Do VLMs Perceive or Recall? Probing Visual Perception vs. Memory with Classic Visual Illusions

Xiaoxiao Sun[1,*] Mingyang Li[1,*] Kun Yuan[2,3] Min Woo Sun[1] Mark Endo[1] Shengguang Wu[1]
Changlin Li[1] Yuhui Zhang[1] Zeyu Wang[1] Serena Yeung-Levy[1,†]

[1]Stanford University [2]University of Strasbourg [3]Technical University of Munich
[*]Equal contribution. [†]Corresponding author.

xxsun@stanford.edu, mli89@stanford.edu, syyeung@stanford.edu

https://sites.google.com/view/vi-probe/

In this appendix, we provide additional materials that complement the main paper and offer a more complete view of our dataset construction, experimental setup, and empirical findings. Specifically, we first present further details about the data in VI-Probe, including the illusion categories and the exact prompts used in our evaluations. We then summarize the models evaluated in the main paper and include additional experimental results that were omitted from the main submission due to space constraints. We also provide more visual examples from VI-Probe to illustrate the diversity and structure of the benchmark. Finally, we include further discussion to help contextualize the main findings and support a more detailed understanding of how visual illusions can be used to probe perception-versus-memory behavior in vision-language models.

A. VI-Probe Details & Examples

This section provides a more detailed overview of VI-Probe, with a particular focus on the benchmark composition, prompting format, and representative examples. Our goal is to make the dataset construction and evaluation protocol fully transparent so that readers can better understand what each test case measures and how the benchmark isolates perception-driven behavior from memory-driven responses. In particular, we clarify the set of illusion categories covered by VI-Probe, the exact natural-language questions posed to the models, and the instruction templates used throughout our experiments.

Table 1 lists the 27 illusion cases currently included in VI-Probe, spanning size, color, and orientation phenomena. The corresponding case-specific questions are provided in Table 2. Together, these materials give a concrete view of how we convert classic visual illusions into standardized VLM evaluation instances while keeping the query format simple and consistent across cases. We will release the images and code used to generate the different data variants, and we also plan to maintain an online resource so that the benchmark can be expanded with additional illusion families and new diagnostic settings over time.

#	Illusion Name	Category
1	Müller Lyer Illusion	size
2	Circle Müller Lyer Illusion	size
3	Ponzo Illusion	size
4	Ponzo Trapezoid Illusion	size
5	Ebbinghaus Illusion	size
6	Ebbinghaus Illusion Rectangular	size
7	Delboeuf Illusion	size
8	Oppel Kundt Illusion	size
9	Irradiation Illusion	size
10	Irradiation Pentagon Illusion	size
11	Circle Ponzo Illusion	size
12	Cornsweet Illusion	color
13	Simultaneous Contrast Illusion	color
14	Munker White Illusion	color
15	Mach Band Illusion	color
16	Mach Band Illusion Case2	color
17	Chubb Illusion	color
18	Cornsweet Illusion Case1	color
19	Hering Illusion	orientation
20	Hering Illusion Vertical	orientation
21	Zöllner Illusion	orientation
22	Zöllner Illusion Vertical	orientation
23	Twisted Cord Illusion	orientation
24	Twisted Cord Illusion Light	orientation
25	Poggendorff Illusion	orientation
26	Poggendorff Horizontal Illusion	orientation
27	Ehrenstein Illusion	orientation

Table 1. List of all Illusions cases, categorized into three groups.

Original Full Prompts used in the experiments are listed

#	Question	Ori-Answer
1	Are the two black lines of equal length?	Yes
2	Are the two black lines of equal length?	Yes
3	Are the two horizontal black lines of equal length?	Yes
4	Are the two horizontal black lines of equal length?	Yes
5	Are the two orange circles the same size?	Yes
6	Are the two orange circles the same size?	Yes
7	Are the two solid circles the same size?	Yes
8	Are the distances between the vertical markers labeled A–B and B–C equal?	Yes
9	Are the left white square and the right black square equal in size?	Yes
10	Are the left white pentagon and the right black pentagon equal in size?	Yes
11	Are the two circles the same size?	Yes
12	Are the two circles of the same color?	Yes
13	Are the two small squares of the same color?	Yes
14	Are the two rectangle the same color?	Yes
15	Is there a boundary in between every adjecent regions?	No
16	Is there a boundary in between every adjecent regions?	No
17	Are the two circles of the same color?	Yes
18	Are the two vertical bands of the same color?	Yes
19	Are the two horizontal lines straight?	Yes
20	Are the two vertical lines straight?	Yes
21	Are the those red lines straight?	Yes
22	Are the those red lines straight?	Yes
23	Are those vertical columns parallel?	Yes
24	Are those vertical columns parallel?	Yes
25	Are the red and black solid diagonal lines aligned?	Yes
26	Are the red and black solid diagonal lines aligned?	Yes
27	Do the squares on the left and right have straight edges?	Yes

Table 2. Question prompts for all illusion cases, together with the ground-truth answer for the original illusion image.

below to clarify the exact instructions given to the models and to facilitate reproducibility. We include the prompt template in full because small wording differences can meaningfully affect VLM behavior, especially in tasks that require careful visual comparison under potentially conflicting semantic priors.

```
Are the two black lines of equal length?
Answer Instructions:
1. Write your reasoning inside
<reasons>...</reasons>.
- Use natural language explanation.
2. Give the final numeric answer inside
<answer>...</answer>.
- Use "1" if yes.
- Use "0" if no.
- Do not write anything else inside <answer>.
```

The first line is adapted to each illusion case according to the questions in Table 2, while the answer format and reasoning instructions remain fixed. This design lets us vary the visual task itself without changing the response protocol, making the benchmark more controlled and allowing differences in model performance to be attributed more directly to the visual content rather than prompt variation.

Visual Comparison Instruction Prompt The following prompt is added to the system prompt for linguistic variation and for isolating the model from prior semantic knowledge. It is intended to explicitly encourage image-grounded reasoning, reduce reliance on memorized illusion templates, and test whether models can shift from prior-driven responses to direct visual comparison when such behavior is instructed.

```
Visual Comparison Instructions:
Base your judgment exclusively on direct
visual perception of the image. Compare the
two targets systematically using only what is
visible in the image itself.
Critical constraints:
• Disregard language priors and linguistic
biases
• Ignore implications from question phrasing
• Set aside world knowledge and assumptions
• Do not rely on typical patterns or
expectations
Required approach:
1. Evaluate the "equal" hypothesis against
visual evidence
2. Evaluate the "not equal" hypothesis
against visual evidence
3. Compare which hypothesis better matches
the observable data
4. Provide a binary answer based solely on
this visual analysis
Your response must be grounded entirely in
what you can directly perceive in the image.
```

Table 3 provides a quick reference for the notation used in Section 3.2 of the paper.

Symbol	Meaning
x^O	Original illusion image
x^P	Perturbed image (control factor inverted)
x^{OC}, x^{PC}	Control (inducers removed)
x^{OH}, x^{PH}	Hinted (visual cues overlaid)
q^f, q^r, q^I	Forward / Reversed / Instructional question
α	Perturbation strength

Table 3. Notation quick reference of VI-Probe

u represents any image of x^O, x^P, x^{OC}, x^{PC}, or x^{OH}, x^{PH}. The metrics for Paraphrase-pair consistency (same vs. different) are calculated as follows:

Polarity-Flip Consistency (PFC).

$$\mathrm{PFC} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbf{1}\big[a(u, q^{\mathrm{r}}) = 1 - a(u, q^{\mathrm{f}})\big].$$

Polarity-Flip Accuracy (PFA).

$$\mathrm{PFA} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbf{1}\big[a(u, q^{\mathrm{f}}) = y^{\mathrm{f}}(u)\big]\, \mathbf{1}\big[a(u, q^{\mathrm{r}}) = y^{\mathrm{r}}(u)\big].$$

Template Fixation Index (TFI).

$$\mathrm{TFI} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbf{1}\big[a(u, q^{\mathrm{r}}) = a(u, q^{\mathrm{f}})\big] = 1 - \mathrm{PFC}.$$

B. Experiment

B.1. Experimental setting

VLMs evaluated using VI-Probe. The main paper evaluates the 15 most recent models from four model families (OpenAI [4], Anthropic [2], Google [3], Qwen3-VL and Qwen2.5-VL series [5]). Earlier versions have also been assessed in the supplementary materials. Model names are provided in Table 4.

Model Series 1			
ChatGPT [1]	GPT-5	GPT-5-mini	GPT-5-nano
Claude [2]	Claude Opus 4.1	Claude Sonnet 4.5	Claude Haiku 4.5
Gemini [3]	Gemini 2.5 Pro	Gemini 2.5 Flash	Gemini 2.5 Flash-Lite
Qwen [5]	Qwen3-VL-235B-A22B	Qwen3-VL-30B	Qwen3-VL-8B
	Qwen2.5-VL-72B	Qwen2.5-VL-32B	Qwen2.5-VL-7B
Model Series 2			
ChatGPT [1]	GPT-4o	GPT-4o-mini	
Claude [2]	Claude 3.5 Sonnet	Claude 3.5 Haiku	

Table 4. List of evaluated models. Current versions and previous versions are tested for consistency.

Accuracy reported in the paper. With the exception of the Polarity-Flip Consistency framework (Sec 4.2.1.), accuracy on all other data types is computed by requiring correct answers to both paired questions, *i.e.* a case is marked correct only when the model answers both the forward and reverse prompts correctly.

B.2. Additional results and analysis

B.2.1. Cross-Generation Consistency: Previous Models

To assess consistency of illusion susceptibility patterns across model generations, we evaluated four predecessor models (Claude-3.5-Sonnet, Claude-3.5-Haiku, GPT-4o, GPT-4o-mini) using the same VI-Probe pipeline. Table 5 reports aggregate accuracy across Size illusion categories, following the same metrics as the main paper.

Pattern replication across model generations. Previous-generation models show qualitatively similar failure modes to their successors:

- **Claude-3.5-Sonnet** replicates the memory-driven profile of Claude-Sonnet-4.5. It has resulting in a significant 74.38% Original→Perturbed drop (86.78→12.40) coupled with high $R = 1.28$. This 6× accuracy ratio (Original/Perturbed = 7.0) confirms a strong template reliance persists across Claude generations.
- **Claude-3.5-Haiku** shows more balanced performance (Original 60.33, Perturbed 35.04), similar to Claude-Haiku-4.5's relatively lower susceptibility. However, its $R = 1.45$ is notably higher than Haiku-4.5's, suggesting older versions may have stronger illusion-specific biases despite lower overall capacity.
- **GPT-4o** shows moderate memory-driven behavior (70.25→20.50, 49.75% drop) with $R = 1.04$, indicating illusion effects barely exceed baseline difficulty. This aligns with the GPT-5 profile, but it demonstrates weaker overall performance, which is consistent with the architectural improvements in GPT-5.
- **GPT-4o-mini** exhibits inverted susceptibility. Original (19.83) < Perturbed (66.94). It mirrors GPT-5-Nano's anomalous pattern. This +47.11% advantage on Perturbed images with strong control performance (OC: 94.21%) confirms the inversion stems from weak illusion-specific templates rather than poor visual processing. When models lack memorized patterns, they use visual analysis and often perform better on perturbed images.

Model	Original	Perturbed	Original Control	Perturbed Control	R
claude-3.5-sonnet	86.78	12.40	90.91	32.89	1.28
claude-3.5-haiku	60.33	35.04	72.73	55.29	1.45
gpt-4o	70.25	20.50	96.69	49.01	1.04
gpt-4o-mini	19.83	66.94	94.21	53.31	1.15

Table 5. Accuracy by image type for previous-generation models. R = (Original − Perturbed) / (OC − PC) quantifies illusion-specific memory interference relative to baseline visual difficulty.

Accuracy on control data as an indicator of model visual capacity. Original Control accuracy serves as a proxy for baseline visual capability: GPT-4o (96.69%) and GPT-4o-mini (94.21%) both exceed Claude-3.5 models (72.73–90.91%), suggesting OpenAI models possess stronger low-level visual processing. Yet this advantage disappears under illusions—Claude-3.5-Sonnet's Original accuracy exceeds GPT-4o, showing that template retrieval can override visual signals even when perceptual capacity is adequate.

R Values Confirm Architecture-Specific Biases. The rank-ordering of R values shows illusion susceptibility is not simply a function of model size or recency (Claude-3.5-Haiku (1.45) > Claude-3.5-Sonnet (1.28) > GPT-4o-mini (1.15) > GPT-4o (1.04)). Instead, it reflects architectural and training choices:

- **Claude models** consistently show $R > 1.2$. This suggests that Anthropic's training pipeline emphasizes pattern matching and alignment with world knowledge, potentially increasing susceptibility to illusions.
- **Previous OpenAI models** show $R \approx 1.0$–1.15, indicating their architectures or training data distributions produce weaker illusion-specific biases.
- The inverted scaling within model families (Haiku > Sonnet, Mini > Base) replicates across generations, confirming that increased capacity systematically amplifies tem-

plate reliance in some model families.

Cross-Generation Takeaways

- **Failure modes persist across versions**: Claude-3.5-Sonnet's 7× accuracy drop mirrors Claude-4.5-Sonnet, confirming architectural biases dominate over incremental improvements.
- **Inverted susceptibility replicates**: GPT-4o-mini (Original 19.83, Perturbed 66.94) mirrors GPT-5-Nano, validating weak-template explanations.
- **Control dissociations are universal**: All models show larger Original→Perturbed drops on illusions than controls, confirming memory interference transcends model generations.

These cross-generation results validate that VI-Probe captures stable, architecture-level phenomena rather than transient artifacts of specific model versions. The consistency of failure patterns across iterations suggests fundamental limitations in how current VLM training paradigms balance perception and memory.

B.2.2. Detailed Analysis of Intervention Effects on Size Illusions

We now present full results for the intervention experiments in Sec. 4.2.4, focusing on Size illusions. We test two strategies: (1) visual hints (alignment marks, measurement grids) and (2) system prompts instructing models to ignore prior knowledge and rely on visual evidence. Tables 6, 7, and 8 show accuracy breakdowns across all 15 models.

Visual Hints: Reinforcing Templates Rather Than Enabling Visual Reasoning. Table 6 shows an asymmetry in how visual hints affect model performance. On Original illusion images, hints improve accuracy for 13 of 15 models (mean gain: +6.2%). The largest improvements occur in perception-limited models: Qwen2.5-VL-7B (+22.31%), Claude-Haiku-4.5 (+15.70%), and Qwen3-VL-235B (+14.87%). For models without strong illusion-specific templates, hints provide useful visual guidance.

However, on Perturbed images (where illusion factors are inverted), visual hints *degrade* performance for 12 of 15 models (mean drop: -6.9%). The largest degradations occur in memory-driven models: Claude-Sonnet-4.5 (-17.85%), Gemini-2.5-Flash (-12.48%), and Qwen3-VL-235B (-9.09%). Visual hints thus act as *pattern-completion cues* that reinforce memorized configurations. When visual evidence contradicts the expected template, hints become misleading anchors that pull predictions toward incorrect memorized patterns.

