' propensity of exploiting test cases. *arXiv preprint arXiv:2510.20270*, 2025.

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L. Instruction-following evaluation for large language models. In *Advances in Neural Information Processing Systems*, 2023.

A. Per-Constraint CSR Heatmaps

Per-constraint CSR heatmaps for every model are shown in Figure 3 (text CSE-30, 30 constraints × 14 models) and Figure 4 (visual CSE-18, 18 constraints × 10 VLMs).

B. Per-Constraint Performance

Table 6 and Table 7 report the per-constraint satisfaction rate and comprehension-maintenance gap for all 30 text and all 18 visual constraints, respectively.

Table 6. Text CSE-30: per-constraint CSR (%) and gap (Δ, in pp) across 14 models. **Bold** marks highest CSR and largest gap per dimension.

Dimension	Constraint	CSR	Δ
Structural	S1 Word count	3.0	**+46**
	S2 Sentence count	**23**	+30
	S3 Paragraph structure	11	+31
	S4 Character count	1.1	+33
	S5 Line count	18	+21
Lexical	L1 Lipogram	38	**+61**
	L2 Mandatory words	60	+5
	L3 Sentence-initial	14	+21
	L4 Acrostic	12	+16
	L5 Forbidden word	**85**	+15
	L7 Vowel ratio	67	+6
Word-level	W1 No repetition	14	+61
	W2 Length range	23	**+70**
	W3 Minimum sentence words	**46**	+33
Relational	R1 Scene graph	**39**	+3
	R2 Transitive order	24	**+8**
	R3 Logic grid	23	+7
Format	F1 Valid JSON	**51**	+0
	F2 Markdown table	38	+4
	F3 Bullet list	38	**+15**
	F4 Code blocks	48	+6
Meta	M1 Hidden message	6.2	+14
	M2 Self-count	1.1	+31
	M3 Divisibility	**12**	**+44**
Numerical	N1 Numeric embedding	14	+2
	N2 No digits	**45**	**+54**
	N3 Distinct numbers	16	+29
Ordering	O1 Monotonic length	11	**+39**
	O2 Alphabetical sentences	**23**	+19
	O3 Alternating length	18	+39

C. Per-Constraint CSR by Model

Table 8 reports per-constraint CSR for each individual model on the ten most difficult constraints, with the excluded DeepSeek V3 results in the final column for completeness.

Table 7. Visual CSE-18: per-constraint CSR (%) and gap (Δ, in pp) across 10 VLMs. Gaps are uniformly small (≤ 4.4pp). **Bold** marks highest CSR per dimension.

Dimension	Constraint	CSR	Δ
Attribute	A1 Color	61.8	+1.2
	A2 Shape	57.4	+2.4
	A3 Size	**62.8**	+2.3
Spatial	S1 Left/right	61.1	+3.0
	S2 Above/below	61.8	+2.5
	S3 Adjacency	53.6	+1.6
	S4 Isolation	54.1	+2.8
	E1 Extremal	**71.0**	+0.6
Relational	R1 Same/different	**60.1**	+2.2
	R2 Comparative size	59.7	**+4.4**
Set/Logical	L1 Threshold	52.6	+2.3
	L2 Disjunction	57.9	+2.4
	L3 Negation	**61.9**	+1.9
	L4 Uniqueness	58.2	+0.0
	L5 Attribute count	35.5	**+4.0**
Counting	C1 Count filter	49.8	+2.2
	C2 Group parity	**53.7**	+3.2
	C3 Plurality	49.8	**+3.3**

D. Decay Model Predictions vs. Actuals

Table 9 compares the multiplicative and floor decay models' predicted CSRs against the observed values at each k, with the held-out k=9−12 rows marked TEST.

E. Failure Mode Details

Table 10 and Table 11 give the per-model breakdown of L1 failures into refusal/meta/genuine categories and, for genuine failures only, the position-of-violation strategy profiles.

Classifier design. Accurately separating meta-preamble from genuine failures required iterative refinement. An initial narrow classifier (keying on explicit preamble markers) identifies only 8 meta-preamble cases. Manual inspection reveals contamination: 4 of 5 top-ranked "genuine" failures are constraint self-narration ("Constraint L1 requires no letter 'k' "). We expand the classifier with a constraint-echo regex matching patterns such as "no letter 'X' ", "avoiding the letter", and "Constraint [L1]" within 60 characters of the violation, reclassifying matches in the first 150 characters as meta-preamble. This moves 106 cases from genuine to meta-preamble (8 → 114), primarily from Claude models (+50, +22, +23 for the three variants). The 150-character threshold prevents reclassifying violations deep in the response where constraint references may be legitimate.

