**Title Idea:** *From Pixels to Profit: A Multimodal, Multitask System for Generating and Optimizing E-Commerce Listings*

**1. Abstract**

* **Problem:** State the high cost and inefficiency of manually creating high-quality e-commerce product listings. Mention common issues like factual errors in AI-generated text and the lack of actionable insights.
* **Solution:** Introduce your system (give it a name, e.g., "ListingGPT" or "Opti-List"). Describe it as an end-to-end, multimodal system that generates complete, structured JSON listings from a single product image.
* **Core Contributions (The 3 Claims):** Concisely state your key contributions.
  1. A **hierarchical generation** process that first predicts structured attributes to ensure factually-grounded descriptive text. **(HG)**
  2. A **holistic visual representation** learned via multi-task learning, enabling a single vision backbone to capture nuanced visual features for improved attribute and price prediction. **(HVR)**
  3. A novel **prescriptive analytics pipeline** that moves beyond generation to provide actionable suggestions for product attribute optimization. **(PA)**
* **Results:** Tease your key findings. "Our system outperforms strong baselines in factual consistency, price prediction accuracy, and provides novel, actionable business insights, as validated by extensive automated and human evaluations."

**2. Introduction**

* **Hook:** Start with the scale of e-commerce and the critical role of the product listing. A good listing drives sales; a bad one is a liability.
* **The Challenge:** Detail the current problems. Manual creation is slow. Existing AI tools often work in silos (one tool for backgrounds, one for text), are prone to "hallucinating" details not in the image, and are purely *descriptive* (they tell you what is) rather than *prescriptive* (they tell you what you should do).
* **Our Approach:** Introduce your system as the solution. Frame your three claims as direct answers to these challenges.
  + "To combat factual hallucinations, we propose a **hierarchical generation (HG)** process..."
  + "To move beyond siloed models and capture subtle visual cues like fabric quality, we employ multi-task learning to develop a **holistic visual representation (HVR)**..."
  + "Finally, to bridge the gap from description to decision-making, we introduce a **prescriptive analytics (PA)** engine..."
* **Contributions Summary:** Clearly list your 3 contributions again in a bulleted list.
* **Paper Structure:** Briefly outline the remaining sections.

**3. Related Work**

* **Multimodal Learning for E-commerce:** Discuss prior work on attribute extraction from images, image-to-text generation for product descriptions, and visual-based price prediction.
* **Hierarchical & Controlled Text Generation:** Situate your **(HG)** claim. Discuss related concepts like chain-of-thought prompting, retrieval-augmented generation, and other methods that use intermediate structured data to guide LLMs and reduce hallucinations.
* **Multi-Task Learning in Computer Vision:** Contextualize your **(HVR)** claim. Cover how MTL has been used to learn more robust and generalizable representations by sharing knowledge across related tasks.
* **AI for Business Intelligence:** Frame your **(PA)** claim. Briefly cover predictive analytics in retail, but highlight that applying this in a *generative, "what-if" simulation context* on product attributes is a novel contribution.

**4. Methodology**

* **Overall Architecture:** Start with a high-level diagram of the entire system, from image input to JSON output (including the prescriptive loop). This is the most important figure in the paper.
* **Section 4.1: The Holistic Visual Representation Backbone (HVR)**
  + Describe the shared vision encoder (e.g., ViT).
  + Detail the multiple heads: the attribute classification head, the price regression head, and the input projection for the text decoder.
  + Explain the joint loss function (e.g., L\_total = α\*L\_attr + β\*L\_price + γ\*L\_text).
* **Section 4.2: Hierarchical Text Generation (HG)**
  + Explain the two-step flow. First, the attribute head predicts a set of structured tags A = {a1, a2, ...}.
  + Describe how these predicted attributes A are then used to condition the text decoder. Are they prepended to the input sequence? Used in cross-attention? Be specific. This directly explains how you ensure factual grounding.
* **Section 4.3: The Prescriptive Analytics Engine (PA)**
  + Explain the "what-if" simulation logic. Detail how you systematically perturb a single predicted attribute (e.g., change color='Blue' to color='Black') while holding others constant.
  + Describe how you feed these new, hypothetical attribute sets back into the trained price prediction component (either the regression head of the MTL model or the XGBoost model) to get a predicted price/rating change.

**5. Experiments**

* **Dataset:** Describe your dataset, its source, size, and preprocessing steps.
* **Evaluation Metrics:** List the metrics you defined in your plan for attributes, price, and text. Explicitly define your "Factual Consistency Metric".
* **Baselines and Ablations:** Formally introduce all the models you are comparing against, using the naming convention from earlier (Baseline-Siloed, Baseline-Hybrid, etc.). Refer back to your core claims to justify why each baseline is important.
* **Implementation Details:** Provide hyperparameters, model sizes, and training infrastructure.

**6. Results and Discussion**

* **Structure this section around your research questions/claims, not just tables of numbers.**
* **6.1: The Value of a Holistic Visual Representation (HVR)**
  + Present a table comparing Ours-MTL against Baseline-Siloed, Ours-Ablate-NoMTL, and critically, Baseline-Hybrid on price prediction (MAE, RMSE) and attribute prediction (F1, EMR).
  + **Narrative:** Argue that your MTL model's performance on price prediction, especially relative to the powerful Baseline-Hybrid, demonstrates its ability to capture subtle visual information beyond what structured attributes alone can convey. Discuss the efficiency gains (training time, inference speed).
* **6.2: Hierarchical Generation Reduces Factual Hallucinations (HG)**
  + Present a table with text generation metrics (BLEU, ROUGE, BERTScore, and **Factual Consistency**). Compare Ours-MTL with Baseline-DirectVLM and Ours-Ablate-NoHier.
  + **Narrative:** The Factual Consistency score is your hero here. Show that your model significantly outperforms baselines on this metric. Include a qualitative example (Figure) showing a baseline hallucinating an attribute ("long sleeves" on a t-shirt) while your model correctly describes it based on the predicted attributes.
* **6.3: Prescriptive Analytics for Actionable Insights (PA)**
  + This section will be more qualitative. Show a compelling example of the final JSON output, complete with the suggestions block.
  + Report the results of your user study (e.g., "X% of users found the output with suggestions more valuable").
  + **Narrative:** Argue that this component transforms the tool from a passive content generator into an active business strategy partner.

**7. Conclusion**

* **Summary:** Briefly restate the problem and your solution.
* **Reiterate Contributions:** Summarize your three key contributions and how your results validated them.
* **Limitations and Future Work:** Acknowledge any limitations (e.g., limited to a single product domain, dependent on dataset label quality). Suggest exciting future directions (e.g., extending to video, incorporating user reviews as a modality, live A/B testing of suggestions).