Only two models show slight improvements on Perturbed images with hints: Qwen3-VL-32B (+3.31%) and Qwen2.5-VL-72B (+0.33%). Both have weak illusion-specific templates (low R in Table 2), so hints can guide visual perception without triggering strong template retrieval.

Model	Original			Perturbed		
	w/o Hints (%)	w/ Hints (%)	Effect (%)	w/o Hints (%)	w/ Hints (%)	Effect (%)
GPT-5	98.35	100.00	+1.65	1.98	0.17	-1.82
GPT-5-mini	89.26	91.74	+2.48	6.78	2.81	-3.97
GPT-5-nano	19.01	33.88	+14.87	66.03	63.72	-2.31
Claude-Opus-4.1	71.90	75.21	+3.31	30.41	24.13	-6.28
Claude-Sonnet-4.5	60.33	64.46	+4.13	41.07	23.22	-17.85
Claude-Haiku-4.5	26.45	42.15	+15.70	72.31	64.63	-7.68
Gemini-2.5-Flash	77.69	81.82	+4.13	22.98	10.50	-12.48
Gemini-2.5-Flash-Lite	42.15	44.63	+2.48	46.86	39.17	-7.69
Qwen3-VL-235B	80.17	95.04	+14.87	12.48	3.39	-9.09
Qwen3-VL-32B	94.21	83.47	-10.74	2.23	5.54	+3.31
Qwen3-VL-8B	63.64	71.90	+8.26	31.49	27.19	-4.30
Qwen2.5-VL-72B	71.07	65.29	-5.78	6.03	6.37	+0.33
Qwen2.5-VL-32B	56.20	61.16	+4.96	20.58	17.27	-3.30
Qwen2.5-VL-7B	53.72	76.03	+22.31	19.26	13.47	-5.79
Qwen2.5-VL-3B	11.57	20.66	+9.09	17.36	14.71	-2.65

Table 6. Guide Effect on Model Performance (Size Illusion). w/o Hints: without visual hints, w/ Hints: with visual hints overlaid. Effect = w/ Hints - w/o Hints. Positive values indicate improvement with visual guidance.

System Prompts: Exposing Binary Mode Switching. Tables 7 and 8 show that instructing models to "ignore prior knowledge and compare carefully" triggers *all-or-nothing mode switching* in memory-driven models. Table 7 demonstrates severe trade-offs for frontier models with $R > 1.2$:

- **GPT-5**: Original drops 84.30% (98.35%→14.05%) while Perturbed surges 63.97% (1.98%→65.95%)
- **GPT-5-mini**: Original drops 28.93% while Perturbed gains 34.62%
- **Gemini-2.5-Flash**: Original drops 19.01% while Perturbed gains 28.26%
- **Claude-Sonnet-4.5**: Original drops 21.49% while Perturbed gains 22.90%

These extreme swings demonstrate that system prompts force binary switching between two mutually exclusive modes: (1) template retrieval (high Original, low Perturbed) and (2) visual analysis (low Original, high Perturbed). For example, GPT-5 achieved an accuracy of 65.95% on Perturbed images with system prompts, compared to only 1.98% without them. This demonstrates that the model has sufficient visual capability when templates are suppressed. The failure stems not from perceptual limits but from the inability to adaptively balance memory and perception.

Conversely, three smaller Qwen models (2.5-VL-3B/7B/32B, shown in blue in Table 3 of the main paper) exhibit *simultaneous gains* on both Original and Perturbed conditions:

- **Qwen2.5-VL-3B**: Original +26.45%, Perturbed +13.30%
- **Qwen2.5-VL-7B**: Original +5.78%, Perturbed +6.36%
- **Qwen2.5-VL-32B**: Original +2.48%, Perturbed +1.57%

Model	Original			Perturbed		
	w/o SP	w/ SP	Diff.	w/o SP	w/ SP	Diff.
GPT-5	98.35	14.05	−84.30	1.98	65.95	+63.97
GPT-5-mini	89.26	60.33	−28.93	6.78	41.40	+34.62
GPT-5-nano	19.01	14.88	−4.13	66.03	77.52	+11.49
Claude-Opus-4.1	71.90	53.72	−18.18	30.41	51.57	+21.16
Claude-Sonnet-4.5	60.33	38.84	−21.49	41.07	63.97	+22.90
Claude-Haiku-4.5	26.45	21.49	−4.96	72.31	77.02	+4.71
Gemini-2.5-Flash	77.69	58.68	−19.01	22.98	51.24	+28.26
Gemini-2.5-Flash-Lite	42.15	30.58	−11.57	46.86	65.45	+18.59
Qwen3-VL-235B	80.17	66.94	−13.23	12.48	25.62	+13.14
Qwen3-VL-32B	94.21	65.29	−28.92	2.23	22.89	+20.66
Qwen3-VL-8B	63.64	57.85	−5.79	31.49	33.88	+2.39
Qwen2.5-VL-72B	71.07	65.29	−5.78	6.03	26.69	+20.66
Qwen2.5-VL-32B	56.20	58.68	+2.48	20.58	22.15	+1.57
Qwen2.5-VL-7B	53.72	59.50	+5.78	19.26	25.62	+6.36
Qwen2.5-VL-3B	11.57	38.02	+26.45	17.36	30.66	+13.30

Table 7. Comparison of Both Correct Accuracy with and without system prompting (SP) on **Original and Perturbed images**.

Model	Original			Perturbed		
	w/o SP	w/ SP	Diff.	w/o SP	w/ SP	Diff.
GPT-5	100.00	100.00	+0.00	59.92	65.37	+5.45
GPT-5-mini	99.17	95.87	−3.30	32.64	58.10	+25.46
GPT-5-nano	25.62	30.58	+4.96	81.65	72.81	−8.84
Claude-Opus-4.1	96.69	95.87	−0.82	67.52	66.36	−1.16
Claude-Sonnet-4.5	92.56	89.26	−3.30	75.45	73.55	−1.90
Claude-Haiku-4.5	90.91	85.12	−5.79	78.10	80.00	+1.90
Gemini-2.5-Flash	100.00	95.04	−4.96	50.50	55.95	+5.45
Gemini-2.5-Flash-Lite	97.52	64.46	−33.06	56.36	77.60	+21.24
Qwen3-VL-235B	100.00	100.00	+0.00	16.28	26.28	+10.00
Qwen3-VL-32B	100.00	100.00	+0.00	15.62	55.70	+40.08
Qwen3-VL-8B	99.17	99.17	+0.00	23.31	29.50	+6.19
Qwen2.5-VL-72B	100.00	97.52	−2.48	20.08	36.36	+16.28
Qwen2.5-VL-32B	93.39	98.35	+4.96	12.73	17.85	+5.12
Qwen2.5-VL-7B	85.12	100.00	+14.88	22.23	22.56	+0.33
Qwen2.5-VL-3B	84.30	99.17	+14.87	6.86	11.57	+4.71

Table 8. Comparison of Correct Accuracy with and without system prompting (SP) on **Original Control and Perturbed Control images**. Difference = w/ SP − w/o SP.

This uniform improvement confirms that weak template stores allow system prompts to enhance visual reasoning without triggering mode collapse. These models benefit from explicit instructions because they lack strong priors to override.

Control Condition Reveals True Visual Capability. Table 8 shows system prompt effects on control images (illusion patterns removed). On Original Control images, most models show minimal changes or slight degradation (mean:−2.1%), with the notable exception of smaller Qwen models showing improvements (Qwen2.5-VL-7B: +14.88%, Qwen2.5-VL-3B: +14.87%). This means flagship models already operate near ceiling on simple comparisons, leaving little room for prompt-driven gains

On Perturbed Control images, system prompts yield broader improvements: 12 out of 15 models gain accuracy (mean: +11.8%). The largest gains occur in models with moderate baseline performance: Qwen3-VL-32B (+40.08%), GPT-5-mini (+25.46%), and Gemini-2.5-Flash-Lite (+21.24%). This indicates that explicit visual comparison instructions help when visual signals are degraded but no interfering templates exist.

Comparing Tables 7 and 8 exposes a critical dissociation: flagship models improve substantially on Perturbed Control (+5.45% to +25.46%) yet collapse on Original illusions (−84.30% to −11.57%). This confirms that template retrieval drives failure under illusions, rather than perceptual limits. **When illusion semantics are absent (controls), the same prompts that suppress templates enable visual processing.**

Takeaways

- Visual hints act as pattern-completion cues, improving template matching (+6.2% on Original) but harming visual updating (−6.9% on Perturbed).
- System prompts expose binary mode switching in flagship models: suppressing templates boosts Perturbed accuracy (+64% for GPT-5) but devastates Original accuracy (−84%).
- Smaller models with weak templates (Qwen2.5-VL-3B/7B/32B) show uniform gains, confirming interventions help when strong priors are absent.
- The control vs. illusion dissociation proves failures stem from memory interference, not perceptual capacity: prompts that improve Perturbed Control by +25% can collapse Original illusions by −84% in the same model.

These results demonstrate that current VLMs lack metacognitive mechanisms to apply instructions selectively. They cannot modulate template reliance based on task demands, instead switching uniformly between retrieval-dominant and perception-dominant modes across all inputs.

Group	Model	Original	OC	Perturbed	PC	Inducer
OpenAI	GPT-5	1.00	1.00	0.01	0.55	1.00
	GPT-5-mini	1.00	1.00	0.00	0.27	1.00
	GPT-5-nano	0.45	0.18	0.65	0.95	1.00
Anthropic	Claude-Opus-4.1	1.00	1.00	0.03	0.88	0.64
	Claude-Sonnet-4.5	1.00	1.00	0.01	0.94	0.91
	Claude-Haiku-4.5	0.55	1.00	0.61	0.99	1.00
Google	Gemini-2.5-Flash	1.00	1.00	0.01	0.54	1.00
	Gemini-2.5-Flash-Lite	0.73	1.00	0.16	0.63	1.00
Qwen	Qwen3-VL-235B-A22B	1.00	1.00	0.99	0.32	0.73
	Qwen3-VL-32B	1.00	1.00	0.72	0.49	1.00
	Qwen3-VL-8B	1.00	1.00	0.25	0.20	1.00
	Qwen2.5-VL-72B	1.00	1.00	0.12	0.21	0.00
	Qwen2.5-VL-32B	1.00	1.00	0.00	0.06	1.00
	Qwen2.5-VL-7B	0.36	1.00	0.06	0.26	0.91
	Qwen2.5-VL-3B	0.00	1.00	0.01	0.04	0.00

Table 9. Model performance on the Müller-Lyer Illusion.

Group	Model	Original	OC	Perturbed	PC	Inducer
OpenAI	GPT-5	1.00	1.00	0.00	0.83	1.00
	GPT-5-mini	1.00	1.00	0.02	0.56	1.00
	GPT-5-nano	0.45	0.00	0.62	0.93	0.00
Anthropic	Claude-Opus-4.1	1.00	1.00	0.00	0.80	0.00
	Claude-Sonnet-4.5	0.82	0.82	0.29	0.93	0.00
	Claude-Haiku-4.5	0.00	0.64	0.96	0.95	0.00
Google	Gemini-2.5-Flash	1.00	1.00	0.00	0.71	0.18
	Gemini-2.5-Flash-Lite	1.00	1.00	0.01	0.72	0.18
Qwen	Qwen3-VL-235B-A22B	1.00	1.00	0.00	0.45	0.00
	Qwen3-VL-32B	1.00	1.00	0.00	0.38	0.00
	Qwen3-VL-8B	1.00	1.00	0.06	0.62	0.00
	Qwen2.5-VL-72B	1.00	1.00	0.06	0.16	0.00
	Qwen2.5-VL-32B	0.73	1.00	0.11	0.03	0.00
	Qwen2.5-VL-7B	0.91	0.64	0.22	0.38	0.00
	Qwen2.5-VL-3B	0.09	1.00	0.91	0.28	0.00

Table 10. Model performance on the Ebbinghaus Illusion.

B.2.3. Case-Specific Analysis: Inducer-Only Tests Expose Pure Template Bias

Tables 9 and 10 present detailed results for Müller-Lyer and Ebbinghaus illusions across Original, Perturbed, and Control variants. Critically, the **Inducer** column tests pure template bias: models are shown *only* the illusion-inducing elements (e.g., arrow heads without line segments, surrounding circles without center circles) and asked the same question ("Are the two lines equal length?"). High Inducer accuracy indicates the model reproduces Original answers despite *no visual evidence*, showing pure hallucination driven by memorized patterns.

Inducer-Only Reveals Extreme Template Bias. All OpenAI and Google models achieve perfect Inducer scores (1.00) on Müller-Lyer, meaning they answer identically to Original images even when the comparison targets are completely absent. GPT-5 demonstrates the most extreme bias: Original (1.00), Inducer (1.00), yet Perturbed (0.01). It generates a false perception in the absence of lines and fails when visual evidence contradicts prior knowledge. Conversely, all Qwen models and Claude models show 0.00 Inducer accuracy on Ebbinghaus, indicating they do not reproduce memorized answers when visual evidence is removed, confirming weaker illusion-specific templates for this case.

Illusion-Specific Patterns. Müller-Lyer triggers stronger memory-driven failures than Ebbinghaus. Flagship models (GPT-5, Gemini-2.5-Flash, Claude-Sonnet-4.5) show 99% Original→Perturbed drops on Müller-Lyer, far exceeding their 45–60% drops on controls, quantifying pure illusion interference. Ebbinghaus exhibits non-monotonic scaling: Claude-Haiku-4.5 achieves 0.96 Perturbed accuracy (best among all models) versus 0.00–0.29 for larger Claude siblings, replicating the main paper's finding that increased capacity amplifies template reliance for certain illusion types.

Exceptional Cases. Two models show unique processing: (1) Qwen3-VL-235B achieves near-perfect accuracy on both Original (1.00) and Perturbed (0.99) Müller-Lyer, with Perturbed accuracy *exceeding* Perturbed Control by +67pp—suggesting reliance on geometric structure rather than semantic templates. (2) GPT-5-Nano exhibits inverted susceptibility (Original $<$ Perturbed) on both illusions, confirming weak semantic priors make canonical configurations harder than perturbed variants.

Key Takeaways

- **Inducer-only = pure hallucination test**: GPT-5/Gemini achieve Inducer=1.00 on Müller-Lyer (answering correctly despite no lines), proving extreme template bias.
- **99% illusion-specific drops**: Müller-Lyer Original→Perturbed drops (GPT-5, Sonnet-4.5) far exceed control drops, isolating memory interference.
- **Non-monotonic scaling**: Claude-Haiku-4.5 outperforms larger siblings (0.96 vs. 0.00–0.29 on Ebbinghaus), confirming capacity amplifies template reliance.
- **Illusion heterogeneity**: No model excels universally; Qwen3-VL-235B dominates Müller-Lyer (0.99) but fails Ebbinghaus (0.00).