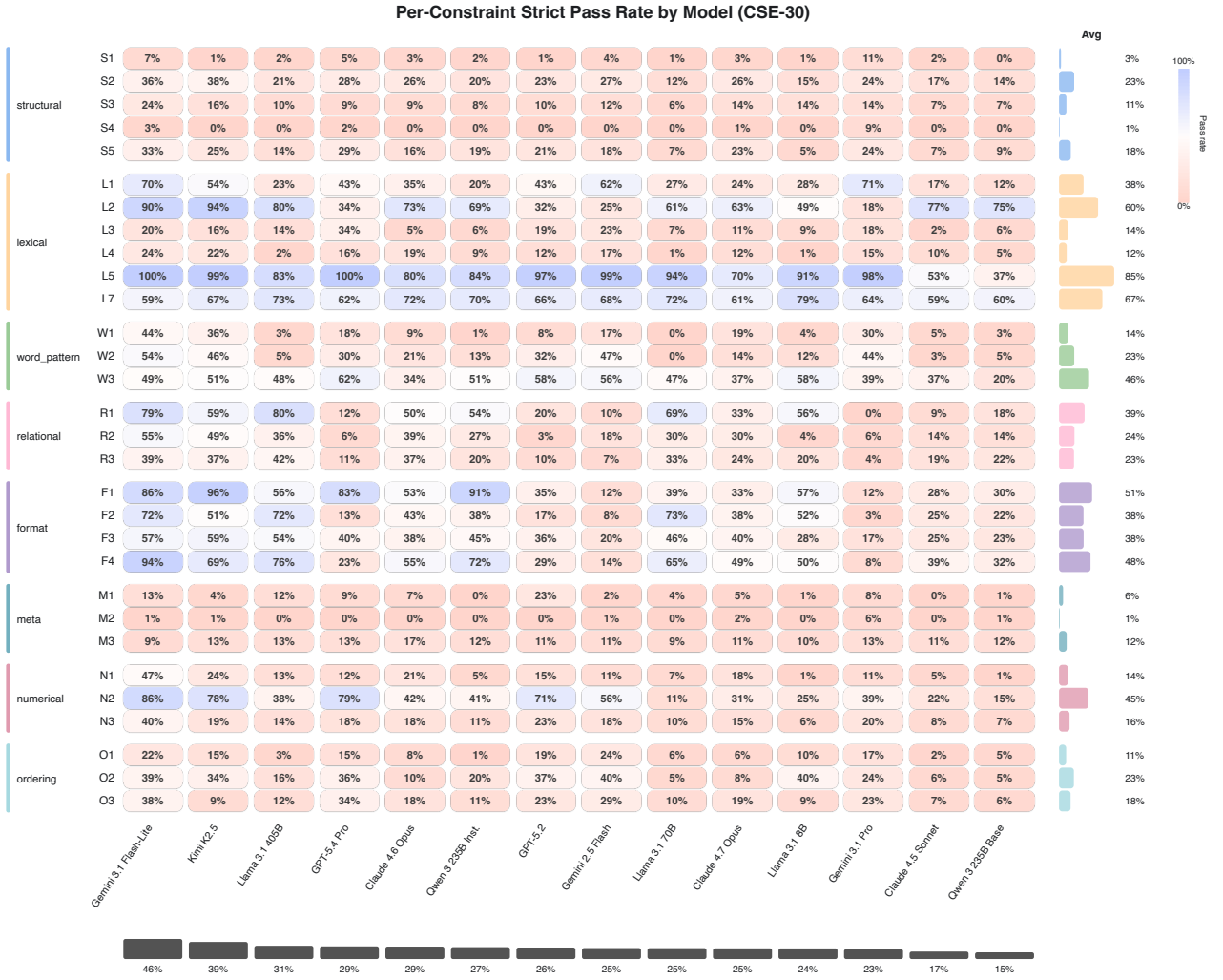


Figure 3. Per-constraint CSR by model (text CSE-30, 30 constraints $\times$ 14 models). Gemini 3.1 Flash-Lite dominates across all dimensions (46.2% overall), while model rankings invert at the constraint level: GPT-5.4 Pro leads at k=1 (82%) but Flash-Lite overtakes by $k \geq 2$. The format–relational cluster is visible as a correlated block (F1–F4, R1–R3): models that fail structured output (JSON, tables) fail relational reasoning simultaneously—output-structure dependency, not cognitive coupling (Section 3.1). Structural constraints (S1–S5) are uniformly hard ($\leq$23%), while lexical constraints (L5 forbidden word: 85%, L7 vowel ratio: 67%) sustain the asymptotic floor. The Qwen 3 Base column shows the cost of no instruction tuning: 16% overall, with near-zero rates on format and relational constraints.

F. Full Constraint Specifications

Table 12 and its continuation give the complete specification — name, exact instruction text injected into prompts, and an example output — for all 30 text constraints.

G. Constraint Validation Protocol

All 30 constraints underwent a two-phase pilot validation before the full evaluation run.

Phase 1 — Parameter calibration (390 probes): Each of the 13 new constraints (added beyond the original 17) is tested at k=1 with 30 probes on Llama 3.1 8B. Constraints with k=1 CSR $< 5\%$ (floor) or $> 95\%$ (ceiling) are flagged for parameter adjustment. Two fixes are applied: W2 (word length range) is widened from $\min = 3$ to $\min = 1$ to allow common single-character words; L7 (vowel ratio) is widened from (0.7, 0.9) to (0.5, 0.9) to include natural English ratios. All other constraints fall within the target range.

Phase 2 — End-to-end pipeline verification (144 probes): 6 probes per k level (k=1−12) are run on Llama 70B and Gemini 2.5 Flash to verify API access, response parsing, and verifier correctness. The phase transition pattern is visible even with 12 observations per k point, confirming that the effect size is large enough to detect with modest sample sizes.

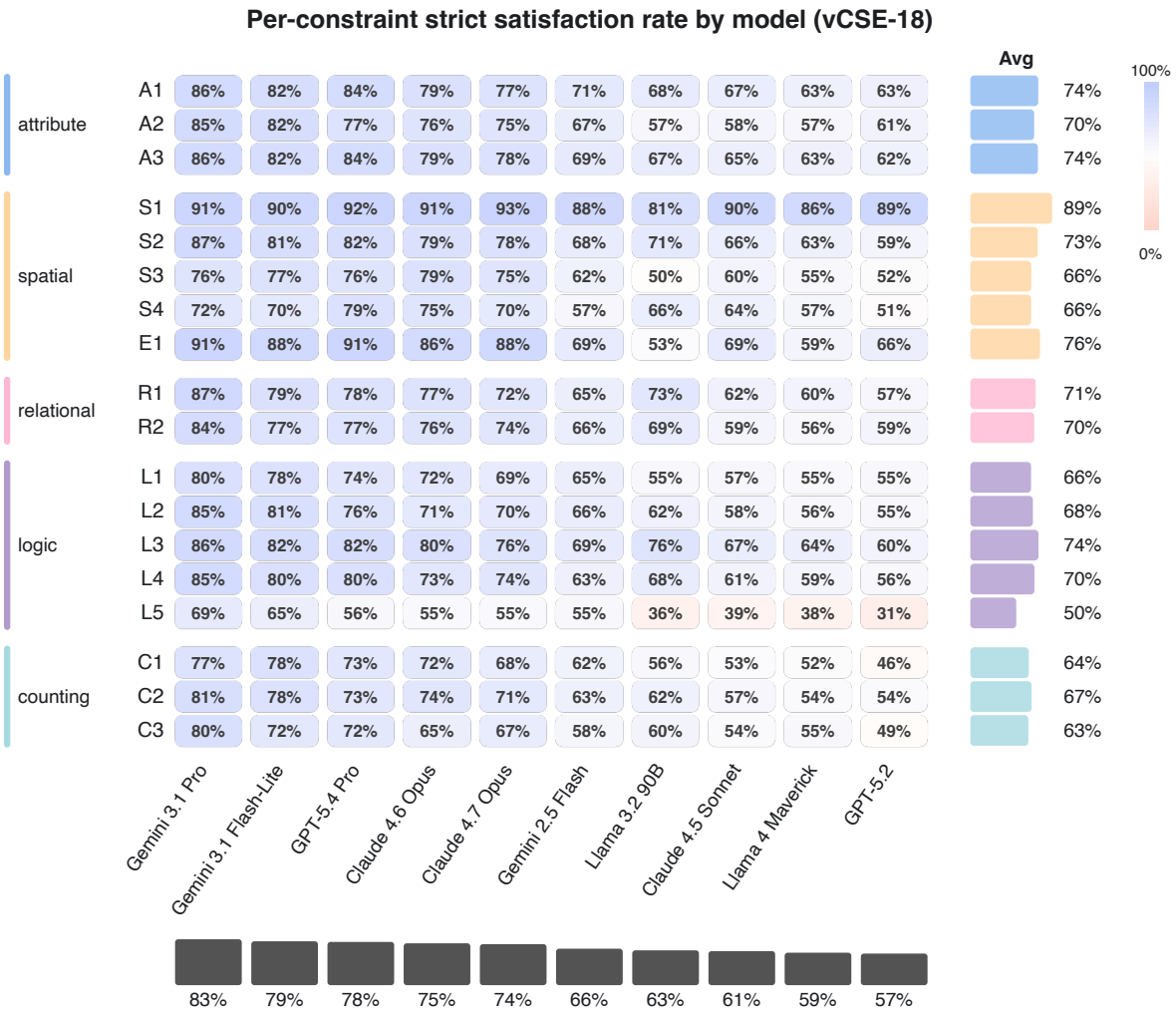


Figure 4. Per-constraint CSR by model (visual CSE-18, 18 constraints × 10 VLMs, corrected for reclassified impossible probes). Gemini 3.1 Pro Preview leads (83.5%), with a 25.1pp spread to GPT-5.2 (58.4%). GPT-5.2 is the weakest visual solver (37.8% on solvable probes) despite being tied for best impossibility detector (90.2%). L5 (attribute count comparison) is the hardest constraint across all models (50.0% corrected), requiring two scene-wide counts and a comparison. E1 (extremal position, 75.9%) is the easiest—a hard-selector that narrows the answer to one object. Unlike the text heatmap (Figure 3), no format–relational cluster is visible: the co-failure structure is monolithic (φ=0.89), with all constraints loading uniformly on a single perception factor (Section 3.2).

H. IRT Validation

A 2PL Item Response Theory model (Li et al., 2025; Balkır et al., 2026) fitted on 355 informative probes (excluding 839 all-pass or all-fail probes) yields ability estimates that strongly correlate with raw CSRs (Spearman $\rho = 0.981$, $p < 10^{-9}$), validating that raw CSRs are a sufficient metric without requiring adaptive infrastructure. The 70% of probes uninformative for IRT model ranking are not wasted: they define the endpoints of the phase transition curve and power constraint-level analyses operating on all 82,628 checks. The main text refers to this appendix for the IRT analysis (Section 4).

IRT rankings diverge meaningfully from per-constraint CSRs for three models: Kimi K2.5 ranks 2nd by per-constraint CSR but 8th by IRT ability ($\Delta = -6$), while Claude 4.7 Opus rises from 9th to 3rd ($\Delta = +6$) and Gemini 3.1 Pro from 12th to 2nd ($\Delta = +10$). This divergence reflects a *breadth vs. precision* distinction: IRT uses strict all-pass scoring (all constraints in a probe must pass), while per-constraint rates average across individual constraints. Kimi passes many individual constraints but fewer complete probes—it is broadly competent but less precise. Claude 4.7 and Gemini Pro pass fewer constraints overall but succeed on complete probes more often, indicating higher compositional precision. This contrast suggests that per-constraint and probe-level metrics capture qualitatively different aspects of instruction-following capability.

Table 8. Per-constraint CSR (%) for each model (top 10 constraints shown), sorted by overall constraint difficulty. **Bold** marks the best model for each constraint. The rightmost column (V3) is DeepSeek V3, which is excluded from main analyses due to data loss, but is included here for completeness.

	Fl-Lite	*405B*	*Opus*	*GPT-5.4*	*Qwen-I*	*GPT-5.2*	*70B*	*Fl-2.5*	*8B*	*Pro*	*Sonnet*	*Qwen-B*	*V3*
L5	**88**	84	83	86	77	83	75	82	70	76	73	65	82
L7	**81**	71	69	71	72	67	65	70	63	60	57	52	62
L2	**76**	68	63	62	62	57	56	60	54	55	48	50	45
F1	**77**	65	47	38	61	29	56	14	46	10	30	27	9
F4	**65**	53	51	48	46	44	42	41	38	37	31	27	14
W3	**66**	50	46	49	52	42	43	49	40	38	36	32	21
N2	**64**	49	42	46	41	39	34	45	31	36	34	30	7
R1	**58**	52	42	9	33	10	43	12	25	4	14	17	5
L1	**57**	44	37	34	36	37	35	33	31	30	28	24	8
F2	**63**	44	35	38	42	31	30	25	25	20	22	15	10

Table 9. Aggregate decay model predictions vs. actual CSRs at each k. Rows marked TEST are held-out (k=9−12). The pure multiplicative model tracks training data well (train MAE = 1.8pp) but diverges at $k \geq 9$, underpredicting by 5–10pp. The floor model ($c = 15.1\%$) matches held-out actuals within 1–2pp. **Bold** marks the model with lower absolute error at each held-out k.

k	Mult	Floor	Actual	Mult err	Floor err	Set
1	56.0	58.2	58.4	−2.4	**−0.2**	Train
2	46.9	46.0	45.4	+1.5	**+0.7**	Train
3	39.3	37.3	36.9	+2.3	**+0.4**	Train
4	32.9	31.0	31.9	**+0.9**	−0.9	Train
5	27.5	26.5	28.5	**−1.0**	−2.0	Train
6	23.0	23.3	21.9	+1.1	**+1.4**	Train
7	19.3	21.0	20.4	−1.1	**+0.6**	Train
8	16.1	19.3	18.2	−2.0	**+1.1**	Train
9	13.5	18.1	19.0	−5.5	**−0.9**	Test
10	11.3	17.2	16.2	−4.9	**+1.0**	Test
11	9.5	16.6	15.5	−6.0	**+1.1**	Test
12	7.9	16.2	18.4	−10.5	**−2.2**	Test

Caveats. M=14 models is below classical IRT recommendations ($N \geq 200$ for stable 2PL estimation). The optimization did not fully converge, with extreme θ values for boundary models. We use IRT as a validation check (which it passes: high discrimination, rank preservation) rather than as a primary analysis.

I. Visual CSE: Constraint Specifications

Table 14 and its continuation give the complete specification — name, instruction text, and deterministic verifier — for all 18 visual constraints.

J. Visual CSE: Validation Protocol

The visual CSE benchmark underwent four validation stages prior to the full evaluation run. Each stage addressed a specific class of potential error; the final probe set reflects all corrections.