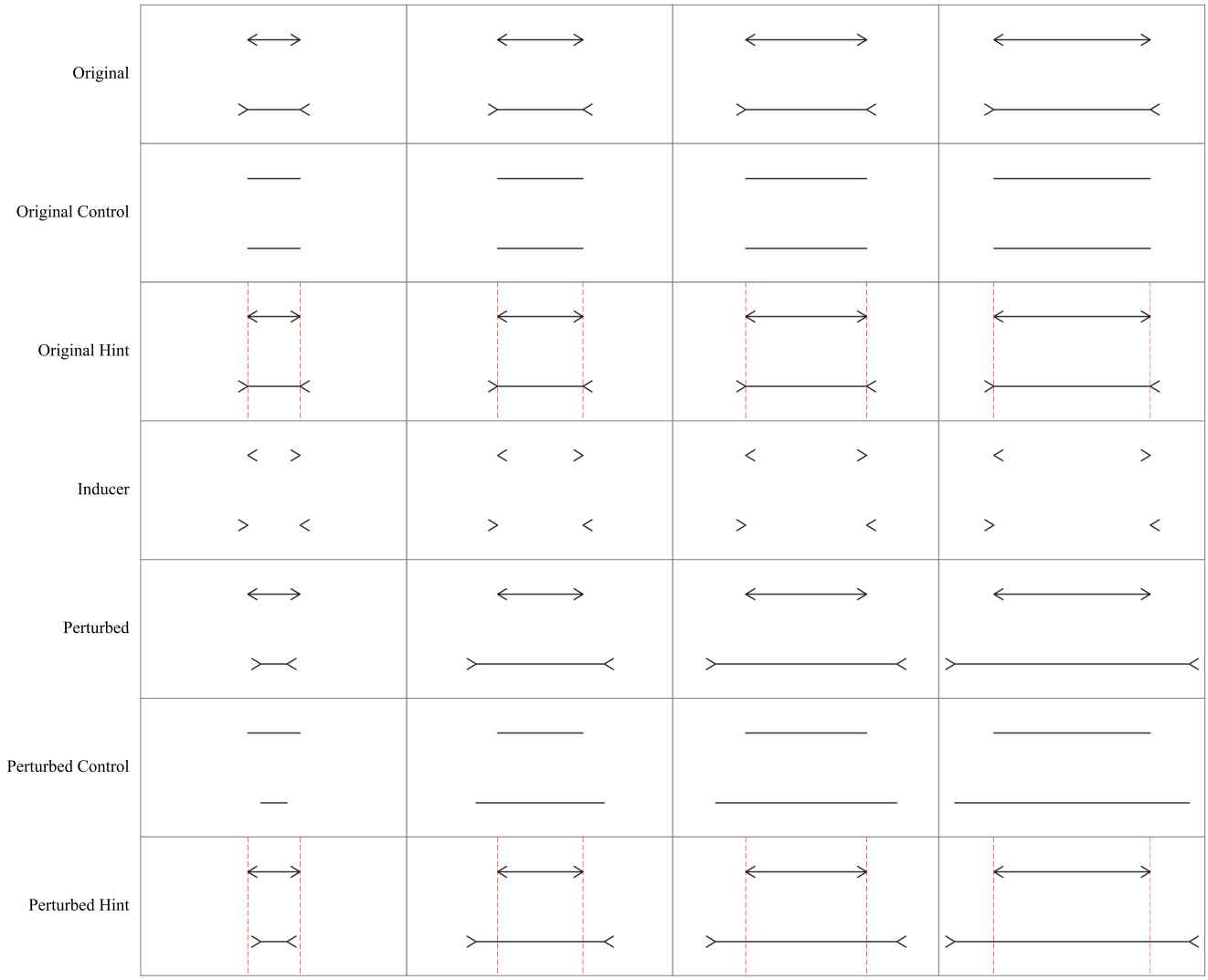
C. Examples of VI-Probe

This section provides visual examples from VI-Probe, illustrating how illusion-inducing elements (inducers) are manipulated across different data types, complete case-level variations for representative illusions, and visual embedding analyses that validate our perturbation pipeline.

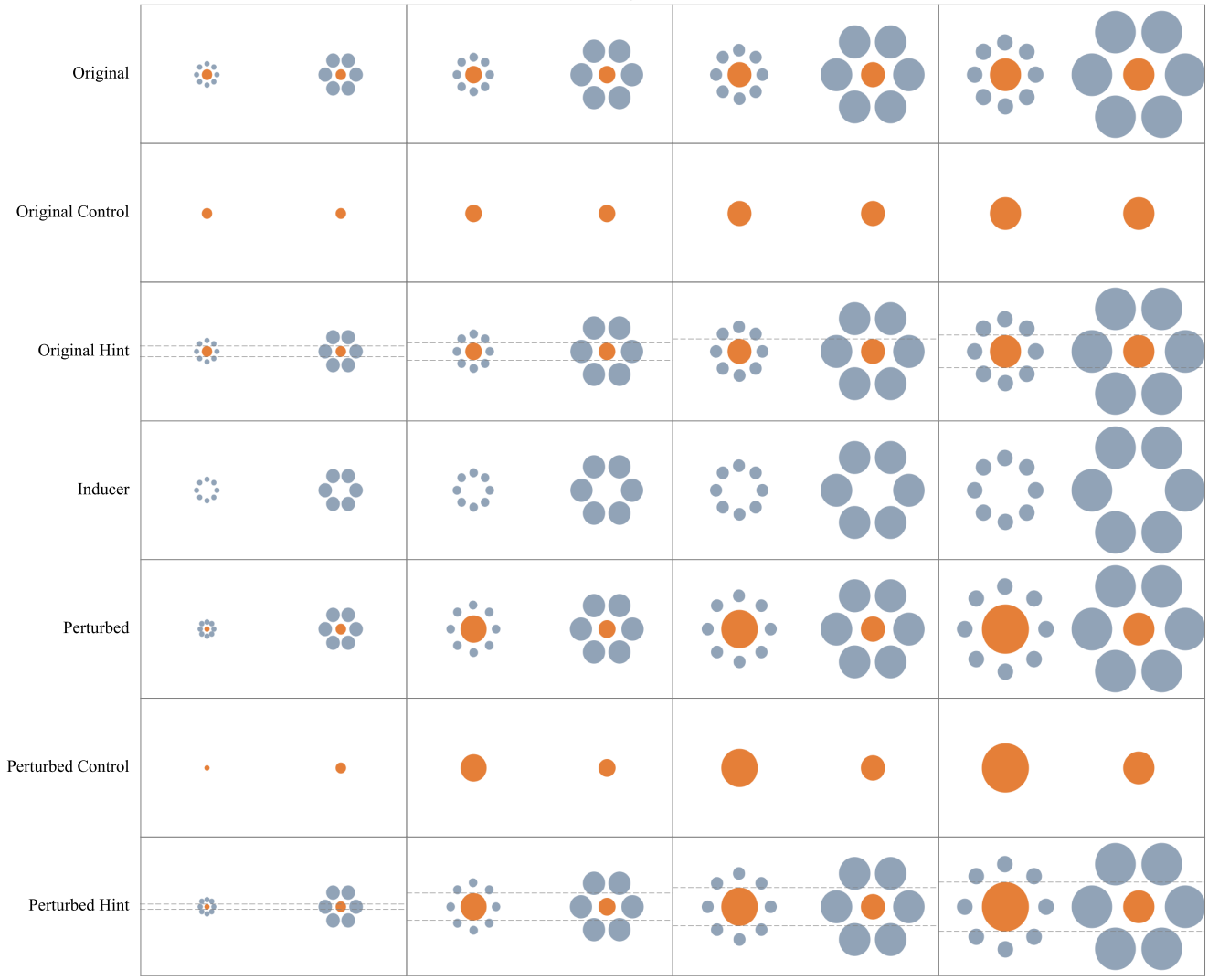
C.1. Illustration of inducer

Figure examples below demonstrate how VI-Probe systematically manipulates illusion-inducing elements across Original, Perturbed, and Control conditions to isolate the specific contribution of each visual factor.