Table 10. L1 (lipogram) failure categorization by model (n = 1,377 total failures after excluding Kimi K2.5 and Claude 4.7 Opus, which have separate audits). "Genuine" = model attempts the constraint and violates it in content; "Refusal" = model identifies infeasibility and the forbidden letter appears in the explanation text; "Meta" = constraint self-narration (restating the rule using the forbidden letter). The refusal rate correlates with RLHF-style safety training intensity: base models (Qwen Base: 12%) rarely refuse; heavily RLHF'd models (GPT: 89–97%) almost always do. The expanded classifier described below reclassifies 110 additional cases from Genuine to Meta in this 12-model subset, primarily from Claude models; the main text (Section 3.1) reports post-expansion totals across all 14 models.

Model	**Genuine**	**Refusal**	**Meta**	**Total**
Claude 4.5 Sonnet	126	10	0	136
Claude 4.6 Opus	104	3	0	107
GPT-5.2	10	**84**	0	94
GPT-5.4 Pro	3	**90**	0	93
Gemini 2.5 Flash	37	25	1	63
Gemini 3.1 Flash-Lite	44	6	0	50
Gemini 3.1 Pro	15	33	0	48
Llama 3.1 405B	81	43	2	126
Llama 3.1 70B	37	82	0	119
Llama 3.1 8B	42	75	1	118
Qwen 3 235B Base	**127**	18	0	145
Qwen 3 235B Inst.	56	76	0	132
Total	**828**	**545**	**4**	**1,377**

Stage 1 — Coordinate system verification (50 scenes, 1,930 pairwise comparisons). The camera projects the 7×7 grid onto screen space with an elevated rear perspective (cam_pos = $[3.5, 7.5, 11.0]$). Grid-x is inverted under this projection: higher grid-x → lower screen-x (visually left). Grid-y is not inverted: lower grid-y → visually above. We verify these mappings empirically across 50 rendered scenes by comparing grid coordinates to screen-space centroids, confirming 100% inversion rate for x (1,930/1,930 pairs)

Table 11. Model failure strategy profiles for genuine L1 failures only ($n = 828$). Median = normalized position of first violation (0 = response start, 1 = end). Early% = fraction of violations in the first 20% of the response. Models partition into *immediate* (median < 0.10) and *gradual* (median $0.15-0.54$) strategies. GPT-5.2 and GPT-5.4 are excluded (3–10 genuine failures each, insufficient for profiling).

Model	Med	Early%	Late%	σ	Strategy
Qwen 3 Base	0.03	79	9	0.24	Immediate
Claude 4.6 Opus	0.07	70	15	0.27	Immediate
Claude 4.5 Sonnet	0.09	79	3	0.16	Immediate
Llama 3.1 8B	0.15	52	7	0.23	Gradual
Llama 3.1 70B	0.16	59	19	0.30	Gradual
Llama 3.1 405B	0.17	51	16	0.27	Gradual
Gemini Flash-Lite	0.43	20	30	0.28	Gradual
Gemini 2.5 Flash	0.46	16	32	0.26	Gradual
Gemini 3.1 Pro	**0.54**	20	**40**	0.29	Gradual

and 0% for y. All spatial verifiers (S1, S2, S3, S4, E1) use grid coordinates only; no verifier references screen-space.

Stage 2 — Verifier–signature consistency audit (18 constraints). Every verifier–signature pair is traced end-to-end to confirm that: (a) the signature assigns a constraint only when the target object satisfies it, (b) the verifier checks the same condition on the model's extracted IDs, and (c) the instruction text from `generate()` matches the verifier's semantics. This audit identifies and resolves a critical bug: the `assemble_prompt` function applies a left↔right text swap to S1 and leftmost↔rightmost to E1 at inference time, double-inverting labels that were already visually correct from the signature. Fix: removed the text swap; set `GRID_X_INVERTED=True` so the signature carries visual semantics end-to-end. All 18 verifier–signature pairs are confirmed consistent after the fix.

Stage 3 — Three-phase pilot (112 API calls). Phase 1 tested each of the 18 constraints at k=1 (2 probes each, 36 total) on Gemini 3.1 Flash-Lite, flagging any constraint at $< 5\%$ (floor) or $> 95\%$ (ceiling). Two constraints flagged as floor (S1 left/right at 0%, R2 comparative size at 0%) are diagnosed as genuine model difficulty (answer-set over-extraction), not verifier bugs—confirmed by tracing the verifier output on specific probes. Phase 2 tested mixed-k probes (k=1−8, 16 total) to verify the extraction pipeline and k-curve shape. Phase 3 ran 20 random probes on 3 models (GPT-5.2, Claude 4.5, Flash-Lite) to verify cross-model consistency. All 112 responses hit the `<answer>` tag extraction tier (100% tag compliance).

Stage 4 — Post-hoc satisfiability audit (1,463 solvable probes). After the full run, a binding check re-verified all solvable probes against the current verifier logic. 349 probes (23.9%) are identified as secretly impossible—100% contained S1, indicating a spatial-constraint composition artifact where the scene-first satisfiability guarantee failed for a specific interaction between S1's directional filter and other constraints at high k. Contamination scaled with k: 0% at k=1 to 49.5% at k=8. These 349 probes are reclassified as impossible; the solvable-only analysis (Section 3.2) excludes them. The remaining 1,114 solvable probes pass the binding check: for each, at least one object in the scene satisfies all k constraints simultaneously.

K. Related Work (Full)

This appendix provides the complete related work discussion summarized in Section 5.

K.1. Compositional Instruction Following

Evaluation benchmarks. IFEval (Zhou et al., 2023) established the paradigm of verifiable instruction following with 25 constraint types and deterministic programmatic checks, but probes typically contain $k \leq 3$ constraints—insufficient to observe compositional degradation. A model scoring 80%+ on IFEval could in principle collapse at k=5−8 on CSE, exposing degradation invisible to proficiency-oriented evaluation. InfoBench (Qin et al., 2024) decomposes each instruction into atomic verifiable criteria and evaluates per-criterion satisfaction via its Decomposed Requirements Following Rate (DRFR). GPT-4 achieves ∼89% DRFR overall, yet more than 10% of requirements remain unfulfilled even for this strongest model, and no model fully satisfies number or linguistic constraints. Performance is highest on content and style, intermediate on format, and lowest on number and linguistic dimensions—a constraint-type hierarchy that foreshadows CSE's depth-of-processing finding, where relational and word-level constraints degrade $2.4\times$ faster than lexical ones. InfoBench operates at the single-instruction level; CSE extends this analysis systematically across k=1 to 12, revealing how per-criterion gaps compound into the phase transition. FollowBench (Jiang et al., 2024) scales to k=5 and introduces the Consistent Satisfaction Level (CSL) metric for tracking partial compliance across difficulty levels, but relies on GPT-4 as a judge for several constraint types, introducing evaluation variance. CSL measures which satisfaction level a model can sustain, which is conceptually adjacent to CSE's k-curve—but CSE isolates where the transition occurs rather than tracking cumulative levels, and uses fully deterministic verification throughout. COLLIE (Yao et al., 2023) is the closest prior work: it pioneered grammar-based constraint specification with formal CFG definitions and corpus-grounded extraction ensuring natural solutions exist, and demonstrated that compositional constraints are harder than individual ones—a finding we confirm and extend.