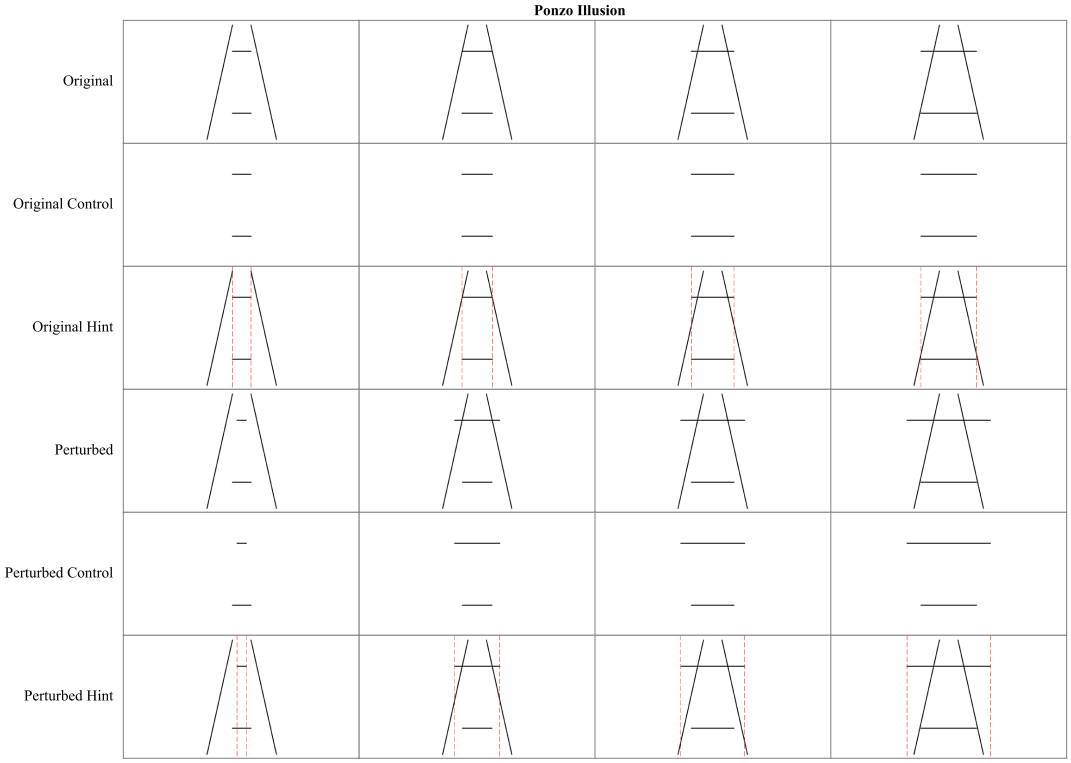
Muller Lyer Illusion

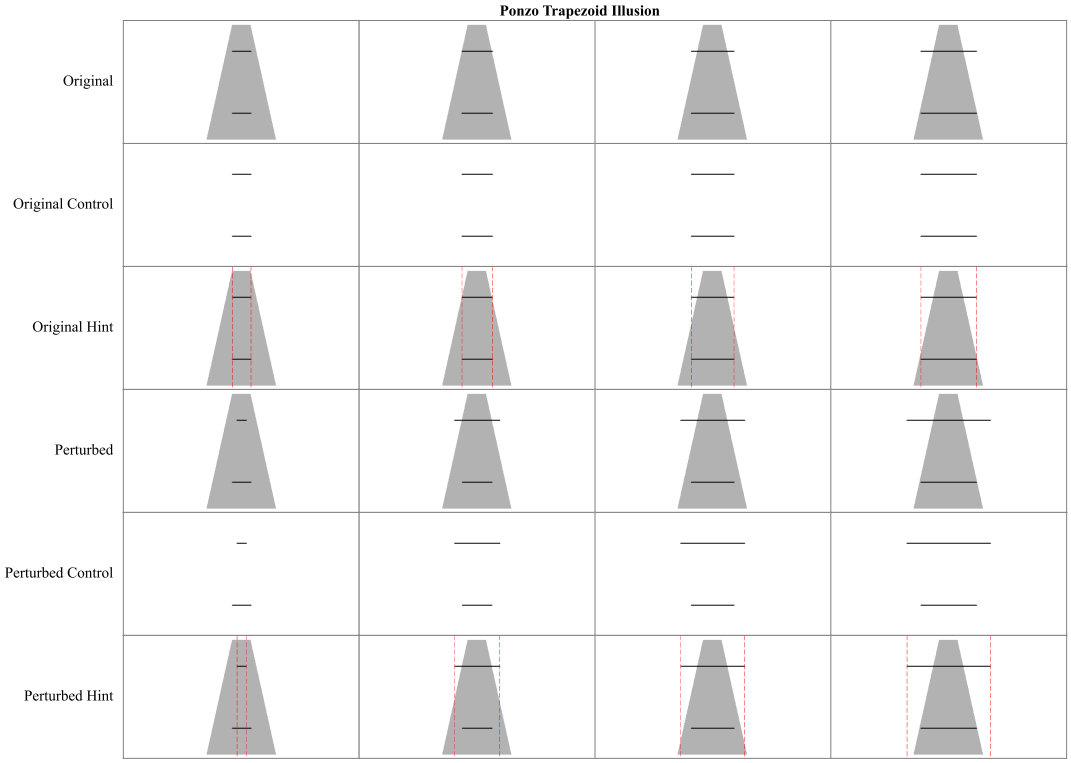


Ebbinghaus Illusion

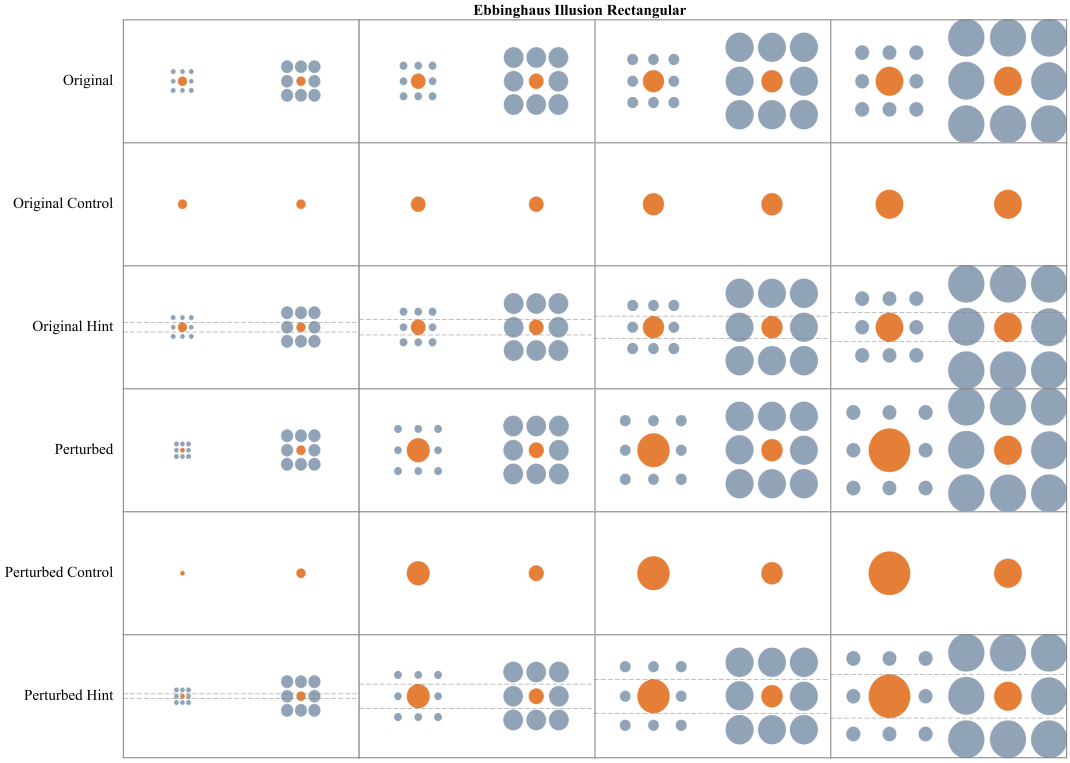


C.2. Visual examples of VI-Probe

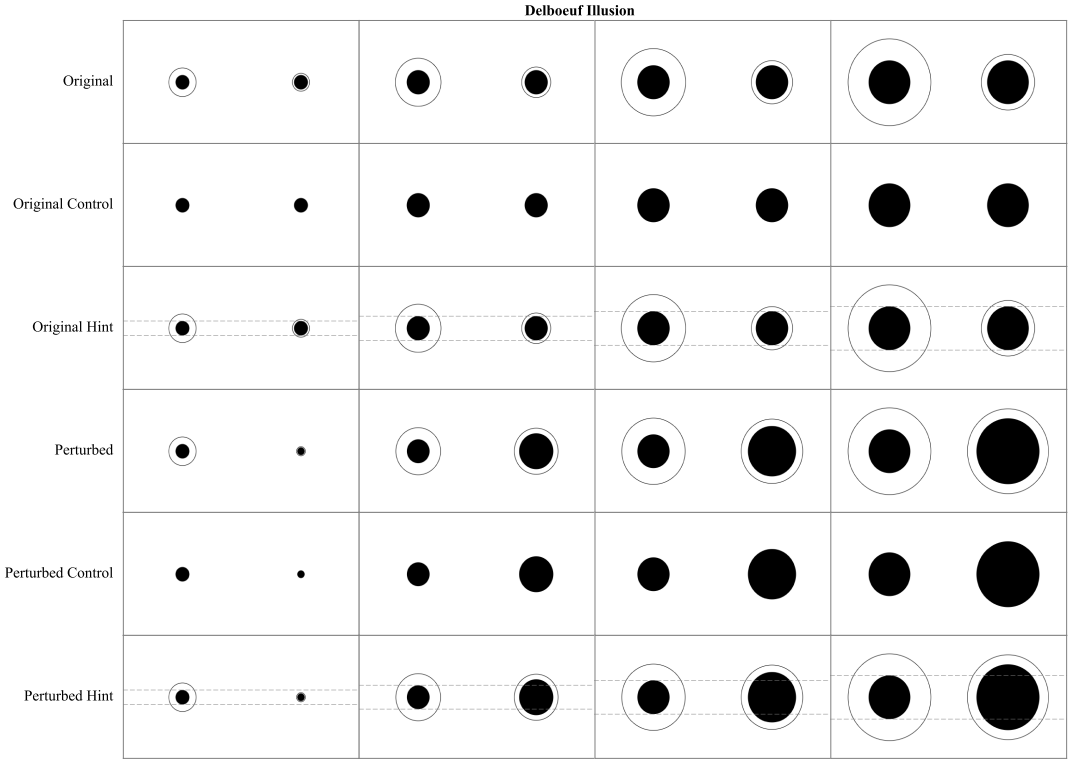




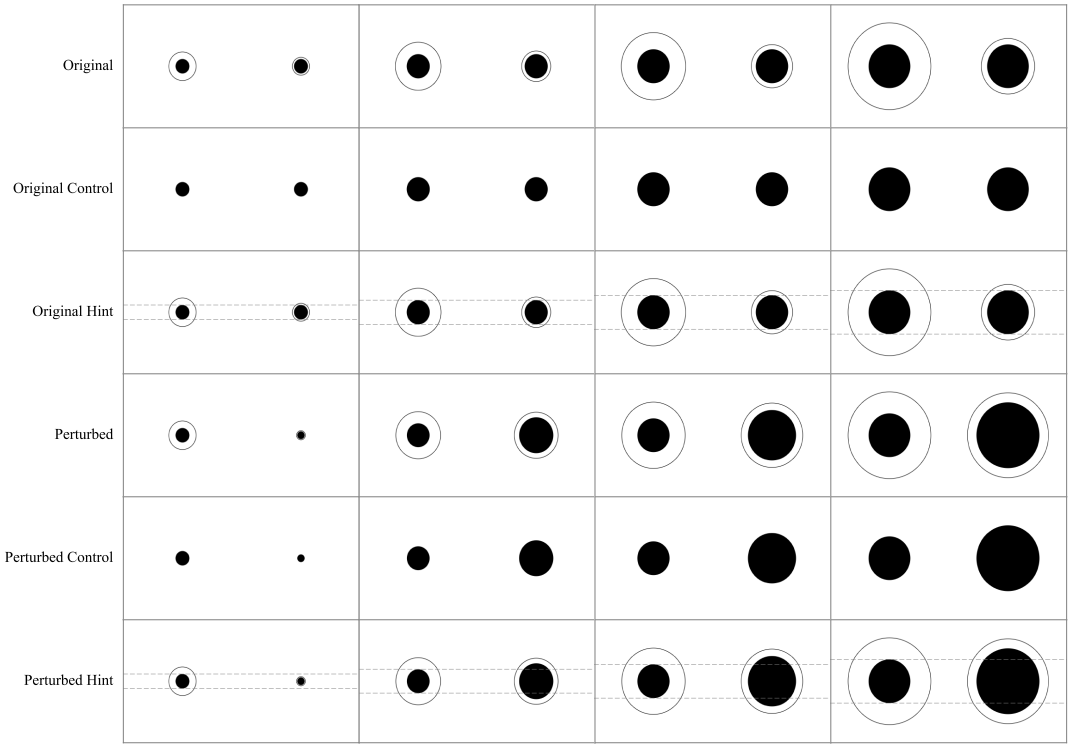
Ebbinghaus Illusion Rectangular



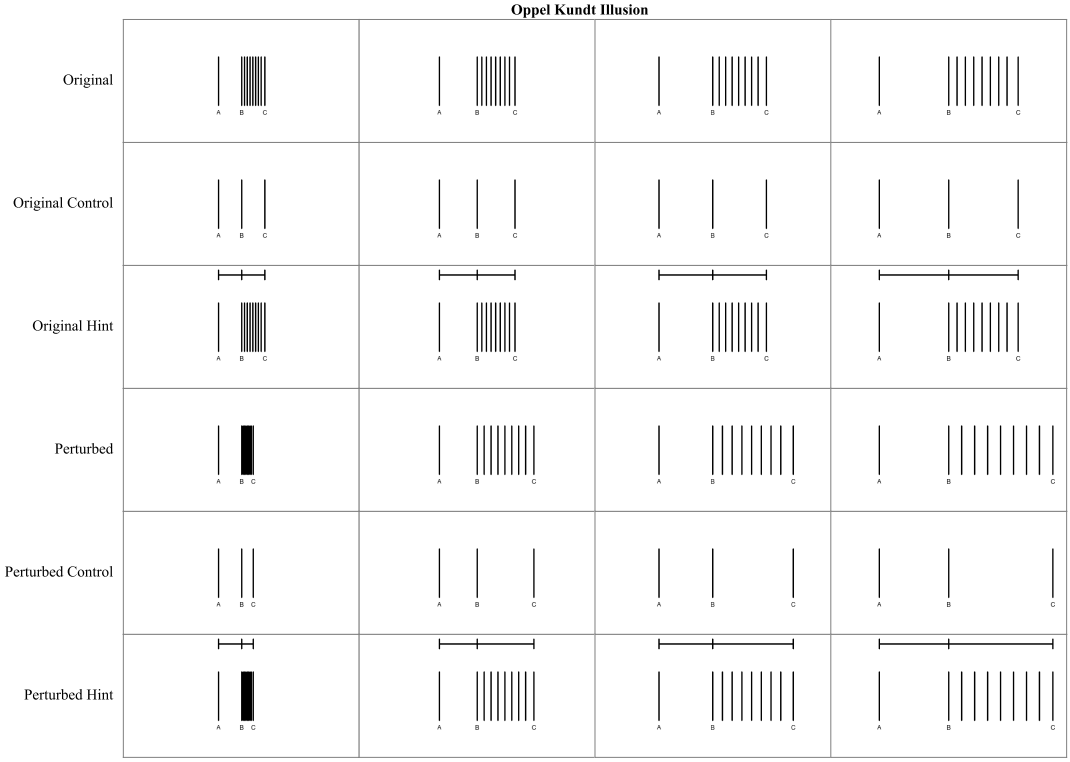
Delboeuf Illusion



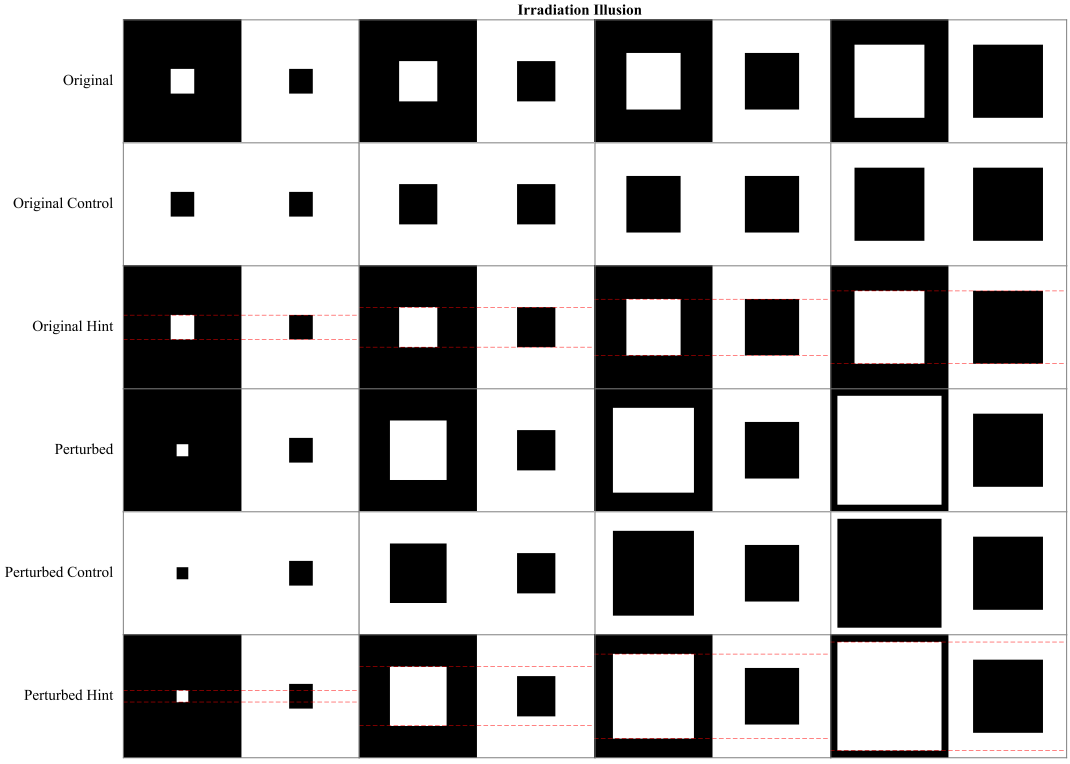
Delboeuf Illusion



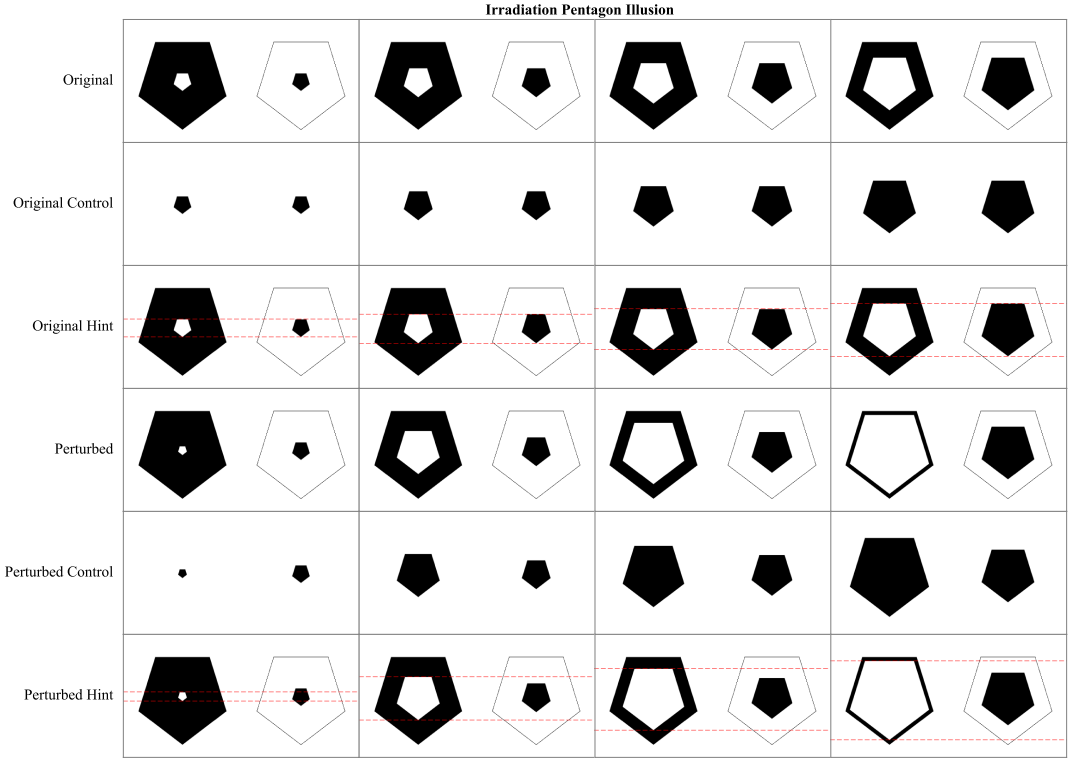