Three key differences: (1) COLLIE uses fixed constraint compositions while CSE systematically varies k from 1–12, enabling the decay characterization that is this paper's central contribution; (2) COLLIE reports that composition is hard but does not quantify *where* capability breaks or what functional form the degradation takes; (3) CSE's dual scoring (binary + continuous) reveals gaps between partial compliance and strict satisfaction that are absent from COLLIE's binary-only evaluation. COLLIE is stronger in formal rigor (CFG grammar) and in its pass@k evaluation design; CSE contributes the systematic scaling analysis and the mechanistic decomposition of why composition fails.

Complex constraint benchmarks. RECAST (Guo et al., 2025) pushes constraint counts to k=13+ with 19 constraint types and proposes RLVC training with constraint-specific reward signals, but does not report per-k performance trajectories, precluding decay analysis. ComplexBench (Wen et al., 2024) models constraint composition *structure*—And, Chain, Selection, Nesting—rather than count, and is thus orthogonal to our analysis of how many constraints can be jointly satisfied. IFBENCH (Pyatkin et al., 2025) demonstrates that models overfit to IFEval's 25 known constraint types and proposes IF-RLVR training for generalization—a different axis (unseen constraints) from our focus (composition scaling). CCTU (Ye et al., 2026) evaluates constrained tool use, finding that no model exceeds 20% task completion under strict multi-constraint adherence—a ceiling remarkably close to CSE's text-only floor of $\sim$16% at $k\geq$9, despite a completely different task domain, suggesting the phase transition may reflect a general property of LLM compositional processing.

Training methods. Conifer (Sun et al., 2024) generates multi-level constraints with GPT-4 and employs curriculum tuning, while AutoIF (Dong et al., 2025) implements self-dialogue with execution-based verification—both focus on training rather than diagnostic evaluation. MulDimIF (Ye et al., 2025) explores how constraint *presentation mode* (example, listing, incorporation) affects following and proposes GRPO training; LexInstructEval (LexInstructEval Authors, 2025) introduces a formal grammar for lexical constraints with bilingual evaluation, finding that procedural depth degrades performance—complementary to our finding that constraint *count* does so. UltraIF (An et al., 2025) scales to 100K+ instructions via automatic constraint decomposition. Neurologic Decoding (Lu et al., 2021; 2022) approaches constraint satisfaction from the inference side via lexically-constrained beam search—complementary to our diagnostic focus.

Mechanistic explanations. ATA (Lv et al., 2026) identifies a mechanistic explanation for compositional degradation: models diffusely attend across all sub-tasks in the input, and inactive sub-tasks cause interference via attention distraction—corroborating our phase transition with performance drops from 60% to 48% as sub-task count increases. Their inference-time attention steering recovers 6–10pp without retraining, demonstrating that the transition is partially addressable through architectural intervention. In reinforcement learning, CERLLA (Cohen et al., 2024) demonstrates that compositional value function representations with Boolean task algebra (AND, OR, NOT) enable sample-efficient learning of 162 language-conditioned tasks. CSE's constraint composition is purely conjunctive (all k constraints must hold simultaneously), but the shared insight is that compositional structure—whether in policy space or constraint space—determines scalability.

K.2. Constraint Scaling and Cross-Modal Parallels

Interaction vs. accumulation. A central question in compositional evaluation is whether difficulty arises from constraint *interactions* (specific pairs that conflict) or from *accumulation* (each additional constraint independently reducing success probability). ComplexBench (Wen et al., 2024) models structural interactions through nesting and branching but does not measure pairwise statistical independence. Order Matters Authors (2025) demonstrate that constraint ordering affects performance by up to 25%, establishing that presentation matters—we find this effect is orthogonal to constraint count (Section 3.1). Ordered CommonGen (Sakai et al., 2025) evaluates whether LLMs can generate sentences with concepts in a specified order across 36 models, finding that even the best model follows the specified order only 75% of the time on a single ordering constraint. CSE's ordering constraints show even lower k=1 rates (O1 monotonic length: 10.9%, O2 alphabetical: 22.8%), and these per-constraint imperfections are precisely what compound multiplicatively to produce the phase transition.

Phase transitions and scaling laws. Our finding that constraints fail approximately independently connects to the theoretical literature on phase transitions in constraint satisfaction. Random k-SAT exhibits a sharp satisfiability threshold driven by clause density and interaction structure (Mertens et al., 2006; Mézard & Montanari, 2009). Our transition is mechanistically simpler: it arises from multiplicative accumulation of independent failure probabilities without requiring interaction structure. This parallels neural scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022), which established that simple parametric models (power laws in compute and data) predict performance with practical accuracy. Favero et al. (2025) provide theoretical support for hierarchical compositional learning: diffusion models learn grammar rules level-by-level, with sample complexity $P_\ell \propto m^{\ell+1}$ for level-ℓ rules—higher-level (more global) rules require polynomially more data. This predicts that con-

straints requiring deeper compositional processing should be more fragile, consistent with our finding that relational constraints (requiring multi-step reasoning) degrade 2.4× faster than lexical constraints (single-step pattern matching) under compositional load (Section 3.1). Our floor model $P(k) = a \cdot e^{-\beta k} + c$ serves an analogous role for constraint composition: three parameters, fitted on k=1−8, predict held-out k=9−12 within 1.4pp—comparable predictive precision from a comparably simple functional form. The "emergent abilities" literature (Wei et al., 2022) documents sharp capability transitions as a function of model scale; we find analogous transitions as a function of compositional load.