Oppel Kundt Illusion



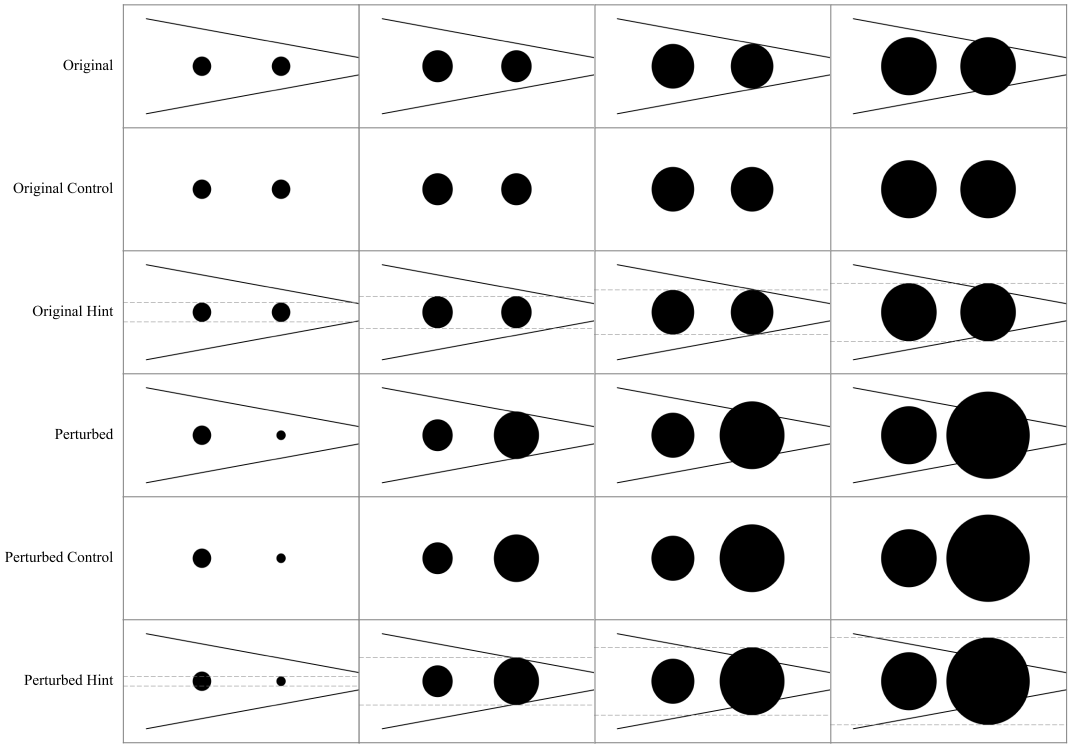
Irradiation Illusion



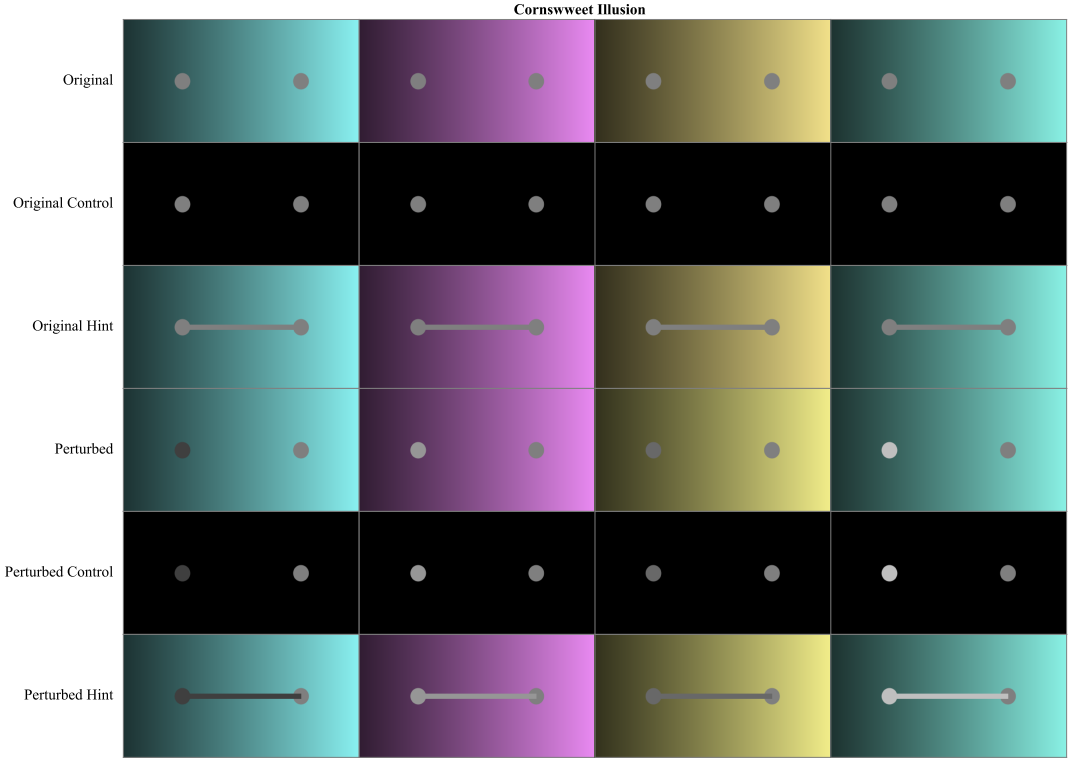
Irradiation Pentagon Illusion

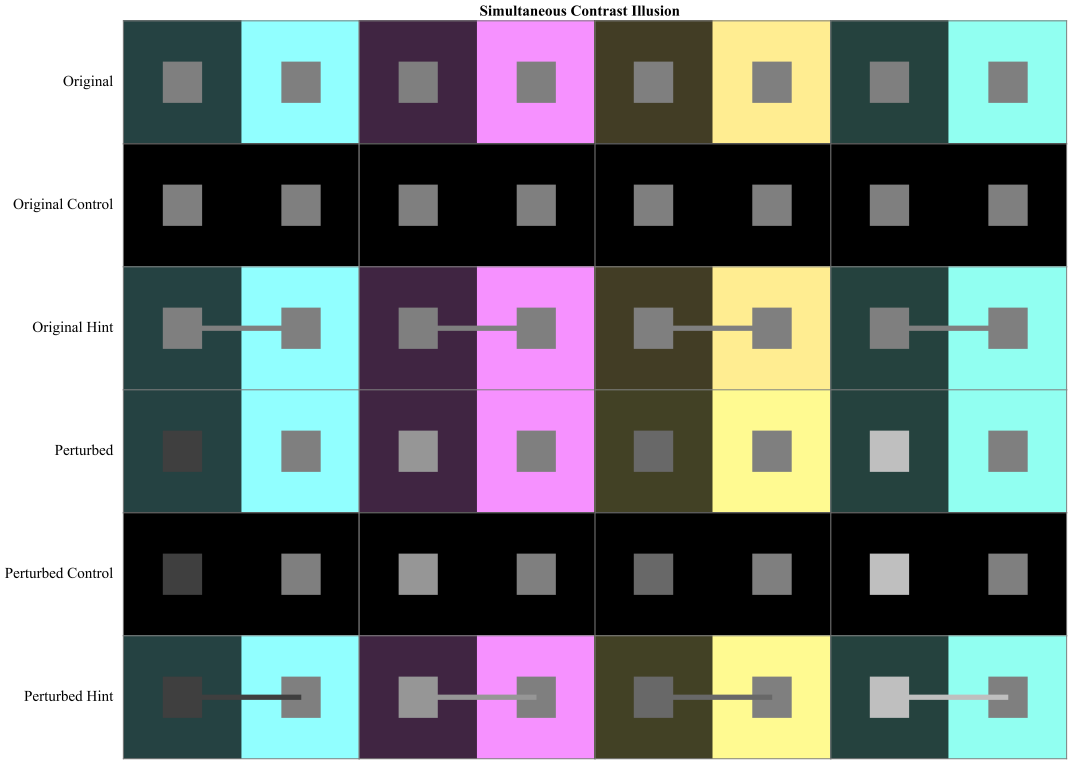


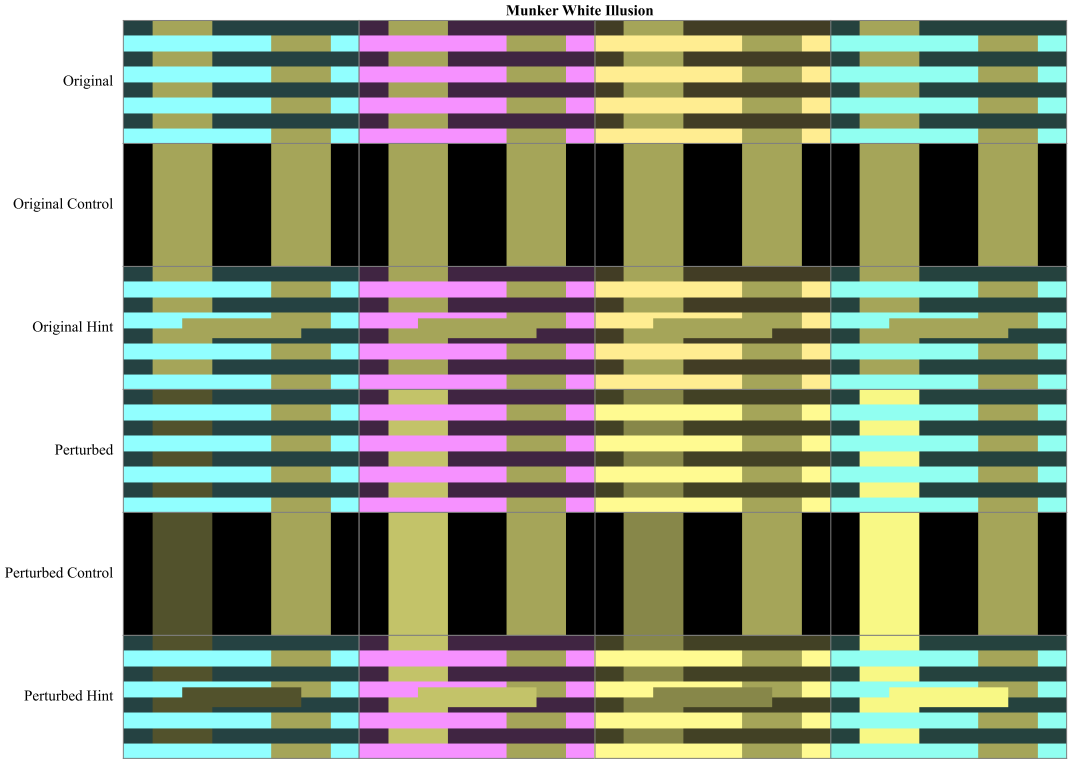
Circle Ponzo Illusion

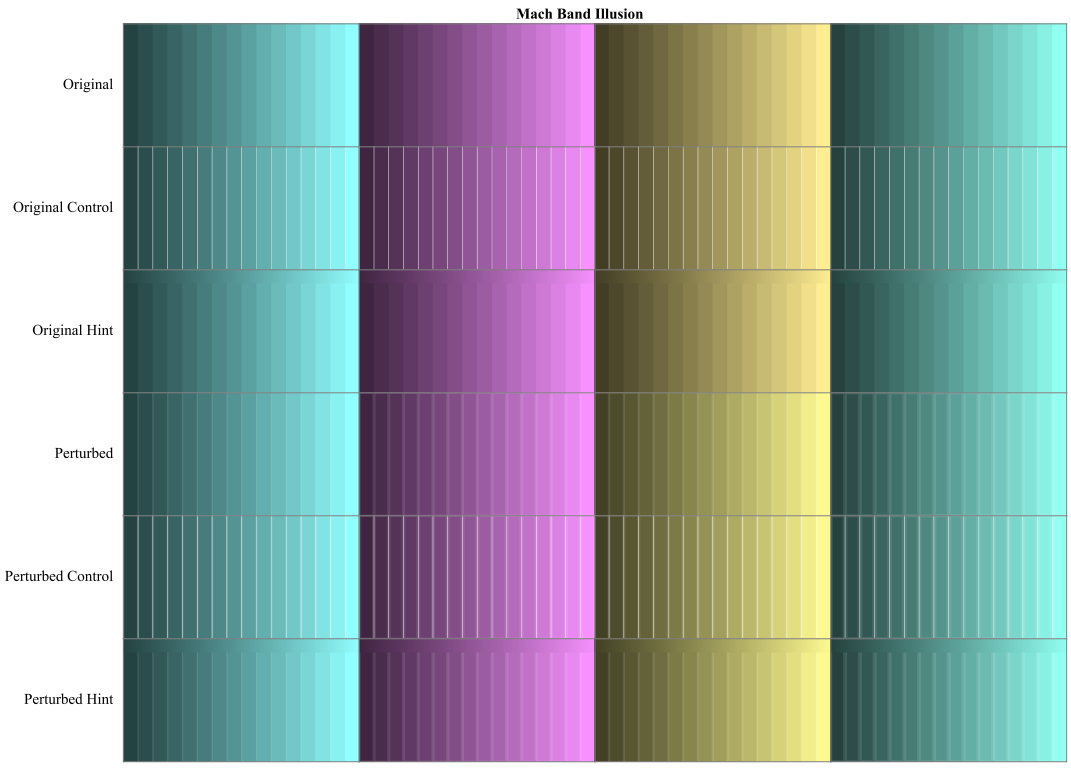


Cornswweet Illusion

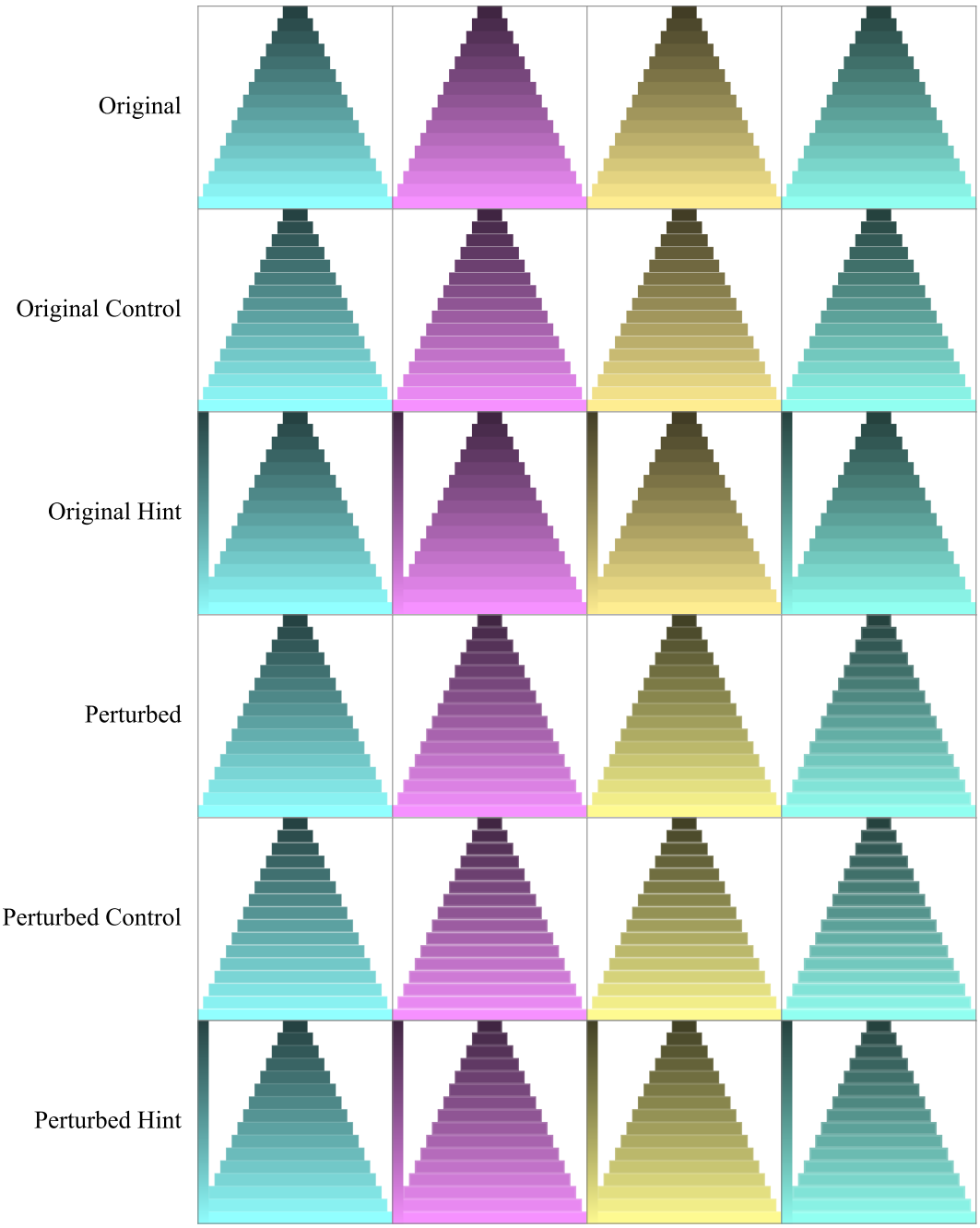


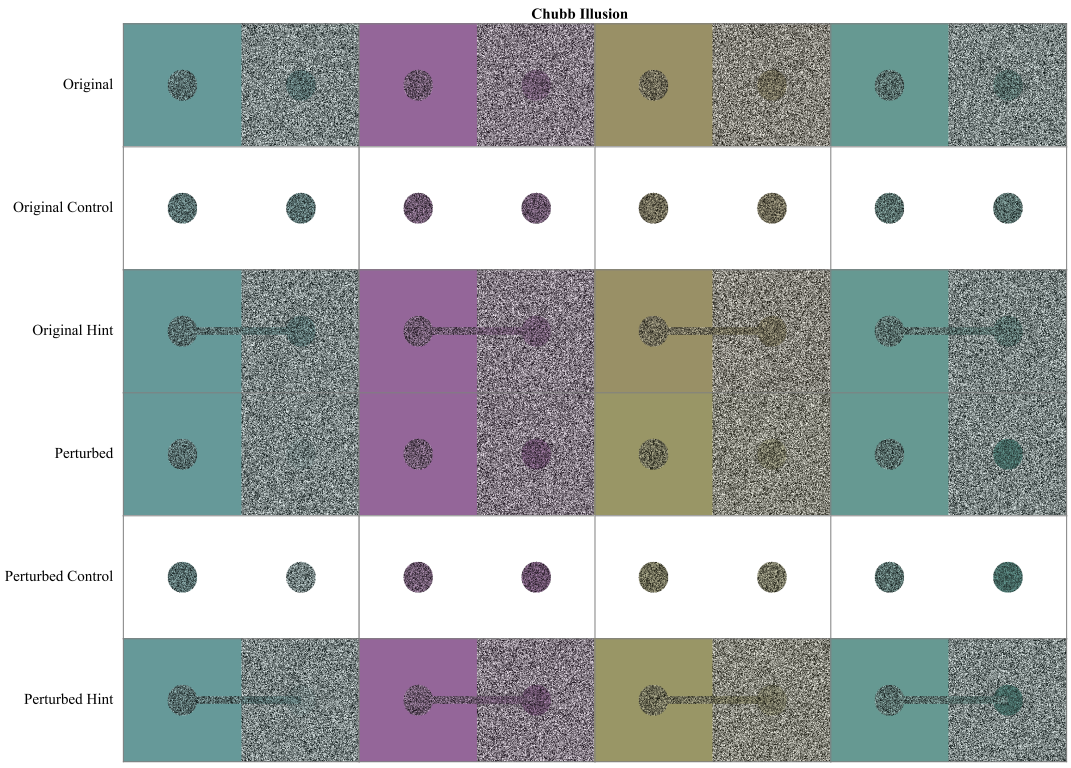

Simultaneous Contrast Illusion
Original
Original Control
Original Hint
Perturbed
Perturbed Control
Perturbed Hint


Munker White Illusion
Original
Original Control
Original Hint
Perturbed
Perturbed Control
Perturbed Hint

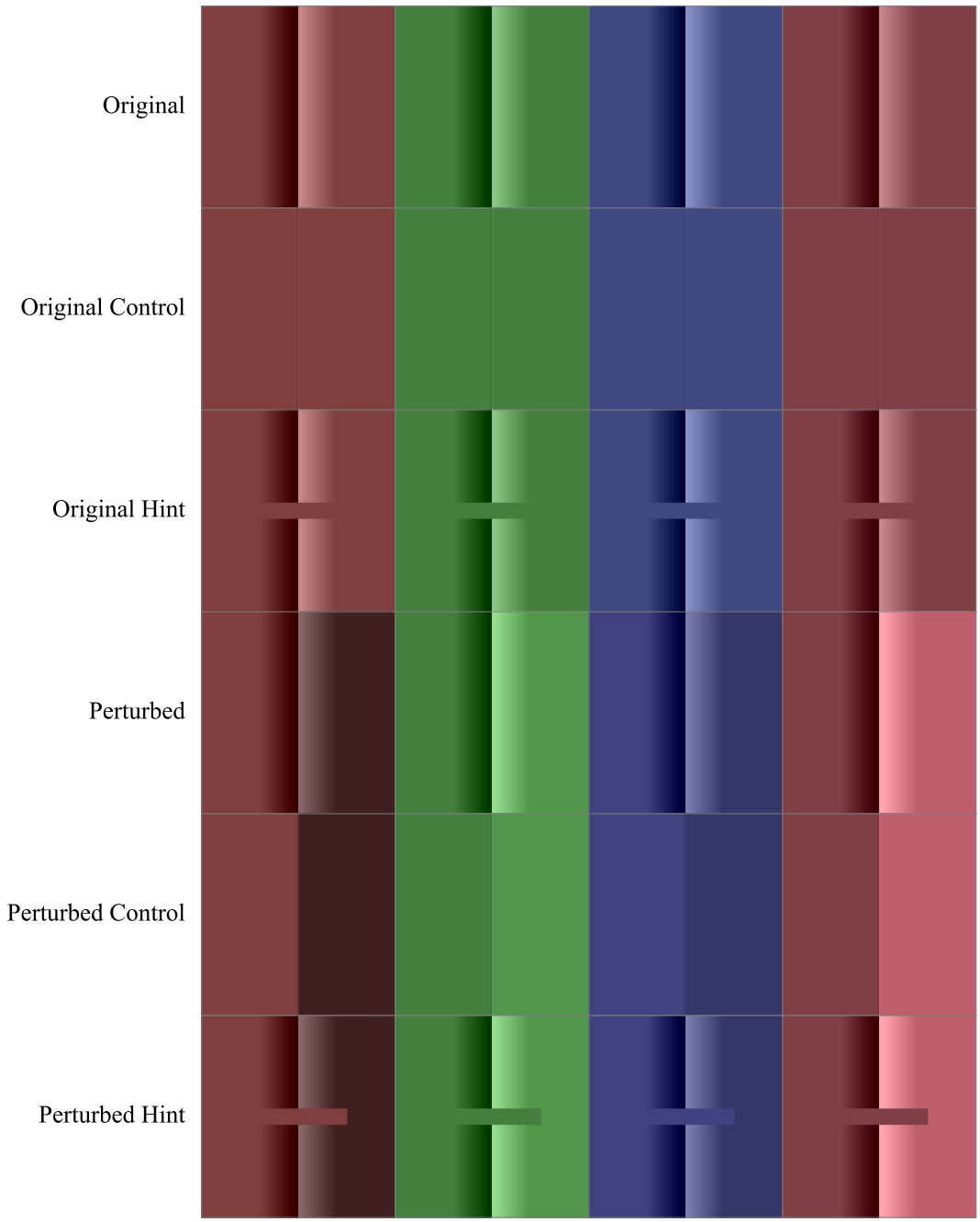

Mach Band Illusion
Original
Original Control
Original Hint
Perturbed
Perturbed Control
Perturbed Hint

Mach Band Illusion Case2

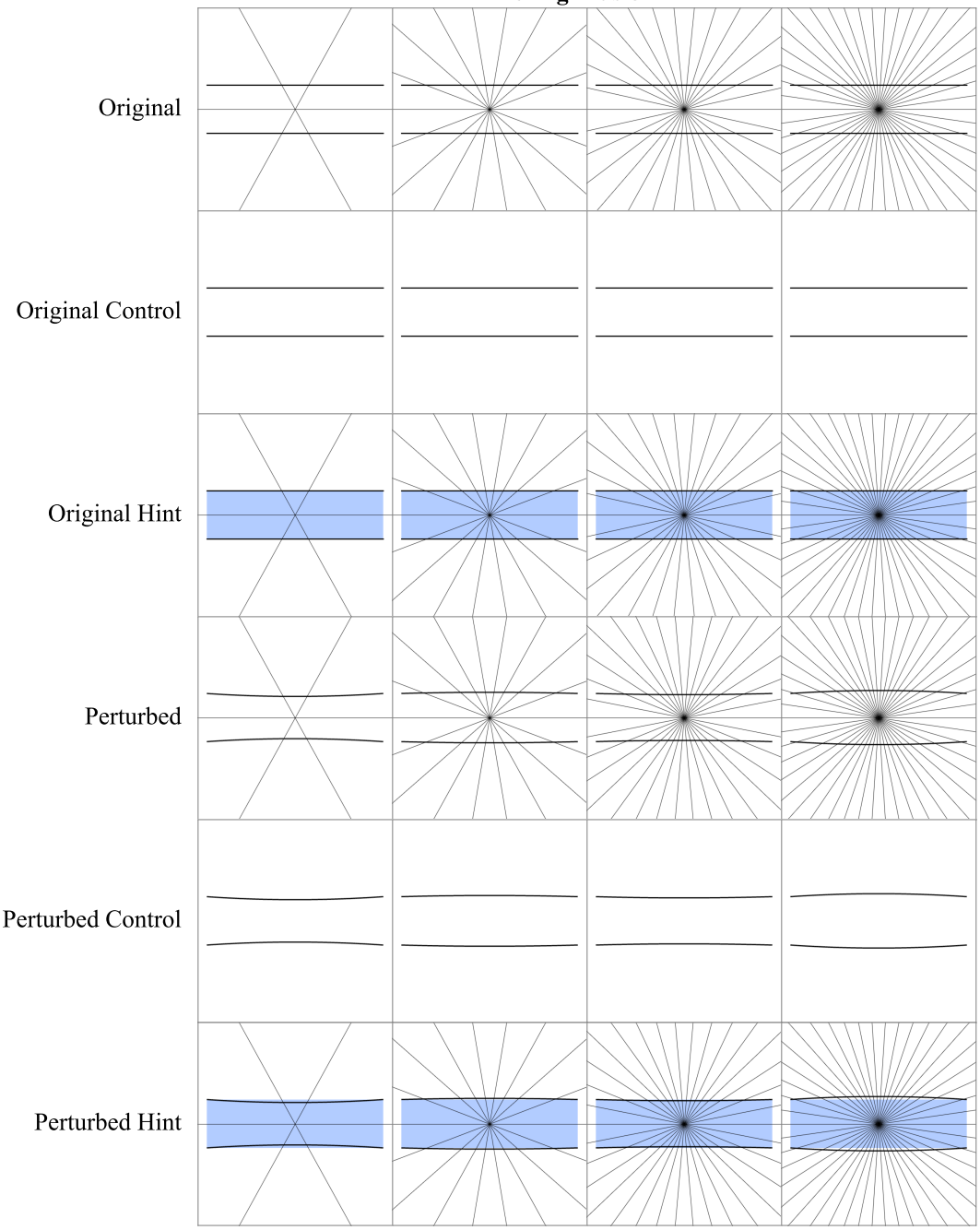



Chubb Illusion
Original
Original Control
Original Hint
Perturbed
Perturbed Control
Perturbed Hint