Cross-modal compositional decay. The decay pattern we observe in text is consistent with cross-modal findings in generation tasks. ConceptMix (ConceptMix Authors, 2024) reports that text-to-image models drop from 83% at k=1 concepts to 8% at k=7—implying $r \approx 0.65$ per added concept, steeper than text CSE's r=0.84 but following the same multiplicative functional form. ConceptMix++ (Gan et al., 2025) shows that prompt optimization can recover up to 20% of lost compositional capability, suggesting the limitation is partially in prompt understanding rather than solely in generation. T2I-CompBench++ (T2I-CompBench Authors, 2023) documents degradation across attribute binding, spatial relationships, and generative numeracy. T2V-CompBench (T2V-CompBench Authors, 2024) shows temporal composition failures in video generation. All of these are *generation* tasks, consistent with our finding that constrained generation produces decay while constrained selection (visual CSE) produces a different pattern (Section 3.2). GenAI-Bench (GenAI-Bench Authors, 2024) evaluates compositional text-to-visual generation, reporting degradation with compositional complexity in image, video, and 3D generation—further evidence that the decay pattern is specific to *generation* tasks across modalities, not to any single modality.

Visual compositionality benchmarks. Our visual CSE uses CLEVR-style (Johnson et al., 2017) procedural scenes with deterministic scene-graph verification. CLEVR demonstrated that controlled synthetic worlds enable exact evaluation of visual reasoning—the same principle underlies our design, but CLEVR asks compositional *questions* (VQA) while we vary the number of simultaneous *constraints* (CSE). Winoground (Thrush et al., 2022) and SugarCrepe (Hsieh et al., 2023) test visio-linguistic compositionality through contrastive image-text pairs. SugarCrepe specifically addresses benchmark hackability—models exploiting spurious correlations rather than genuine compositionality—a concern that CSE's deterministic verification against scene graphs eliminates by construction. Both find that strong VLMs struggle with compositional understanding, consistent with our finding that visual constraint satisfaction is substantially harder than text at baseline (best VLM: 79.3% vs. text median ∼80% at k=1). A related asymmetry appears in grammar induction: Li et al. (2022) show that LLM embeddings alone outperform multimodal models on unsupervised constituency parsing, suggesting that LLMs internalize enough compositional structure from text to make visual grounding redundant for linguistic tasks. This may partially explain why text CSE's constraints—which map onto syntactic and structural properties already captured in the model's internalized grammar—produce near-independent failures ($\varphi \approx 0.02$), while visual constraints, which require perceptual verification against an external scene with no analogous internalized grammar, produce correlated failures ($\varphi \approx 0.89$). BYO-Eval (BYO-Eval Authors, 2025) demonstrates that procedurally generated synthetic stimuli can isolate specific skills with high synthetic-to-real correlation ($r > 0.99$ for 6/8 models)—methodologically parallel to our procedural probe generation in both modalities.

L. Co-Failure Analysis: Additional Detail

Within- vs. across-dimension. Within-dimension co-failure (mean $|\varphi| = 0.035$) exceeds across-dimension (0.018), but this difference does not reach significance (Mann-Whitney $p = 0.13$) and is entirely driven by the format and relational clusters. Structural–structural ($\bar{\varphi} = 0.010$), lexical–lexical ($\bar{\varphi} = 0.003$), and all other within-dimension pairs are at noise level. Constraints within the same cognitive dimension do not interfere with each other more than constraints across dimensions, outside the shared structured-output dependency.

No shortcut through constraint selection. The co-failure analysis contains no synergistic pairs ($\varphi < -0.05$, except four marginal cases, the strongest at -0.077). The most interpretable is L1–L2 ($\varphi = -0.063$): lipogram and mandatory words are mildly antagonistic when the required words contain the banned letter—genuine interference, but weak. If "compatible" constraint pairs existed—combinations that are easier together than independently—they would appear as negative φ values. They do not.

M. Degradation Hierarchy: Additional Detail

Four constraints have slopes > -0.02 and half-lives > 12—essentially flat across the entire k range: L1 (lipogram), L5 (forbidden word), L7 (vowel ratio), and M3 (divisibility). These "survivors" form the asymptotic floor. Two—M3 (123% retention at k=12) and L1 (91% retention)—exhibit *paradoxical improvement* under compositional load, both benefiting from the longer outputs that compositional prompts tend to elicit: more words increase the chance of hitting a count divisible by 3, and more text dilutes any

single early lipogram violation.

Two control correlations confirm that the hierarchy is not an artifact of difficulty. Degradation rate is uncorrelated with intrinsic difficulty ($\rho(k{=}1 \text{ rate}, \text{slope}) = -0.030$, $p = 0.88$). The comprehension-maintenance gap Δ (Equation (1)) also does not predict compositional vulnerability ($\rho = +0.196$, $p = 0.30$): L1 (lipogram) has the largest gap (Δ=61pp; 98.7% score vs. 37.7% CSR) yet degrades slowly. The phase transition depends on CSR alone, not on how models fail.

A caveat: the fast degradation of relational constraints may be partially inflated by their co-failure coupling with format constraints ($\bar{\varphi} = 0.116$, Section 3.1). A probe containing both a relational and a format constraint fails both when structured output fails, inflating the apparent relational degradation rate beyond its independent vulnerability. The depth-of-processing interpretation—that relational constraints require more sustained computation—is one explanation consistent with the data, but shared structured-output dependency is an alternative that the current analysis cannot rule out.

N. Failure Mode Decomposition (Full Analysis)

This appendix provides the complete failure mode decomposition summarized in Section 3.1.

The comprehension-maintenance gap for L1 (lipogram) is among the most striking in the benchmark: 37.7% CSR but 98.7% average score (+61.0pp gap). This appears to suggest near-perfect compliance with occasional slips—a "resource depletion" narrative. We test this directly through position-of-first-violation analysis on all $n = 1{,}420$ L1 failures across 14 models.

Three failure modes. L1 failures decompose into refusals (39%, $n = 557$), meta-preamble violations (8%, $n = 114$), and genuine content failures (53%, $n = 749$). Refusals are concentrated in GPT models (89–97% refusal rate). Forbidden letter frequency modulates the split: 73% of failures on 'e' (most common in English) are refusals, while only 35–41% on rare letters 'j' and 'k' are refusals—models rationally calibrate to letter difficulty.

Refusal as a family-specific strategy. The refusal decision is not simply "smarter models know when to quit." Llama scaling *inverts*: 8B refuses 64%, 70B refuses 69%, but 405B refuses only 34%—larger models attempt more, opposite to GPT where both sizes refuse at 89–97%. The Qwen 3 Base/Instruct pair provides a controlled comparison: Base refuses 12%, Instruct refuses 58%. Instruction tuning nearly $5\times$ the refusal rate, demonstrating that RLHF training teaches models *when to refuse*, not just how to follow instructions.