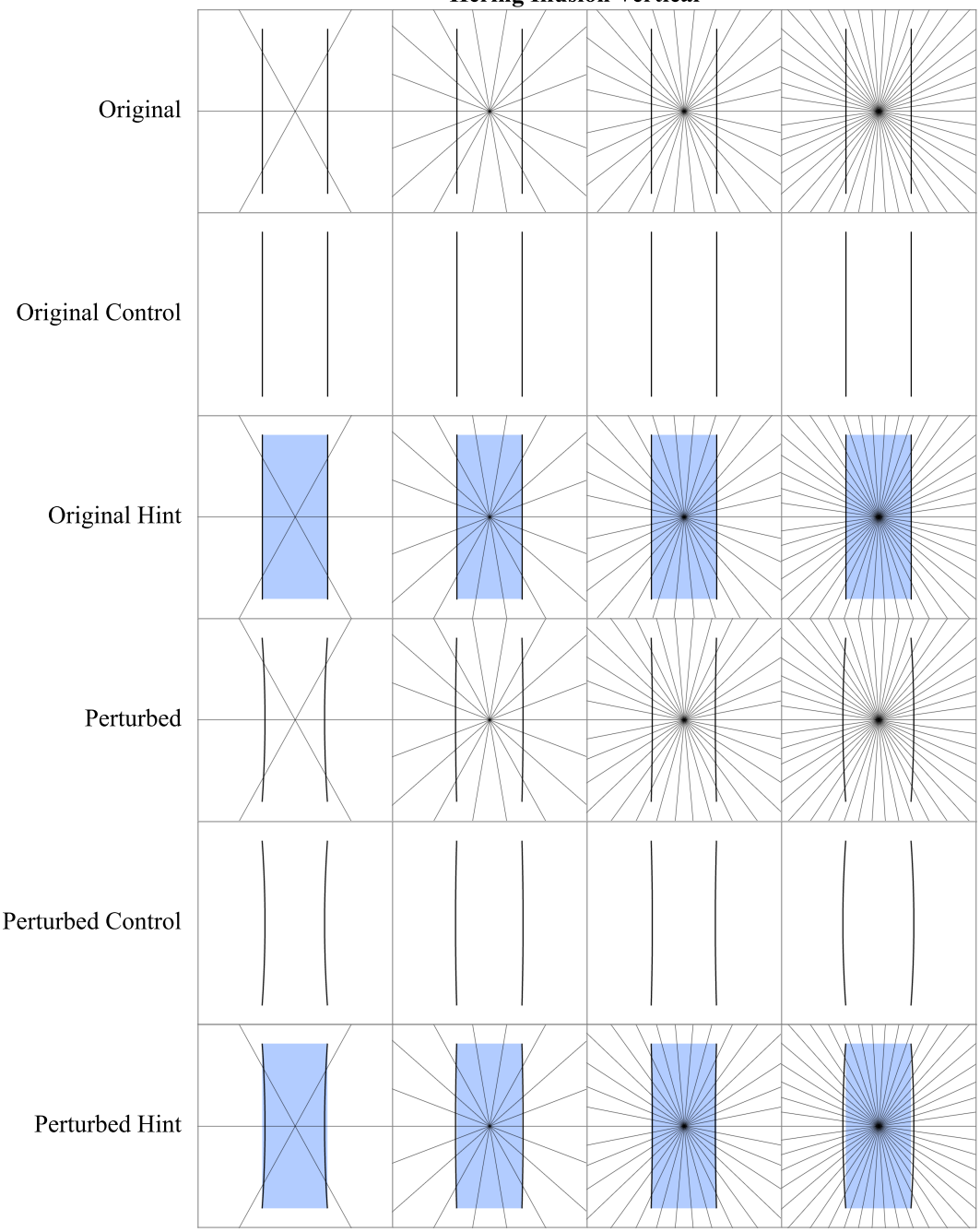
Cornswweet Illusion Case1



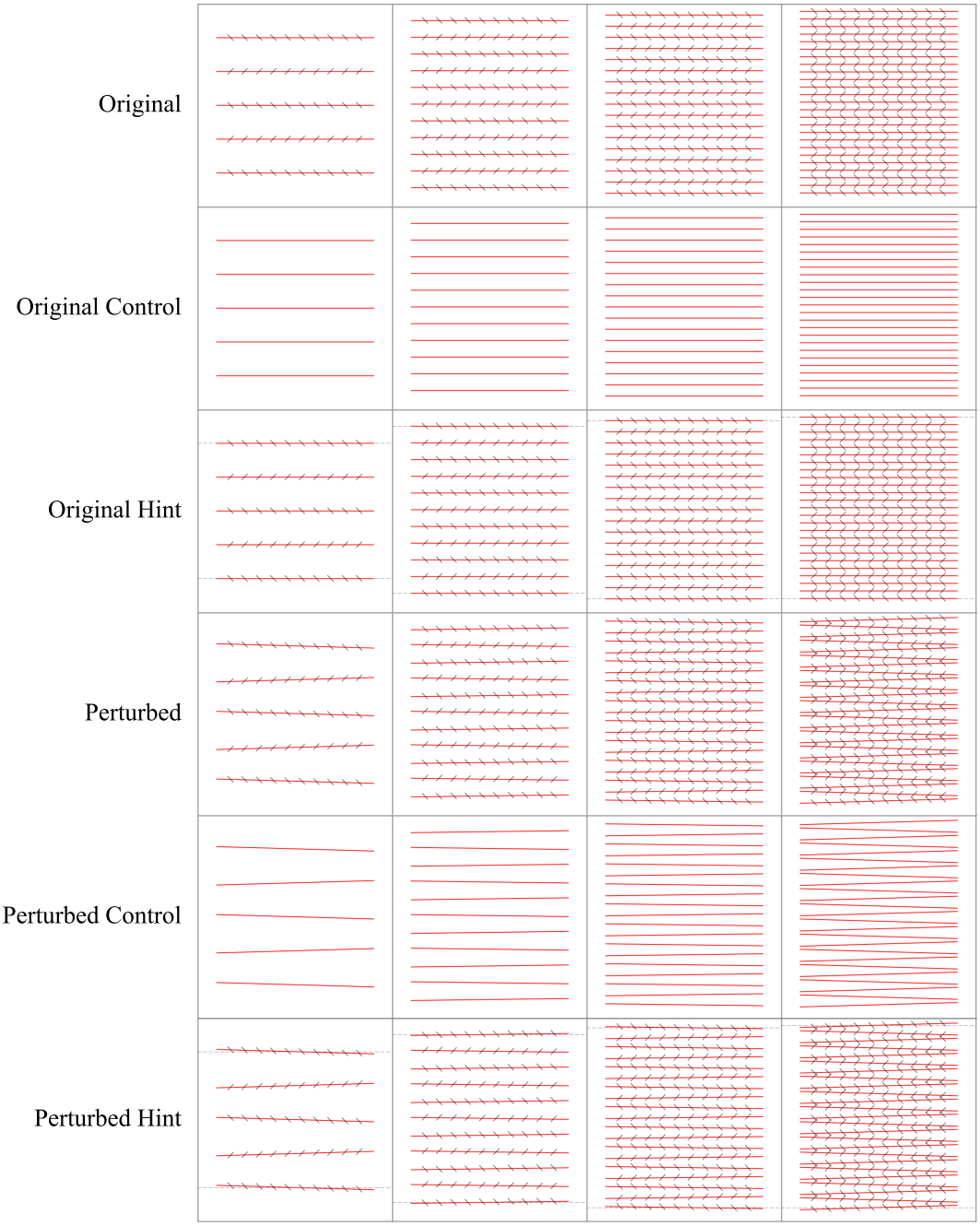
Hering Illusion



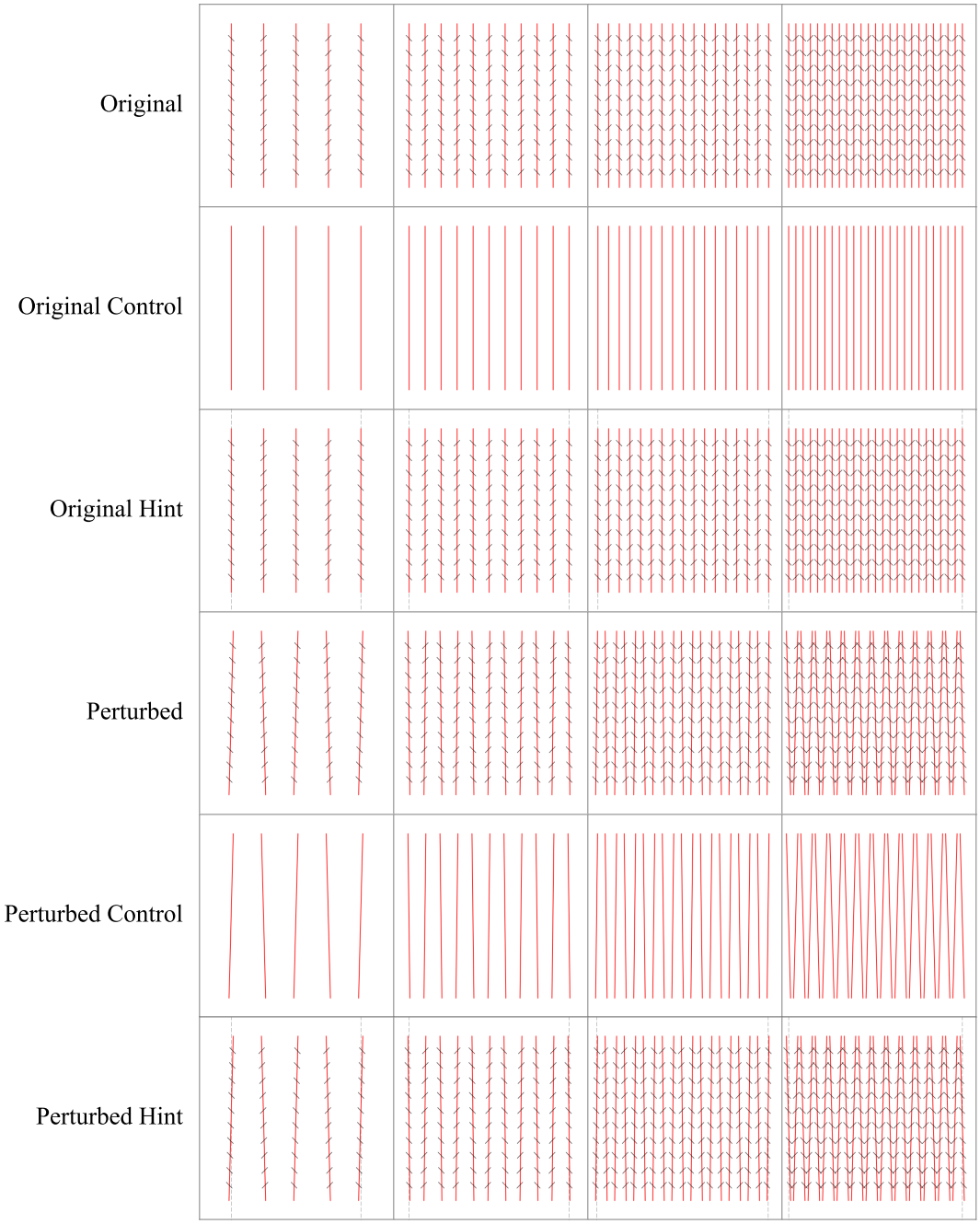
Hering Illusion Vertical



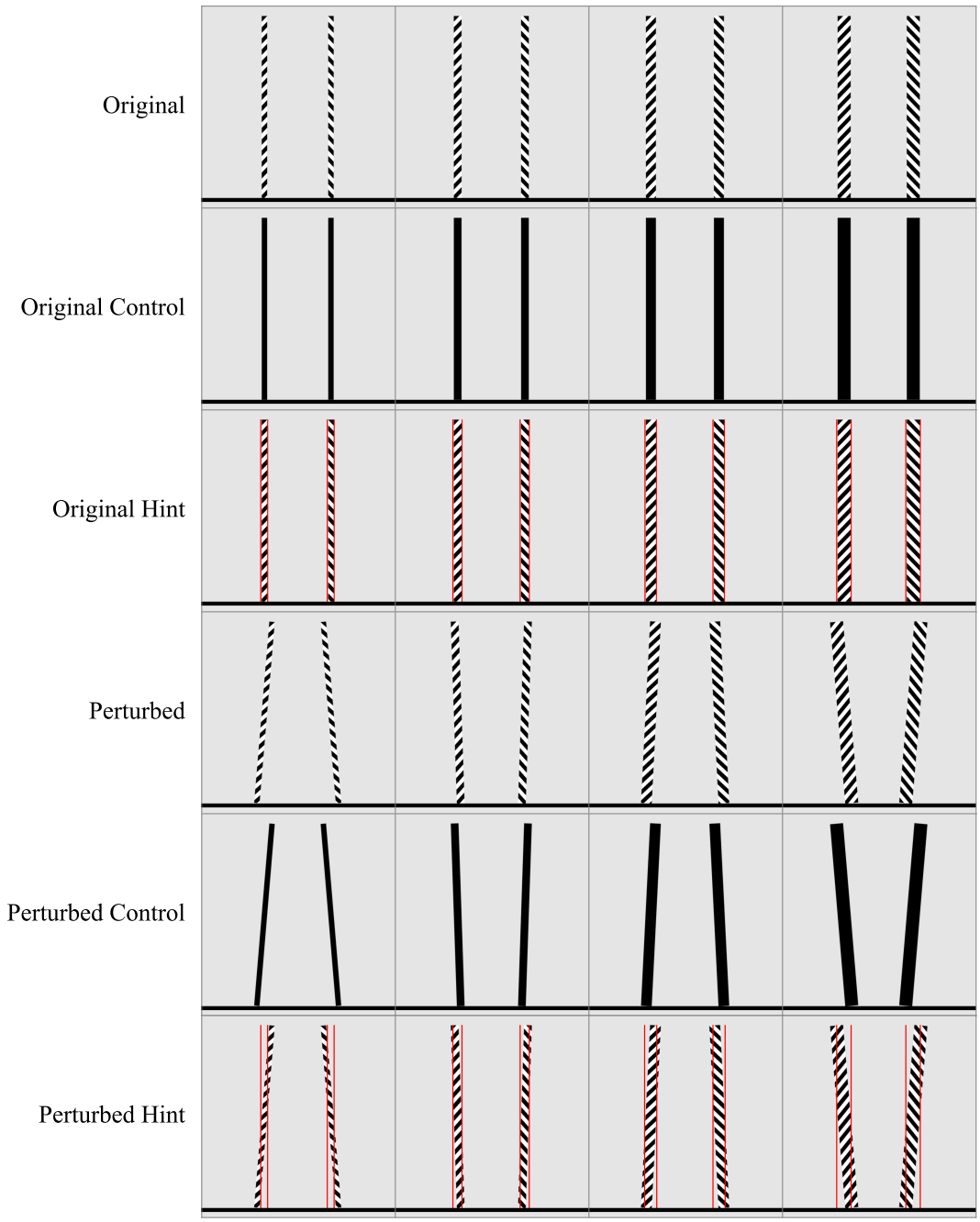
Zollner Illusion



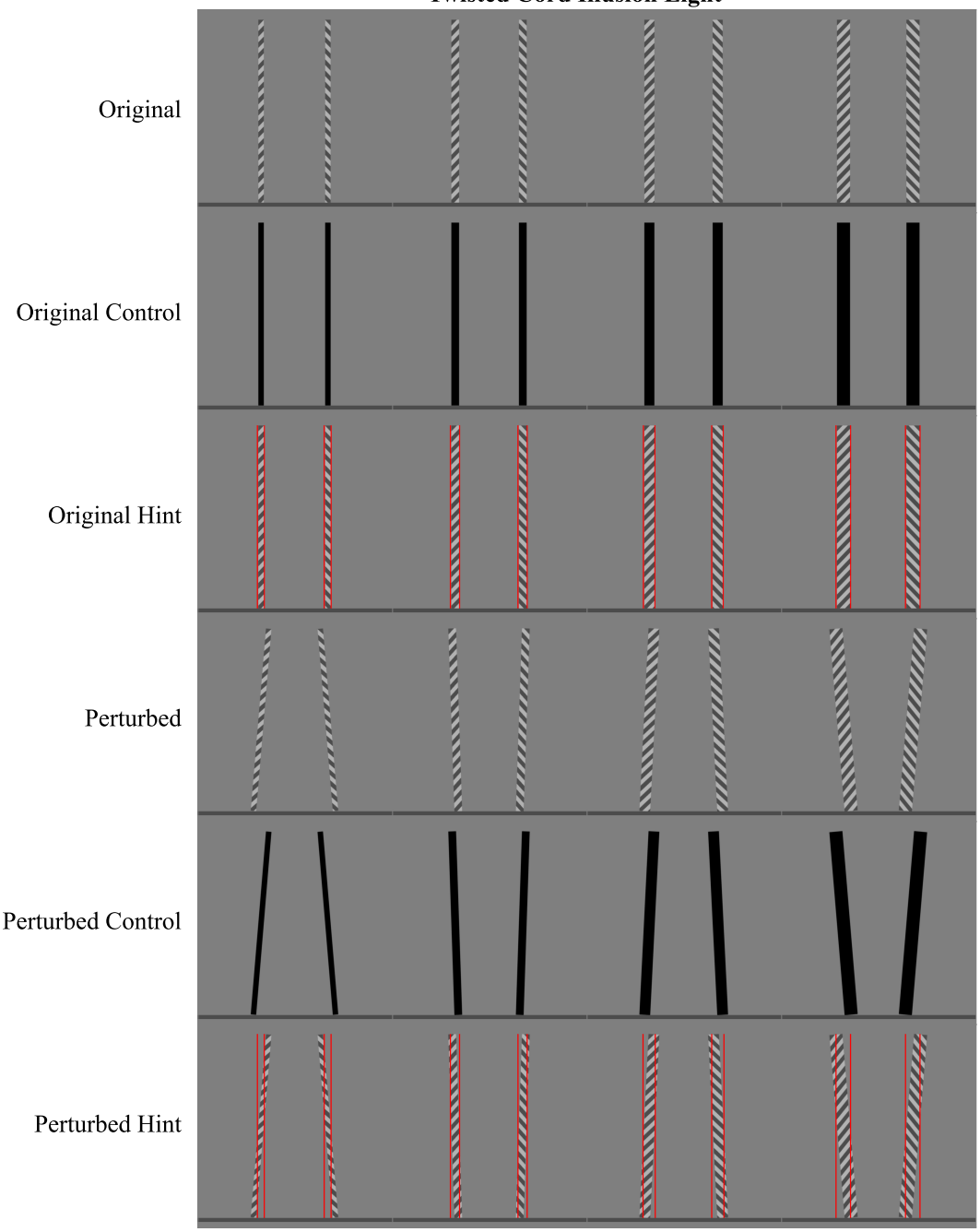
Zollner Illusion Vertical



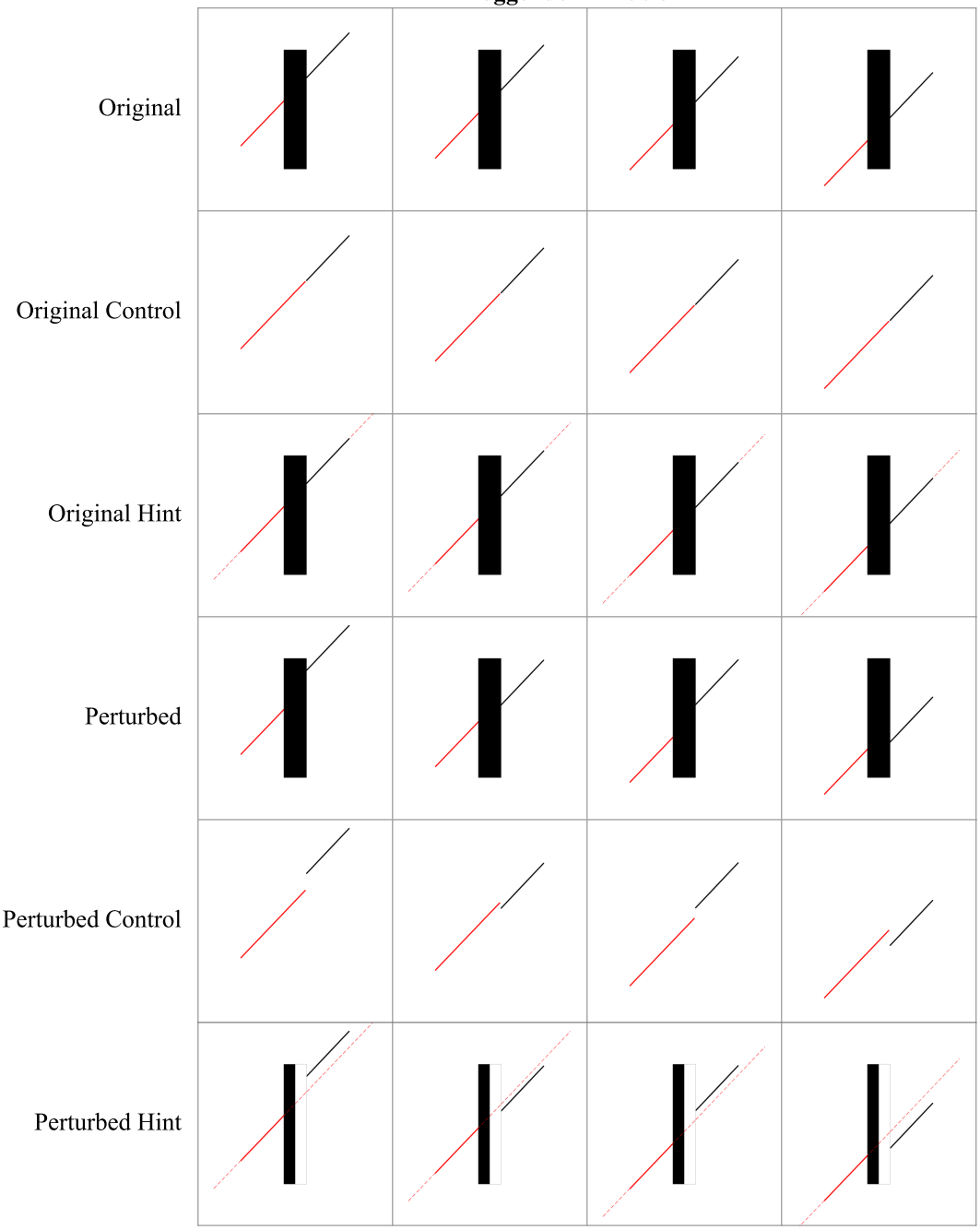
Twisted Cord Illusion



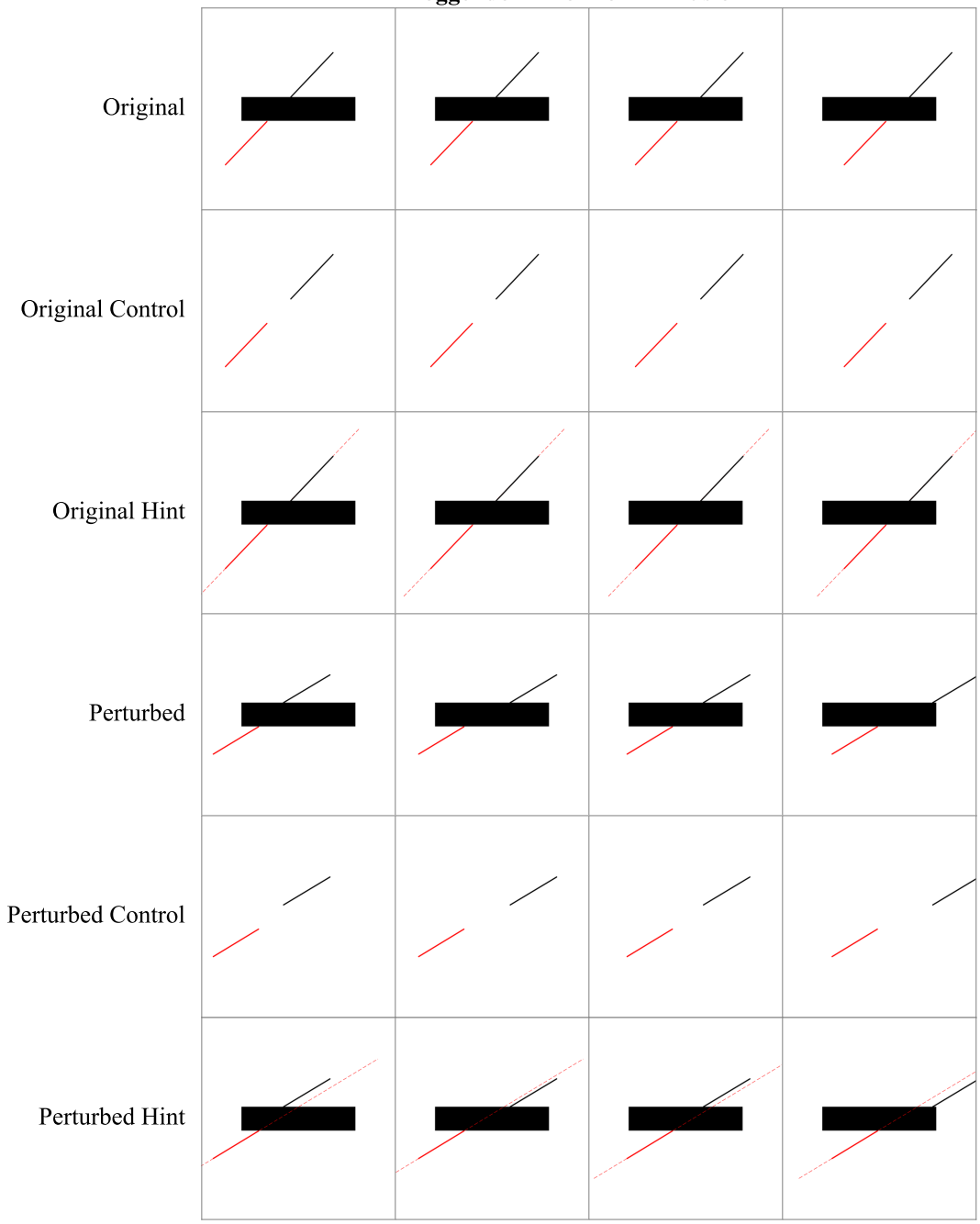
Twisted Cord Illusion Light



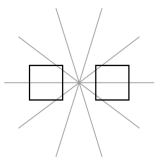
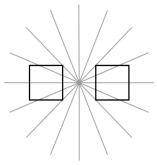
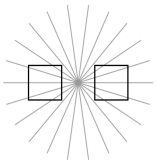
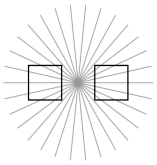
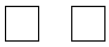
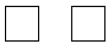
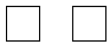
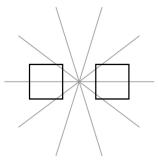
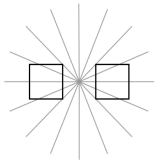
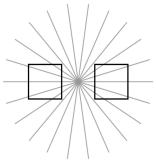
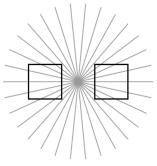
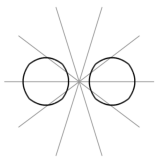
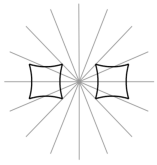
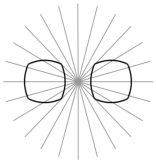
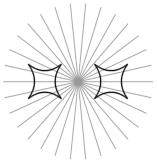
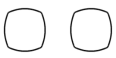
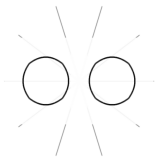