The comprehension paradox. Meta-preamble failures ($n = 114$, 8% of all failures) reveal models restating the forbidden-letter rule in their own words (e.g., "ensuring no letter 'v' appears in my response"), using the forbidden letter in the restatement. The model has a correct internal representation of the constraint but does not apply it to its own meta-commentary. This failure mode is overwhelmingly Claude-specific: 99/114 cases (87%) come from Claude models. No Kimi or Qwen Base failures are meta-preamble.

Predominantly engagement failure. Among genuine content failures ($n = 749$), violations cluster *early*: median normalized position $= 0.177$, with 52% in the first 20% of the response and only 7% in the last 20%. This pattern is stable across $k{=}1{-}12$—adding more constraints does not change *where* in the response the violation occurs, only *whether* the model engages the constraint at all. The high average score (98.7%) is not evidence of near-perfect compliance followed by a slip: it reflects the low baseline frequency of individual letters in English. A response containing 'e' once in the first sentence still scores $\sim$96% because 96% of characters are not 'e'. Models fail to *engage* the constraint from onset, not to *maintain* it through generation. Normalized position correlates with response length ($\rho = -0.49$), but this reflects a stable *absolute* violation position (median 210 characters, $\approx$40 words) regardless of total output length. Models decide whether to honor the lipogram within the first paragraph.

A two-dimensional failure space. The refusal decision and the violation position are independent axes, not a single continuum. Restricting to genuine content failures and analyzing per-model medians reveals four tiers: *Immediate* (median <0.05): Qwen 3 Base (0.03)—a base model with no instruction-tuning preamble behavior. *Early* (median $0.10{-}0.24$): all three Claude models (0.13–0.14) and all three Llama models (0.15–0.24), which engage the constraint but violate within the first few sentences. *Distributed* (median $0.35{-}0.43$): Kimi K2.5 and Gemini Flash-Lite, with failures spread across the response. *Maintain-then-slip* (median $0.44{-}0.55$): Qwen 3 Instruct (0.44), Gemini 2.5 Flash (0.46), Gemini 3.1 Pro (0.55), where 30–43% of violations occur in the last 20% of the response—consistent with genuine resource depletion during generation. GPT models cannot be placed on this axis (3–8 genuine failures each; their dominant strategy is refusal). The failure space is thus two-dimensional: models first decide whether to attempt (refusal axis, driven by alignment training), then—among those that attempt—fail at characteristic positions (depth axis, reflecting processing strategy). Detailed per-model failure profiles appear in Section E.

O. Visual CSE: Degradation Hierarchy

Per-constraint degradation slopes range from -0.073 (S1, fastest) to -0.009 (A2, slowest). No constraint reaches its half-life within k=1$-$8. Set/logical constraints degrade fastest (mean slope -0.047); attribute constraints degrade slowest (-0.028), a 1.7$\times$ ratio—smaller than text CSE's $2.4\times$ (relational vs. lexical). L5 (attribute count comparison, 34.4% on solvable probes) is the hardest constraint, requiring a double-count-and-compare operation across the full scene.

Per-constraint profiles. Three patterns emerge: classic degraders with high k=1 and monotonic drop (L4: 95.5% $\rightarrow$ 51.5%, $-$44pp; A1: 92.7% $\rightarrow$ 54.1%, $-$39pp); flat constraints stable across k (L1: 51.8% $\rightarrow$ 50.3%, -1.5pp); and inverted constraints that improve with k (R2: 20.0% $\rightarrow$ 56.0%, +36pp; S1: 45.5% $\rightarrow$ 54.2%, +9pp). The inverted pattern reflects answer-set cardinality: at k=1, the correct answer may be 5–7 objects, and models include incorrect boundary objects in their answer; at $k\geq$2, additional constraints narrow the valid set, reducing over-inclusion.

P. Visual CSE: Impossibility Detection

409 impossible probes (60 deliberate + 349 from the S1 composition artifact) test whether models detect unsatisfiable constraint combinations. Nine of 10 models detect impossibility above 84%; Llama 3.2 90B is the outlier at 64.5%. Detection improves with k: 18% at k=2 to 91.2% at k=8—more constraints provide more opportunities to identify conflicts.

Solving ability and calibration are inversely correlated. GPT-5.2 is the weakest solver (37.8% solvable) but tied for best detector (90.2%). Gemini Pro is the strongest solver (79.3%) but an average detector (84.4%). The delta between combined and solvable CSRs quantifies the calibration contribution: GPT-5.2 gains +20.6pp from impossibility detection; Gemini Pro gains only +4.2pp. This dissociation reveals that visual constraint satisfaction and constraint-conflict detection draw on partially independent capabilities.

Table 12. Complete constraint taxonomy (30 constraints, 8 dimensions), part 1 of 2. The INSTRUCTION column shows the exact text injected into probe prompts; parameters in braces are instantiated from difficulty-calibrated ranges. Marginal CSRs range from 85% (L5) to 1% (S4); see Table 6 for per-constraint difficulty.

ID	Name	Instruction	Example Output
STRUCTURAL (S1–S5): constraints on document-level counts and segmentation			
S1	Exact word count	Your response must contain exactly {N} words.	"The sun rises over quiet hills…" *(50 words)*
S2	Exact sentence count	Your response must contain exactly {N} sentences.	"Stars form in nebulae. Gravity pulls…" *(5 sent.)*
S3	Paragraph structure	Your response must contain exactly {N} paragraphs, each with $\geq$2 sentences.	*(3 paragraphs, blank-line separated)*
S4	Exact character count	Your response must contain exactly {N} characters (incl. spaces/punctuation).	"Renewable energy can…" *(exactly 200 chars)*
S5	Exact line count	Your response must contain exactly {N} non-empty lines.	"Line one.\nLine two.\n…" *(5 lines)*
LEXICAL (L1–L7): character- and token-level constraints on vocabulary			
L1	Forbidden letter	Your response must not contain the letter '{c}' anywhere. No exceptions.	"Big clouds float across dusk skits" *(no 'e')*
L2	Mandatory word inclusion	Your response must include all of the following words: {list}.	*(contains "algorithm", "telescope")*
L3	Sentence-initial letter	Every sentence must begin with a word starting with '{c}'.	"Solar panels convert. Sustainable methods reduce."
L4	Initial acrostic	{N} sentences; first letters spell "{WORD}".	"**B**right stars. **R**ivers flow. **A**nd mountains."
L5	Forbidden word	Must not contain the word "{word}" (case-insensitive).	"A cat sat on a mat…" *(no "the")*
L7	Vowel–consonant ratio	Vowel/consonant ratio must be between {lo} and {hi}.	"Open areas are alive…" *(ratio* $\in [0.6, 1.0]$*)*
WORD-LEVEL (W1–W3): constraints on individual word properties across the response			
W1	All-unique words	Every word must be unique (case-insensitive).	"Renewable energy transforms global infrastructure"
W2	Word length range	Every word must be {N}–{M} characters long.	"The sun can warm our cold dark land" *(1–8 chars)*
W3	Min sentence word count	Every sentence must contain $\geq${N} words.	"Renewable energy sources are rapidly transforming…"
RELATIONAL (R1–R3): structured reasoning over entities and their relationships			
R1	Scene graph output	Output JSON with "objects" and "relations" keys reproducing all attributes and relations.	{''objects'': {''A'': {...}}, ''relations'': [...]}
R2	Transitive ordering	Given pairwise comparisons, list all items in order.	"1. Portland 2. Amsterdam 3. Toronto"
R3	Logic grid puzzle	Given clues about people/pets/colors, output solution as JSON.	{''Alice'': {''pet'': ''dog''}}

Table 13. Complete constraint taxonomy (continued), part 2 of 2. Format, Meta, Numerical, and Ordering dimensions.

ID	Name	Instruction	Example Output
FORMAT (F1–F4): constraints on output syntax and document structure			
F1	Valid JSON with schema	Entire response must be a single valid JSON object with keys {list}.	{``title'': ``...'', ``body'': ``...''}
F2	Markdown table	Include a table with {C} columns and {R} data rows.	| City | Pop | Area |
F3	Bullet list	Bulleted list with exactly {N} items. Use "- " format.	"- Solar is renewable\n- Wind generates power"
F4	Fenced code blocks	Exactly {N} fenced code blocks (triple backticks).	```\nprint('hello')\n```
META (M1–M3): self-referential and steganographic constraints			
M1	Terminal acrostic	{N} sentences; last word of each starts with a letter spelling "{WORD}".	*(last-word initials spell "LIGHT")*
M2	Self-referential count	Include "This text contains exactly N words" where N is correct.	"...This text contains exactly 42 words."
M3	Word count divisibility	Total word count must be divisible by {N}.	"Stars light the sky at dusk..." *(35 words,* $35 \bmod 7 = 0$*)*
NUMERICAL (N1–N3): constraints on numeric content within the response			
N1	Embedded integer sum	Include exactly {K} integers summing to {T}.	"45 nations...30 percent...35 on track" *(sum = 110)*
N2	No digits	No numerical digits (0–9). Write numbers as words.	"forty-five nations...thirty percent"
N3	Exact distinct number count	Exactly {N} distinct numbers (as digits).	"In 2024, roughly 35 countries invested 100B"
ORDERING (O1–O3): sequential constraints on sentence arrangement			
O1	Monotonic sentence lengths	Each sentence strictly {dir} in word count.	"Sun shines. Energy grows fast. Sources power many."
O2	Alphabetical sentences	First word of each sentence in alphabetical order.	"Agriculture thrives. Bridges connect. Cities grow."
O3	Alternating length	Alternate short ($\leq${N}) and long ($\geq${M}) sentences.	"Sun shines. Renewable energy sources are transforming..."

Table 14. Visual CSE constraint taxonomy (18 filters, 5 dimensions). All verifiers execute deterministically against the hidden scene graph.

ID	Name	Instruction	Verifier
ATTRIBUTE (A1–A3): single-property perceptual filters			
A1	Color match	The target object is {color}.	`obj.color == required`
A2	Shape match	The target object is a {shape}.	`obj.shape == required`
A3	Size match	The target object is {size}.	`obj.size == required`. Two categories (small/large).
SPATIAL (S1–S4, E1): geometric relations on grid coordinates			
S1	Left/right	The target object appears to the {left/right} of object {ref} in the image.	`obj.x > ref.x` (left) or < (right). Grid-x inverted under perspective; `GRID_X_INVERTED=True`.
S2	Above/below	The target object appears {above/below} object {ref} in the image.	`obj.y < ref.y` (above) or > (below). Grid-y not inverted.
S3	Adjacency	The target object is directly next to object {ref} (horizontally or vertically, not diagonally).	Manhattan distance = 1: `|obj.x − ref.x| + |obj.y − ref.y| == 1`.
S4	Isolation	The target object has no objects directly next to it.	No neighbor at Manhattan distance = 1. Mutually exclusive with S3 by construction.
E1	Extremal	The target object is the {leftmost/rightmost/ topmost/bottommost} object in the scene.	`obj.x == max(all_x)` for leftmost (inverted). Tie exclusion: assigned only when a unique extremal object exists.
RELATIONAL (R1–R2): cross-object attribute comparison			
R1	Same/different	The target object has the {same/different} {attr} as object {ref}.	`getattr(obj, attr) == getattr(ref, attr)` for "same"; `!=` for "different". Attributes: color, shape, size.
R2	Comparative size	The target object is {larger/smaller} than object {ref}.	`SIZE_ORDER[obj] > SIZE_ORDER[ref]` for "larger". Strict inequality; equal sizes excluded.

Table 15. Visual CSE constraint taxonomy (continued). Set/Logical and Counting dimensions.

ID	Name	Instruction	Verifier
SET / LOGICAL (L1–L5): logical operations on attribute predicates			
L1	Threshold logic	The target must satisfy at least 2 of these 3 properties: {color}, {shape}, {size}.	Count matching predicates ≥ 2. Always 3 predicates sampled from target's attributes.
L2	Disjunction	The target's {attr} must be {v1} OR {v2}.	`obj.attr in {v1, v2}`. Attributes: color or shape.
L3	Negation	The target must not be {v1} and not {v2} (by {attr}).	`obj.attr not in {v1, v2}`. Complement of L2.
L4	Uniqueness	The target must be the ONLY {value} object (by {attr}) in the scene.	Exactly 1 object has the value AND model returned that ID. Hard selector.
L5	Attr. count cmp.	The target's color appears strictly more times than its shape in the scene.	`color_count > shape_count`. Two scene-wide counts + comparison. Hardest constraint (35.5%).
COUNTING (C1–C3): scene-wide enumeration and arithmetic			
C1	Counting filter	The target's {attr} appears exactly {N} times in the scene.	`count == N`. Only assigned when $N \geq 2$ (count $= 1$ is L4).
C2	Group parity	The number of objects sharing the target's {attr} is {odd/even}.	`count % 2`.
C3	Plurality	The target's {attr} is the most common {attr} in the scene.	`count == max_count`. Signature requires unique plurality (no ties) and `max_count` ≥ 2.