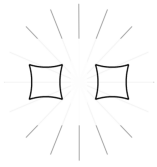
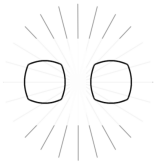
Poggendorff Illusion



Poggendorff Horizontal Illusion



Ehrenstein Illusion

Original				
Original Control				
Original Hint				
Perturbed				
Perturbed Control				
Perturbed Hint				

C.3. More examples of visual embeddings

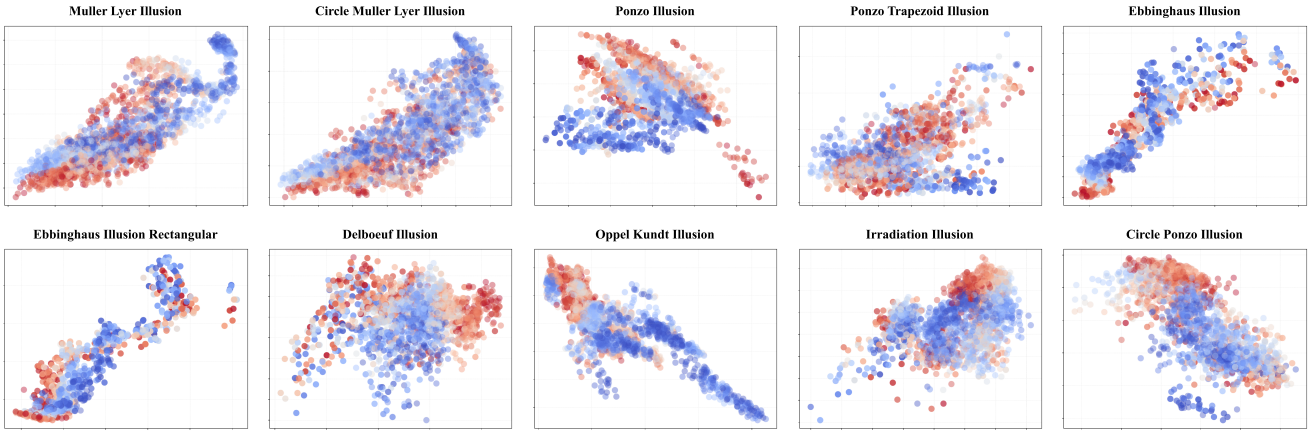


Figure 1. Visualization of Qwen2.5-VL-72B embeddings on size cases color-coded by perturbation strength (excluding the case shown in the main paper). The clear separation between perturbation levels validates the data generation pipeline.

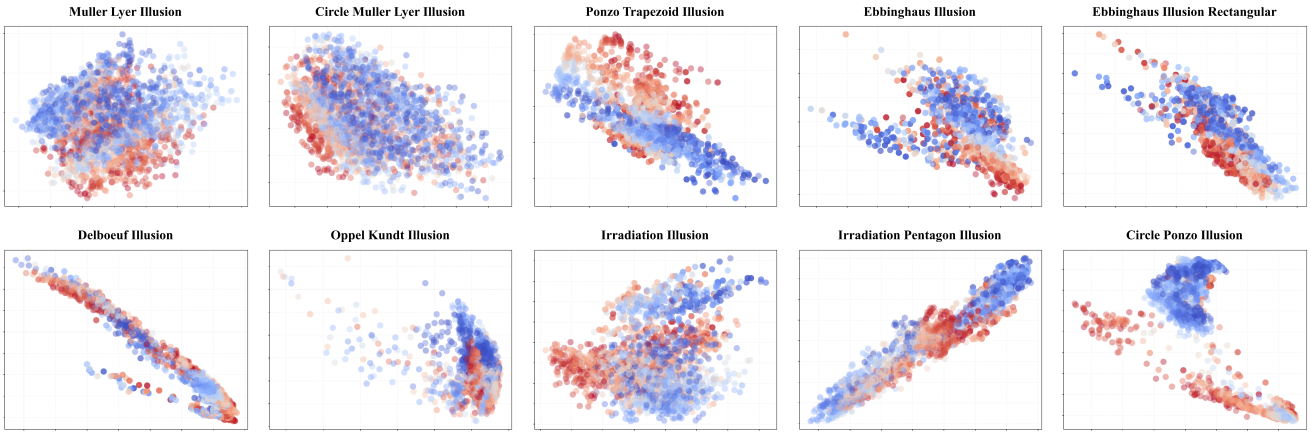


Figure 2. Visualization of Qwen2.5-VL-3B embeddings on size cases color-coded by perturbation strength (excluding the case shown in the main paper). The clear separation between perturbation levels validates the data generation pipeline.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Anthropic. System card: Claude opus 4 & claude sonnet 4. Technical report, Anthropic, 2025. 3

[3] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 3

[4] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. 3

[5] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3